



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



CLADAG 2013

9th SCIENTIFIC MEETING OF THE CLASSIFICATION
AND DATA ANALYSIS GROUP OF THE
ITALIAN STATISTICAL SOCIETY

September 18 - 20, 2013

University of Modena and Reggio Emilia
San Geminiano Complex - Modena, Italy

Book of Abstracts

Editors: Tommaso Minerva, Isabella Morlini, Francesco Palumbo

CLEUP

ISBN: 9788867871179

Patronage



Dipartimento di
Comunicazione ed Economia

Dipartimento di
Scienze Fisiche, Informatiche e Matematiche

'Vc dng'qh'E qpvgpvu'

''

- Tommaso Agasisti, Patrizia Falzetti** pag. 2
Socioeconomic sorting and test scores: an empirical analysis in the Italian junior secondary schools
- Dario Albarello, Vera D'Amico** pag. 9
Empirical testing of probabilistic seismic hazard models
- Federico Andreis, Pier Alda Ferrari** pag. 15
A proposal for the multidimensional extension of CUB models
- Morten Arendt Rasmussen, Evrim Acar** pag. 19
Data fusion in the framework of coupled matrix tensor factorization with common, partially common and unique factors
- Luigi Augugliaro, Angelo M. Mineo** pag. 20
Estimation of Sparse Generalized Linear Models: the dglars package
- Antonio Balzanella, Lidia Rivoli, Elvira Romano** pag. 24
A comparison between two tools for data stream summarization
- Lucio Barabesi, Giancarlo Diana, Pier Francesco Perri** pag. 28
Gini Index Estimation in Randomized Response Surveys
- Francesco Bartolucci, Federico Belotti, Franco Peracchi** pag. 32
A test for time-invariant individual effects in generalized linear models for panel data
- Erich Battistin, Carlos Lamarche, Enrico Rettore** pag. 36
Identification of the distribution of the causal effect of an intervention using a generalised factor model
- Matilde Bini, Lucio Masserini** pag. 37
Internal effectiveness of educational offer and students' satisfaction: a SEM approach
- Matilde Bini, Leopoldo Nascia, Alessandro Zeli** pag. 41
Groups heterogeneity and sectors concentration: a structural equation modeling for micro level analysis of firms
- Giuseppe Boari, Marta Nai Ruscone** pag. 45
Use of Relevant Principal Components to Define a Simplified Multivariate Test Procedure of Optimal Clustering
- Giuseppe Boari, Gabriele Cantaluppi, Angelo Zanella** pag. 49
Some Distance Proposals for Cluster Analysis in Presence of Ordinal Variables

Laura Bocci, Donatella Vicari <i>A general model for INDCLUS with external information</i>	pag. 53
Paola Bongini, Paolo Trivellato, Mariangela Zenga <i>The financial literacy and the undergraduates</i>	pag. 57
Riccardo Bramante, Marta Nai Ruscone, Pasquale Spani <i>Credit risk measurement and ethical issue: some evidences from the italian banks</i>	pag. 61
Pierpaolo Brutti, Lucio Ceccarelli, Fulvio De Santis, Stefania Gubbiotti <i>On the Stylometric Authorship of Ovid's Double Heroïdes: An Ensemble Clustering Approach</i>	pag. 65
Silvia Caligaris, Fulvia Mecatti and Patrizia Farina <i>Causal Inference in Gender Discrimination in China: Nutrition, Health, Care</i>	pag. 69
Giorgio Calzolari, Antonino Di Pino <i>Self-Selection and Direct Estimation of Across-Regime Correlation Parameter</i>	pag. 73
Maria Gabriella Campolo, Antonino Di Pino, Ester Lucia Rizzi <i>Modern Vs. Traditional: A cluster-based specification of gender and familistic attitudes and their influence on the division of labour of Italian couples</i>	pag. 77
Gabriele Cantaluppi, Marco Passarotti <i>Clustering the Four Gospels in the Greek, Latin, Gothic and Old Church Slavonic Translations</i>	pag. 81
Carmela Cappelli, Francesca Di Iorio <i>Regression Trees for change point analysis: methods, applications and recent developments</i>	pag. 85
Roberto Casarin and Marco Tronzano and Domenico Sartore <i>Bayesian Stochastic Correlation Models</i>	pag. 89
Rosalia Castellano, Gennaro Punzo, Antonella Rocca <i>Evaluating the selection effect in labour markets with a low female participation</i>	pag. 93
Paola Cerchiello, Paolo Giudici <i>A statistical based H index for the evaluation of e-markets</i>	pag. 97
Annalisa Cerquetti <i>Bayesian nonparametric estimation of global disclosure risk</i>	pag. 101
Enrico Ciavolino, Roberto Savona <i>The Forecasting side of Sovereign Risk: a Generalized Cross Entropy Approach</i>	pag. 105

Nicoletta Cibella, Tiziana Tuoto, Luca Valentino <i>What data tell you that models can't say</i>	pag. 109
Roberto Colombi, Sabrina Giordano <i>Multiple Hidden Markov Models for Categorical Time Series</i>	pag. 114
Pier Luigi Conti, Daniela Marella <i>Asymptotics in survey sampling for high entropy sampling designs</i>	pag. 118
Claudio Conversano, Massimo Cannas, Francesco Mola <i>On the Use of Recursive Partitioning in Casual Inference: A Proposal</i>	pag. 122
Franca Crippa, Marcella Mazzoleni, Mariangela Zenga <i>Keeping the pace with higher education. A fuzzy states gender study</i>	pag. 128
F. Cugnata, C. Guglielmetti and S. Salini <i>CUB model to validate FACIT TS-PS measurement instrument</i>	pag. 133
Rosario D'Agata, Venera Tomaselli <i>Multilevel Approach in Meta-Analysis of Pre-Election Poll Accuracy</i>	pag. 137
Alfonso Iodice D'Enza and Angelos Markos <i>Low-dimensional tracking of association structures in categorical data</i>	pag. 141
Giulio D'Epifani <i>Self-censored Categorical Responses A device for recovering latent behaviors</i>	pag. 145
Pierpaolo D'Urso, Marta Disegna, Riccardo Massari <i>Tourism Market Segmentation with Imprecise Information</i>	pag. 150
Utkarsh J. Dang, Salvatore Ingrassia, Paul D. McNicholas and Ryan Browne <i>Cluster-weighted models for multivariate response and extensions</i>	pag. 154
Cristina Davino, Domenico Vistocco <i>Unsupervised Classification through Quantile Regression</i>	pag. 158
F. Marta L. Di Lascio, Simone Giannerini <i>A copula-based approach to discover inter-cluster dependence relationships</i>	pag. 162
Josè G. Dias, Sofia B. Ramos <i>Hierarchical market structure of Euro area regime dynamics</i>	pag. 166
Drago Carlo, Balzanella Antonio <i>Consensus Community Detection: a Nonmetric MDS Approach</i>	pag. 170
Fabrizio Durante, Roberta Pappad`a and Nicola Torelli <i>Clustering financial time series by measures of tail dependence</i>	pag. 174

Marco Enea, Antonella Plaia <i>Influence diagnostics for generalized linear mixed models: a gradient-like statistic</i>	pag. 178
Enrico Fabrizi, Maria R. Ferrante, Carlo Trivisano <i>Joint estimation of poverty and inequality parameters in small areas</i>	pag. 182
Giorgio Fagiolo, Andrea Roventini <i>Macroeconomic Policy in DSGE and Agent-Based Models</i>	pag. 187
Salvatore Fasola, Mariangela Sciandra <i>New Flexible Probability Distributions for Ranking Data</i>	pag. 191
Maria Brigida Ferraro, Paolo Giordani <i>A new fuzzy clustering algorithm with entropy regularization</i>	pag. 195
Camilla Ferretti, Piero Ganugi, Renato Pieri <i>Mobility measures for the dairy farms in Lombardy</i>	pag. 199
Silvia Figini, Marika Vezzoli <i>Model averaging and ensemble methods for risk corporate estimation</i>	pag. 203
Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, Agustín Mayo-Iscar <i>New proposals for clustering based on trimming and restrictions</i>	pag. 207
Andreas Geyer-Schulz, Fabian Ball <i>Formal Diagnostics for Graph Clustering: The Role of Graph Automorphisms</i>	pag. 211
Massimiliano Giacalone, Angela Alibrandi <i>An overview on multiple regression models based on permutation tests</i>	pag. 215
Francesca Giambona, Mariano Porcu <i>The determinants of Italian students' reading scores: a Quantile Regression analysis</i>	pag. 219
Paolo Giordani, Henk A.L. Kiers, Maria Antonietta Del Ferraro <i>The R Package ThreeWay</i>	pag. 223
Giuseppe Giordano, Ilaria Primerano <i>Co-occurrence Network from Semantic Differential Data</i>	pag. 227
Paolo Giudici <i>Financial risk data analysis</i>	pag. 231
Silvia Golia, Anna Simonetto <i>A Comparison between SEM and Rasch model: the polytomous case</i>	pag. 237
Anna Gottard <i>Some considerations on VCUB models</i>	pag. 241

Francesca Greselin, Salvatore Ingrassia <i>Data driven EM constraints for mixtures of factor analyzers</i>	pag. 245
Leonardo Grilli, Carla Rampichini, Roberta Varriale <i>Predicting students' academic performance: a challenging issue in statistical modelling</i>	pag. 249
Luigi Grossi, Fany Nan <i>Robust estimation of regime switching models</i>	pag. 255
Kristian Hovde Liland <i>Variable selection in sequential multi-block analysis</i>	pag. 259
Maria Iannario <i>Robustness issues for a class of models for ordinal data</i>	pag. 260
Maria Iannario, Domenico Piccolo <i>A class of ordinal data models in R</i>	pag. 264
Salvatore Ingrassia, Antonio Punzo <i>Parsimony in Mixtures with Random Covariates</i>	pag. 268
Hiroshi Inoue <i>International Relations Based on the Voting Behavior in General Assembly</i>	pag. 272
Carmela Iorio, Massimo Aria, Antonio D'Ambrosio <i>Visual model representation and selection for classification and regression trees</i>	pag. 276
Monia Lupparelli, Luca La Rocca, Alberto Roverato <i>Log-Mean Linear Parameterizations for Smooth Independence Models</i>	pag. 284
Marica Manisera, Paola Zuccolotto <i>Nonlinear CUB models</i>	pag. 288
Marica Manisera, Marika Vezzoli <i>Finding number of groups using a penalized internal cluster quality index</i>	pag. 292
Daniela Marella, Paola Vicard <i>Object-Oriented Bayesian Network to deal with measurement error in household surveys</i>	pag. 296
Angelos Markos, Alfonso Iodice D'Enza, Michel Van de Velden <i>Beyond tandem analysis: joint dimension reduction and clustering in R</i>	pag. 300
F. Martella and M. Alfò <i>A biclustering approach for discrete outcomes</i>	pag. 304

Mariagiulia Matteucci, Stefania Mignani, Roberto Ricci <i>A Multidimensional IRT approach to analyze learning achievement of Italian students</i>	pag. 309
Sabina Mazza <i>Extending the Forward Search to the Combination of Multiple Classifiers: A Proposal</i>	pag. 314
Fulvia Mecatti, M. Giovanna Ranalli <i>Plug-in Bootstrap for Sample Survey Data</i>	pag. 318
Alessandra Menafoglio, Matilde Dalla Rosa and Piercesare Secchi <i>A BLU Predictor for Spatially Dependent Functional Data of a Hilbert Space</i>	pag. 322
Maria Adele Milioli, Lara Berzieri, Sergio Zani <i>Comparing fuzzy and multidimensional methods to evaluate well-being at regional level</i>	pag. 326
Michelangelo Misuraca, Maria Spano <i>Comparing text clustering algorithms from a multivariate perspective</i>	pag. 331
Cristina Mollica, Luca Tardella <i>Mixture models for ranked data classification</i>	pag. 335
Isabella Morlini, Stefano Orlandini <i>Cluster analysis of three-way atmospheric data</i>	pag. 339
Roberto Nardecchia, Roberto Sanzo, Margherita Velucchi, Alessandro Zeli <i>Productivity transition probabilities: A microlevel data analysis for Italian manufacturing sectors (1998-2007)</i>	pag. 345
Andrea Neri, Giuseppe Ilardi <i>Interviewers, co-operation and data accuracy: is there a link?</i>	pag. 349
Akinori Okada, Satoru Yokoyama <i>Nonhierarchical Asymmetric Cluster Analysis</i>	pag. 353
Marco Perone Pacifico <i>SuRF: Subspace Ridge Finder</i>	pag. 357
Andrea Pagano, Francesca Torti, Jessica Cariboni, Domenico Perrotta <i>Robust clustering of EU banking data</i>	pag. 361
Giuseppe Pandolfo, Giovanni C. Porzio <i>On depth functions for directional data</i>	pag. 365
Andrea Pastore, Stefano F. Tonellato <i>A generalised Silhouette-width measure</i>	pag. 369

Fulvia Pennoni, Giorgio Vittadini <i>Hospital efficiency under two competing panel data models</i>	pag. 373
Alessia Pini, Simone Vantini <i>The Interval-Wise Control of the Family-Wise Error Rate for Testing Functional Data</i>	pag. 377
Mariano Porcu, Isabella Sulis <i>Detecting differences between primary schools in mathematics and reading achievement by using schools added-value measures of performance</i>	pag. 381
Antonio Punzo, Paul D. McNicholas, Katherine Morris, Ryan P. Browne <i>Outlier Detection via Contaminated Mixture Distributions</i>	pag. 387
Emanuela Raffinetti, Pier Alda Ferrari <i>New perspectives for the RDI index in social research fields</i>	pag. 392
Monia Ranalli, Roberto Rocci <i>Mixture models for ordinal data: a pairwise likelihood approach</i>	pag. 396
Marco Riani, Andrea Cerioli, Gianluca Morelli <i>Issues in robust clustering</i>	pag. 400
Stéphane Robin <i>Deciphering and modeling heterogeneity in interaction networks</i>	pag. 404
Rosaria Romano, Francesco Palumbo <i>Partial Possibilistic Regression Path Modeling</i>	pag. 409
Renata Rotondi <i>Classification of composite seismogenic sources through probabilistic score indices</i>	pag. 413
Gabriella Schoier <i>On Wild Bootstrap and M Unit Root Test</i>	pag. 417
Luca Scrucca <i>On the implementation of a parallel algorithm for variable selection in model-based clustering</i>	pag. 421
Paolo Sestito <i>The Role of Learning Measurement in the Governance of an Education System: an Overview of the Issues</i>	pag. 425
John Shawe-Taylor, Blaz Zlicar <i>Novelty Detection with Support Vector Machines</i>	pag. 430
Nadia Solaro <i>Multidimensional scaling with incomplete distance matrices: an insight into the problem</i>	pag. 431

Luigi Spezia, Cecilia Pinto <i>Markov switching models for high-frequency time series: flapper skate's depth profile as a case study</i>	pag. 435
Ralf Stecking, Klaus B. Schebesch <i>Data Privacy in Credit Scoring: Evaluating SVM Approaches Based on Microaggregated Data</i>	pag. 439
Isabella Sulis, Francesca Giambona, Nicola Tedesco <i>Analyzing university students' careers using Multi-State Models</i>	pag. 443
Luca Tardella, Danilo Alunni Fegatelli <i>BBRecap for Bayesian Behavioural Capture-Recapture Modeling</i>	pag. 447
Cristina Tortora, Paul D. McNicholas, Ryan P. Browne <i>Mixtures of generalized hyperbolic factor analyzers</i>	pag. 451
Giovanni Trovato <i>Testing for endogeneity and country heterogeneity</i>	pag. 455
Joaquin Vanschoren and Mikio L. Braun, Cheng Soon Ong <i>Open science in machine learning</i>	pag. 461
Valerio Veglio <i>Logistic Regression and Decision Tree: Performance Comparisons in Estimating Customers' Risk of Churn</i>	pag. 465
Maurizio Vichi <i>Robust Two-mode clustering</i>	pag. 469
Vincenzina Vitale <i>Hierarchical Graphical Models and Item Response Theory</i>	pag. 470
Sara Viviani <i>Extending the JM library</i>	pag. 474
Adalbert F.X. Wilhelm <i>Visualisations of Classification Tree Models: An Evaluative Comparison</i>	pag. 478

Socioeconomic sorting and test scores: an empirical analysis in the Italian junior secondary schools

Tommaso Agasisti, Politecnico di Milano School of Management,
e. tommaso.agasisti@polimi.it

Patrizia Falzetti, Invalsi - Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione. E. patrizia.falzetti@invalsi.it

Abstract. Recent research shows that Italian schools are segmenting classes' composition by students' ability and/or socioeconomic status (SES), the so-called "informal tracking". We propose an indicator to measure this phenomenon, defined as the variance between classes (within each school) in the Index of Economic, Social and Cultural Status (*ESCS_Var_Within*). By means of Instrumental Variables, we investigate the causal effects of this indicator on students' achievement. The results show that informal tracking is likely to exert a negative effect on achievement scores for Reading, but not Mathematics.

1 Introduction

Recent debates are focused on the renewed attention to the "tracking" phenomenon, which is the tendency of grouping students by ability into different classes and groups. While this practice has been abandoned during the eighties, on the basis of egalitarian claims, the resurgence of it has been justified on the ground of potential higher efficiency in teaching to homogeneous student groups.

Theoretically, the "segmentation" between classes is not possible in Italy; indeed, the law imposes to school principals to organize classes to be characterized by "equal-heterogeneity", in other words to minimize between-classes variance in terms of students' prior results and socioeconomic background and maximize within-classes variance. Some recent evidence, however, demonstrates that the practice of segmentation between classes is instead still adopted by many school principals, and refers to this phenomenon as "informal tracking" (Ferrer-Esteban, 2011). Also, such segmentation appears to be based more on students' socioeconomic status than on their prior achievement, and the two measures are likely to be correlated.

We derived, for each Italian junior secondary school a measure of "segmentation" in the socioeconomic composition of classes, defined as the extent to which there is variance between classes in the (average value of the) indicator used to measure the students' socioeconomic status (SES). The index, named ESCS (Economic, Social and Cultural Status) is computed analogously to the OECD's procedure, which is to consider the parents' occupation and education, possession of some kinds of goods, etc. The variable of interest measures the "dispersion" (variance) in the average ESCS between classes of the same school; more specifically, *ESCS_Var_Within*, for the k th school, is calculated as:

$$ESCS_Var_Within_k = \frac{1}{N_k} \sum (ESCS_{jk} - \overline{ESCS_k})^2 \quad (1)$$

where j indicates the j th class within the k th school, N_k is total number of students in the school, and $\overline{ESCS_k}$ is the ESCS average in the school.

If Italian schools actually respected the legislative principle of “equal-heterogeneity” in the composition of classes, then $ESCS_Var_Within$ should be close to zero. The actual distribution of the index is reported in the figure 1. As can be seen, the most consistent part of the distribution relies in the interval [0;10%]; however, there are schools in which this indicator is higher, even >20% (see also the statistics by percentiles of this variable).

<Figure 1> around here

The research question of this paper can be formulated as follows: *which is the causal effect of socioeconomic segmentation between classes on students’ achievement?* The next section 2 illustrates the methodology, with special reference to the IV strategy proposed, together with a brief explanation of our dataset. The section 3 contains the preliminary results.

2 Methodology and data

We estimate an educational production function of the following general type:

$$Y_{ijkt} = \alpha_0 + \lambda Y_{ijk(t-1)} + \alpha_1 X_{1ijkt} + \alpha_2 X_{2jkt} + \alpha_3 X_{3kt} + \varepsilon_{ijkt} \quad (2)$$

where Y_{ijkt} is the score in Reading (Mathematics) of the i th student, attending the j th class and the k th school (time t); $Y_{ijk(t-1)}$ is the test score obtained in the year before ($t-1$), X_{1ijkt} is a vector of student-level characteristics, X_{2jkt} is a vector of class-level variables, and X_{3kt} contains the school-level variables. As prior achievement is not available for all the students, a modified version of equation w is also estimated, in which only student, class and school-level *contemporaneous* characteristics are considered. We refer to the latter as “baseline model”, and to the former as “restricted sample”.

Among the school-level factors (vector X_{3kt}), we include our variable of interest, $ESCS_Var_Within$, which measures the variance in the ESCS index between classes. The other variables are listed in table 1, where we report also short notes about each of them and descriptive statistics. Data refer to all the around 500,000 Italian students, attending the around 5,000 junior secondary schools in the 2011/12 year.

<Table 1> around here

As $ESCS_Var_Within$ is likely to be endogenous with respect to achievement (as students are not randomly assigned to schools with a specific index of the variable), we adopt an instrumental variables (IV) approach (Angrist & Pischke, 2008). Specifically, we looked for a variable, which is related to the endogenous variable but not to the outcome. Following the intuition of Collins & Gan (2013), we used the variance between classes within each school in the year before (2010/11). Under the hypothesis that sorting students across classes on the basis of their socioeconomic background is not a random phenomenon, but a deliberate choice made by school’s managers, the index must be correlated with that of the prior year – which in turn is by definition

uncorrelated with the academic performance of current students. Mathematically, we substitute a prediction of $ESCS_VAR_Within$ in the equation 2, obtained by estimating the following first-stage equation:

$$ESCS_VAR_Within_{kt} = \beta_0 + \delta Y_{ijk(t-1)} + \beta_1 X_{1ijkt} + \beta_2 X_{2jkt} + \beta_3 X_{3kt} + \eta Z_{ijkt} + \varepsilon_{1ijkt} \quad (3)$$

where Z_{ijkt} is the “exclusion restriction” (the “Instrument”), which acts as a source of randomness in treatment assignment, that is the student being “assigned” to a school where $ESCS_VAR_Within$ is high or low. Using such a (two-stage) approach, the estimated coefficient for $ESCS_Var_Within$ in the main equation can be considered an unbiased measure of the (causal) effect of the socioeconomic segmentation between classes on students’ results.

With the aim of considering intra-school correlations in educational activities, the standard errors are clustered at school-level (we prefer this option to the clustering at class-level because our variable of interest is measured at school-level).

3 Preliminary results

The preliminary results of our analysis are presented in table 2. In the columns (a) and (b), baseline estimations are reported, about Reading and Mathematics, respectively; the estimates based on the “restricted” sample (those students for whom prior achievement is available) are instead illustrated in the columns (c) and (d). Standardized (beta) coefficients are reported.

<Table 2> around here

All the coefficients related to the student, class and school level appears with the expected sign. Prior achievement (as measured by the INVALSI test at grade 5) has a statistically significant and strong impact on grade 6 results (around 0.5 standard deviations for both Reading and Mathematics). Immigrant status is negatively related to achievement, and this is true more for Reading than for Mathematics, with virtually no distinction between first and second generation immigrants. The index of Economic, Social and Cultural Status (ESCS) is positively associated with achievement, and the magnitude of this effect is high (around 0.15 s.d.). Also peer-effects have a role, as the positive coefficient attached to class-average ESCS suggests. Many variables at class level control for irregularities in the testing procedure (an index of “cheating propensity”, a dummy for classes selected to be part of a selected sample, and the proportion of students who took part to the test). Dummies indicating if the school is located in the Central or Southern Italy confirm the well-know achievement gap between the areas of the country, with students attending a school in the South obtaining, on average, a score in the standardized test which is around 0.1 s.d. lower (around 1,5 points out of 100), all else equal, when compared to a student attending a school in the North.

Turning to our variable of interest, there is not estimated effect of $ESCS_Var_Within$ on Mathematics test score, while a negative effect exerted on Reading test score (magnitude: 0.05 s.d.) is estimated. It is possible that less diverse classes exacerbate the difficulties of students in Reading activities, which are more related to the day-by-day social interactions.

In subsequent steps, we will investigate if the impact of *ESCS_VAR_Within* is heterogeneous on subpopulations of students and schools, and particularly (i) schools in different geographical areas (northern, central and southern Italy), (ii) students attending public and private schools.

References

1. Angrist, J., and Pischke S. (2008), *Mostly Harmless Econometrics: An Empiricists' Companion*, Princeton University Press, Princeton, NJ.
2. Collins, C.A., Gan, L. (2013), Does sorting students improve scores? An analysis of class composition, NBER Working Paper n. 18848.
3. Ferrer-Esteban, G. (2011), Beyond the traditional territorial divide in the Italian Education System – effects of system management factors on performance in lower secondary school, Fondazione Agnelli Working Paper, 43.

Socioeconomic sorting and test scores

Table 1: Descriptive statistics

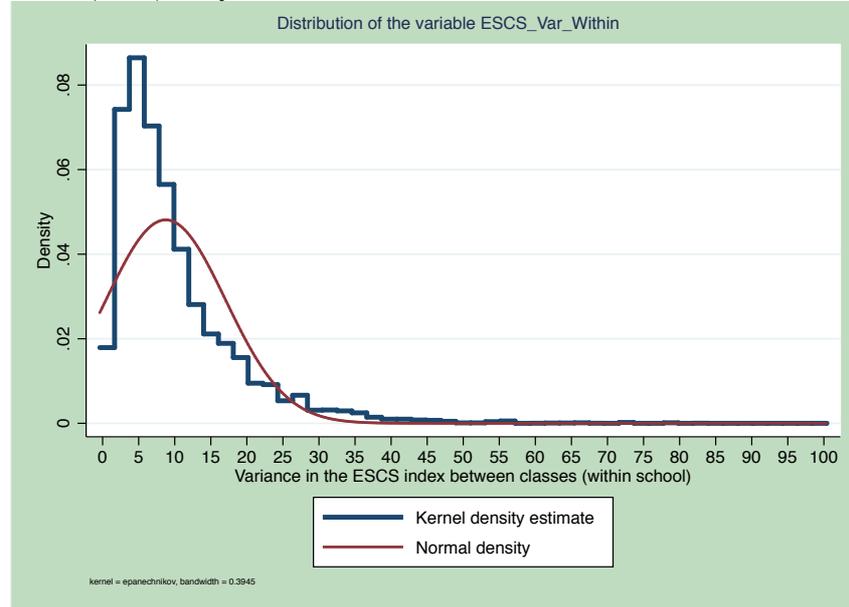
Variable	Mean	Std. Dev.	Min	Max	Obs
Student-level characteristics					
Prior achievement (grade 5)	74.216	13.834	0	100	241955
Female student	0.489	0.500	0	1	510032
1st generation immigrant	0.055	0.228	0	1	469517
2nd generation immigrant	0.049	0.216	0	1	469517
Early-enrolled student	0.022	0.145	0	1	510019
Late-enrolled student	0.073	0.260	0	1	510019
Socioeconomic background (index ESCS; mean=0, stdev=1)	0.137	1.045	-3.114	2.673	468203
Student who does NOT live with both parents	0.137	0.344	0	1	490272
Student who has siblings	0.842	0.364	0	1	490722
Classroom-level characteristics					
Cheating propensity	0.074	0.193	0	1	510873
Classroom selected to be part of the "controlled" sample	0.077	0.266	0	1	510933
Class-average socioeconomic background (index ESCS)	0.136	0.553	-2.670	2.264	469822
Proportion of females in the classroom	43.579	10.641	0	116.667	510757
Proportion of 1st generation immigrants in the classroom	4.865	6.342	0	100	470367
Proportion of 2nd generation immigrants in the classroom	4.342	7.243	0	96.296	470367
Proportion of Early-enrolled students in the classroom	1.954	4.035	0	56.000	510757
Proportion of Late-enrolled students in the classroom	6.401	6.456	0	100	510757
Proportion of disabled students in the classroom	5.417	5.482	0	67	510933
Number of students in the classroom	23.145	3.574	1	35	510933
Proportion of students who took the test	94.262	7.601	3.704	100	510933
Class with " <i>tempo-pieno</i> "	0.024	0.153	0	1	470367
School-level characteristics					
Number of classrooms in the school	6.482	3.079	2	17	510933
Number of students in the school	149.638	77.554	11	428	510933
Average number of students per class, in the school	22.695	2.976	3.667	32.333	510933
School located in Northern-West Italy	0.253	0.435	0	1	510933
School located in Central Italy	0.177	0.382	0	1	510933
School located in Southern Italy	0.391	0.488	0	1	510933
School located in PON area	0.333	0.471	0	1	510933
<i>Istituto Comprensivo</i>	0.606	0.489	0	1	510933
Private school	0.029	0.168	0	1	510933
Variance in the ESCS index between classes (ESCS_Var_Within)	8.758	8.290	0	100	480751

Table 2: Preliminary results

	Baseline results		Restricted sample	
	Reading (a)	Math (b)	Reading (c)	Math (d)
	b-coeff	b-coeff	b-coeff	b-coeff
Student-level characteristics				
Prior achievement (grade 5)			0.487*** (97.76)	0.500*** (102.59)
Female student	0.080*** (54.99)	-0.081*** (-52.49)	0.074*** (42.98)	-0.064*** (-37.21)
1st generation immigrant	-0.107*** (-51.71)	-0.045*** (-25.31)	-0.055*** (-24.54)	-0.016*** (-8.82)
2nd generation immigrant	-0.073*** (-43.12)	-0.043*** (-26.73)	-0.053*** (-28.92)	-0.026*** (-16.06)
Early-enrolled student	-0.011*** (-8.38)	-0.010*** (-7.40)	-0.009*** (-4.83)	-0.006*** (-3.62)
Late-enrolled student	-0.146*** (-81.77)	-0.107*** (-69.51)	-0.042*** (-17.47)	-0.025*** (-13.12)
Socioeconomic background (index ESCS; mean=0, stdev=1)	0.244*** (136.79)	0.227*** (122.57)	0.155*** (66.33)	0.136*** (61.64)
Student who does NOT live with both parents	-0.036*** (-23.69)	-0.043*** (-28.54)	-0.024*** (-13.53)	-0.027*** (-16.42)
Student who has siblings	-0.027*** (-18.96)	0.000 (0.11)	-0.016*** (-9.43)	0.004* (2.15)
Classroom-level characteristics				
Cheating propensity	0.166*** (53.61)	-0.168*** (-45.88)	0.084*** (19.71)	-0.120*** (-30.52)
Classroom selected to be part of the "controlled" sample	0.005* (2.06)	-0.008* (-2.83)	0.007** (2.60)	-0.008** (-2.80)
Class-average socioeconomic background (index ESCS)	0.003 (0.54)	0.011* (1.98)	0.016** (2.92)	0.019*** (4.62)
Proportion of females in the classroom	-0.003 (-1.16)	0.009** (2.68)	-0.002 (-0.50)	0.010** (2.68)
Proportion of 1st generation immigrants in the classroom	0.012*** (3.95)	-0.003 (-0.89)	0.014*** (3.64)	0.005 (1.28)
Proportion of 2nd generation immigrants in the classroom	0.013*** (4.26)	0.007 (1.71)	0.005 (1.70)	-0.003 (-0.86)
Proportion of Early-enrolled students in the classroom	0.013*** (3.50)	0.000 (-0.07)	0.009** (1.97)	-0.002 (-0.42)
Proportion of Late-enrolled students in the classroom	-0.029*** (-8.52)	-0.027*** (-7.26)	-0.014** (-3.47)	-0.013** (-3.25)
Proportion of disabled students in the classroom	-0.005 (-1.93)	0.000 (0.04)	0.001 (0.27)	0.006 (1.59)
Number of students in the classroom	0.029*** (7.96)	0.017*** (4.10)	0.024*** (5.23)	0.016** (3.17)
Proportion of students who took the test	0.063*** (15.91)	0.023*** (5.65)	0.033*** (7.99)	0.010* (2.54)
Class with "tempo-pieno"	-0.001 (-0.56)	0.001 (0.31)	-0.003 (-0.78)	0.000 (0.03)
School-level characteristics				
Number of classrooms in the school	-0.002 (-0.07)	-0.018 (-0.48)	-0.016 (-0.40)	-0.047 (-1.17)
Number of students in the school	0.018 (0.56)	0.032 (0.82)	0.039 (0.92)	0.066 (1.56)
Average number of students per class, in the school	-0.003 (-0.41)	0.002 (0.19)	-0.012 (-1.26)	-0.011 (-1.07)
School located in Northern-West Italy	0.018*** (5.79)	0.006 (1.32)	0.020*** (5.22)	0.004 (0.83)
School located in Central Italy	-0.010** (-2.78)	-0.036*** (-7.67)	-0.021*** (-4.73)	-0.040*** (-8.78)
School located in Southern Italy	-0.068*** (-8.62)	-0.103*** (-10.78)	-0.070*** (-6.25)	-0.082*** (-7.34)
School located in PON area	-0.049*** (-6.29)	0.021* (2.17)	-0.052*** (-4.85)	0.005 (0.47)
Istituto Comprensivo	0.001 (0.23)	0.012 (1.93)	0.003 (0.36)	0.004 (0.57)
Private school	-0.025*** (-6.53)	-0.015*** (-3.36)	-0.021*** (-4.05)	-0.016** (-2.94)
Variance in the ESCS index between classes (ESCS_Var_Within)	-0.024 (-1.16)	0.023 (0.85)	-0.050* (-2.20)	-0.019 (-0.77)
N	423,513	424,087	221,224	247,351
R ²	0.220	0.161	0.382	0.360
adj. R ²	0.220	0.161	0.382	0.360

Figure 1: The distribution of the variable of interest, *ESCS_Var_Within*

Panel A. (Kernel) density – distribution

Statistics of *ESCS_Var_Within*, by percentiles

Percentiles	Value of ESCS Var Within
1%	0.030
5%	0.637
10%	1.398
25%	3.291
50%	6.414
75%	11.278
90%	18.266
95%	24.044
99%	37.493
Mean	8.588
Std. Dev.	7.965

Empirical testing of probabilistic seismic hazard models

Dario Albarello¹, Vera D'Amico²

Abstract The problem of assessing relative reliability of different procedures for long term earthquake forecasting by semi-empirical approaches (probabilistic seismic hazard assessment) is addressed. It is argued that ex-ante procedures (logic-trees) commonly used on purpose and based on expert judgments present important limitations. An ex-post empirical testing procedure based on a coherent probabilistic formulation is proposed that allows to overcome these problems. Recent applications of these procedures in the Italian area have shown their actual feasibility.

1 Introduction

Long term (tens of years) earthquake forecasting (seismic hazard assessment) plays a basic role in the definition of effective strategies for seismic risk reduction. Since available knowledge about the seismogenic process are presently inadequate to select univocally future seismic scenarios, several situations are considered possible from the physical point of view. The major task of seismologists in this context is the assessment the likelihood level to be associated to each possible scenario (Probabilistic Seismic Hazard Assessment or PSHA). This is generally formalized in terms of distributions that associate at any possible level of the seismic ground motion (the seismic “scenario”) an exceedence probability (“seismic hazard”).

This assessment can be performed by considering different pieces of information both concerning deterministic aspects (seismic sources location and geometry, geodetic strain field, etc.) and statistical evidence (past seismic history, scale relationships, etc.). Available approaches (e.g., Bender and Perkins, 1987; Frankel, 1995; Woo, 1996; Pace et al., 2006; D'Amico and Albarello, 2008; Peresan et al., 2011) mainly differ for the

¹ Dario Albarello, Università degli Studi d Siena, dario.albarello@unisi.it

² Vera D'Amico, INGV, Sez. di Milano, vera.damico@pi.ingv.it

balance between deterministic and statistical evidence used in each case to evaluate the likelihood of the possible future seismic scenarios.

In this context, several alternative PSHAs resulting from alternative computational schemes (in terms of basic assumptions, use of available information, etc.) co-exist each “ex-ante” plausible and internally consistent but resulting in quite different hazard evaluations (see, e.g., Pace et al., 2011). To this multiplicity of alternative evaluations one must add the high sensitivity of some of these computational procedures to information characterized by high or not well defined uncertainty (e.g., geometry of seismogenic sources, etc.). In general, uncertainty relative to the choice among alternative models is defined as “epistemic” to distinguish it from the one (“aleatory”) related to the inherent randomness of the seismogenic process (SSHAC, 1997).

2 The “Logic-tree” approach

The presence of different PSHAs for the same area poses a number of problems to stake-holders responsible for political decisions and risk reduction strategies, that are forced to choice among “equivalent” hazard evaluations. To manage this situation, “logic-tree” approaches (e.g., SSHAC 1997) have been proposed to manage this kind of uncertainty (“epistemic”).

In this approach, any PSHA procedure is split into basic constitutive elements (seismicity rates assessment, attenuation relationships, etc.) are identified. Each of these elements constitute a node. At each of these nodes possible alternatives are considered and each of them constitutes a possible way to define a PSHA procedure. Any specific sequence of “nodes” and links (a “branch” of the logic “tree”) defines a specific procedure and identifies a possible hazard estimate. To each possible choice at any node a degree of belief is associated by considering expert judgements provided by a panel of experts. The degree of belief associated to each considered PSHA procedure is simply obtained by the product of the likelihoods along the considered branch.

This procedure is in principle appealing and apparently “objective” or “democratic” since allows the combination of different opinions. However, it leaves completely unresolved the problem of assessing each of this “opinions” that is completely “ex-ante” with respect the final outcome of the model and only depend on the degree of belief attributed to the opinions of each expert in the panel. Some critical aspects on this regards have been discussed by Krinitzky (1993).

Other problems for the logic-tree approach are inherent to the computational scheme adopted to evaluate considered PSHA procedures. This approach, in fact, implicitly assumes that all possible alternative are considered (the tree is complete) and that these are mutually exclusive (to warrant the applicability of the multiplication rule). Two possibilities actually exist. In one case, one considers incomplete relatively small “trees” that only account for a limited number of alternatives. This situation of course provide unsatisfactory results since it requires the agreement of all experts in excluding any alternative considered “ex-ante” totally unreliable and this is the source of endless discussions in the relevant scientific community. The other possibility is constructing enormous logic trees characterized by thousands of branches (e.g., Abrahamson et al., 2002) that may result actually unmanageable. In any case the

degrees of belief associated to each possible branch of the logic-tree is “de-facto” a probability distribution. However, since PSHA outcomes are on their turn probability distributions (e.g., describing the exceedance probability of any level of ground shaking), the ontological status of such combination of probability distributions is unclear. This triggered a debate about what the final outcome of the logic-tree approach should be (Bommer and Sherbaum, 2008).

Another problem is related to the idea, implicit in the logic-tree approach, that reliability of any PSHA estimate ultimately relies on the reliability of each constitutive element and not on the resulting outcomes. In fact, one can demonstrate (Rabinowitz and Steinberg 1991; Stirling and Petersen 2006) that a significant interplay exists between these elements and that the overall reliability is more than just a simple combination of single reliabilities relative to individual components.

3 An alternative approach

An alternative approach that can result formally coherent is based on the use of a Bayesian formulation where likelihoods to be attributed to each PSHA procedure is assessed as a whole.

A generic i -th procedure for PSHA is denoted by H_i . We assume that N of such procedures actually exist. Each hazard estimate relative to a ground shaking parameter g can be considered as a conditional probability in the form $P(g|H_i)$. In this context, the unconditional hazard estimate $P(g)$ can be given in the form

$$P(g) = \sum_{i=1}^N P(H_i)P(g|H_i) \quad (1)$$

where the probabilities $P(H_i)$ are the degrees of belief associated to the methodology H_i . It is worth to note that the distribution $P(g)$ preserves its proper meaning and includes epistemic uncertainty associated to the presence of a multiplicity of PSHA procedures.

In this view, the definition of the final outcome $P(g)$ of the N available PSHA procedures relies on the definition of the values to be attributed to $P(H_i)$. This task can be achieved by testing each PSHA model against actual seismic occurrences during a fixed control interval.

Testing aims at the definition of a level of reliability $P(H_i)$ to be associated to the model H_i , after that a set of occurrences e_s (e.g., the fact that during a “control” time interval at a number S of sites a given PGA threshold has been exceeded or not), has been examined. The set of occurrences e_s represents the control data set. In the epistemic view of probability (e.g., De Finetti, 1974), $P(H_i)$ is considered to represent the degree of belief in the model H_i .

The reliability $P(H_i)$ after that the set of S occurrences e_s is known, can be expressed in terms of the conditional probability $P(H_i|e_s)$. In this conditions, the Bayes theorem holds by stating that

$$P(H_i|e_s) = \frac{1}{P(e_s)} P^*(H_i) P(e_s|H_i) \quad (2)$$

In this formalization, $P(H_i|e_s)$ is the *ex-post* reliability evaluation of the PSHA model H_i and $P^*(H_i)$ is the corresponding *ex-ante* evaluation. The term $P(e_s|H_i)$ represents the Likelihood of the outcome e_s in the case that the H_i model is applied. This term actually

represents the probability that the model H_i associates to the “scenario” e_s : in other terms it represents the “forecast” of the model about that specific scenario. The term $1/P(e_s)$ is a simple normalization factor. In the case that M mutually exclusive competing PSHA models exist and that this set is complete, one has

$$P(e_s) = \sum_{i=1}^M P(H_i)P(e_s|H_i) \quad (3)$$

This formalization enlightens a basic aspect, i.e. the fundamental role played by the likelihood $P(e_s|H_i)$ to perform relative evaluations of the competing models. It also puts in evidence that *ex-ante* evaluations could play an important role but they are not sufficient to judge the relative feasibility of the model under study.

Actually, it is well known that Likelihood is of basic importance in the whole field of statistics (Edwards 1972) and its application in the field of seismic forecast evaluations is well established (e.g., Kagan and Jackson, 1994; Jackson, 1996; Schorlemmer *et al.* 2007; Kagan, 2009). In this view, the above likelihood evaluation becomes the “core” of a generalized PSHA procedure able to include epistemic uncertainty within a coherent frame.

4 Testing PSHA procedures

The basic tool for the definition of the likelihood required in the formulation here proposed that avoids drawbacks implicit in the logic-tree approach, is empirical testing. In practice, the required estimates are provided by comparing the relevant outcomes (e.g. the peak ground acceleration or PGA expected at any fixed exceedence probability) with observations available for a given control period. Whenever possible, these procedures should be performed in a “perspective” way, i.e., to concern observations not considered in the model parameterization. Since probabilistic models are of concern, each resulting in a probabilistic distribution associated to any ground motion parameter (hazard function), the assessment of the above mentioned likelihoods can be also seen as a typical “probability scoring” problem developed and applied in other contexts (e.g., Lind, 1996).

A direct procedure to test any PSHA model can be delineated as follows. Given the PSHA computational model H_i and the set of sites $E_{\Delta t^*}$ where ground shaking has been monitored during the control interval Δt^* , the model’s Likelihood L_i can be estimated from the control sample $E_{\Delta t^*}$. If the events e_s are mutually independent and if, over the duration of the control period, a total of N^* out of S sites have experienced ground shaking $>g_0$, then we have

$$P(e_s|H_i) = L_i = \left\{ \prod_{s=1}^{N^*} P(e_s|H_i) \right\} \left\{ \prod_{s=N^*+1}^S [1 - P(e_s|H_i)] \right\} \quad (4)$$

It is worth noting that the reliability of the hypothesis of mutual independence of the considered occurrences has to be evaluated in the frame of the considered PSHA model: $P(e_s|H_i)$ is a feature of the model H_i and not of the underlying seismogenic process.

The above approach has been recently applied for testing short term (e.g., the “L-test” is the terminology adopted in the RELM/CSEP projects (Schorlemmer and Gerstenberger, 2007; Schorlemmer *et al.*, 2007) and long term earthquake forecasting

(Albarelo and D'Amico, 2008) and has been shown to represent the most straightforward way for testing PSHA procedures. In the frame of the Bayesian view described in the previous section, it can also represent a basic tool for reliable hazard assessment in the presence of alternative testable PSHA procedures.

5 Conclusions

In the frame of a coherent Bayesian approach, the problem of providing seismic hazard estimates in the presence of alternative approaches can be defined that overcomes major drawbacks of the logic-tree approach commonly used on purpose. The basic element of the proposed approach is the empirical assessment of the degree of belief associated to each procedure in the form of a likelihood value. This is obtained by testing outcomes of the considered procedures (each considered as a whole) against seismic occurrences observed at a number of selected sites during a control interval. The actual feasibility of this last approach is testified by a recent applications in the Italian territory and elsewhere.

References

1. Abrahamson, N.A., Birkhauser, P., Koller, M., Mayer-Rosa, D., Smit, P., Sprecher, C., Tinic, S., Graf, R.: PEGASOS — a comprehensive probabilistic seismic hazard assessment for nuclear power plants in Switzerland. Proceedings of the Twelfth European Conference on Earthquake Engineering, London, Paper, vol. 633 (2002).
2. Albarelo D., D'Amico V.: Testing probabilistic seismic hazard estimates by comparison with observations: an example in Italy. *Geophys.J.Int.*, 175, 1088–1094 (2008) doi: 10.1111/j.1365-246X.2008.03928.x
3. Bender, B. and Perkins, D.M.: *SEISRISK III: a computer program for seismic hazard estimation*. USGS Bulletin 1772, 48 pp (1987)
4. Bommer J.J and Scherbaum F.: The Use and Misuse of Logic Trees in Probabilistic Seismic Hazard Analysis, *Earthq. Spectra*, Vol. 24, No. 4, pp. 997-1009 (2008), doi: 10.1193/1.2977755
5. D'Amico V., Albarelo D.: SASHA: a computer program to assess seismic hazard from intensity data. *Seism.Res.Lett.*, 79, 5, 663-671 (2008)
6. De Finetti B.: *Theory of probability*. Wiley, New York (1974)
7. Frankel, A.: Mapping seismic hazard in the Central and Eastern United States, *Seism. Res. Letts.*, v. 66, no. 4, pp. 8-21 (1995).
8. Edwards, A.W.F.: *Likelihood*. Cambridge Univ. Press, 235 pp. (1972)
9. Jackson, D.D.: Hypothesis testing and earthquake prediction. *Proc. Nat. Acad. Sci. USA*, vol. 93, 3772-3775 (1996)
10. Krinitzky, E.L.: Earthquake probability in engineering – part I: the use and misuse of expert opinion, *Eng. Geol.*, 33, 257-288 (1993).
11. Kagan Y.Y.: Testing long-term earthquake forecasts: likelihood methods and error diagrams. *Geophys.J.Int.*, (2009) doi.10.1111/j.1365-246X.2008.04064.x
12. Kagan, Y.Y. and Jackson, D.D.: Long-term probabilistic forecasting of earthquakes, *J. Geophys. Res.*, 99, 13685-13700 (1994)
13. Lind, N.C., Validation of probabilistic models, *Civ. Eng. Syst.*, 13, 175-183 (1996)

14. Pace B., Peruzza L., Lavecchia G., Boncio P.: Layered Seismogenic Source Model and Probabilistic Seismic-Hazard Analyses in Central Italy. *Bull Seism Soc Am* 96(1):107–132 (2006)
15. Pace B., Albarello D., Boncio P., Dolce M., Galli P., Messina P., Peruzza L., Sabetta, F., Sanò T., Visini F.: Predicted Ground Motion after the L'Aquila 2009 earthquake (Italy, Mw6.3): input spectra for Seismic Microzoning. *Bull. Earthquake Eng.*, 9, 199-230, (2011) DOI 10.1007/s10518-010-9238-y
16. Peresan A, Zuccolo E, Vaccari F, Gorshkov A, Panza GF.: Neo-Deterministic Seismic Hazard and Pattern Recognition Techniques: Time-Dependent Scenarios for North-Eastern Italy, *Pure Appl. Geophys.*, Vol. 168, nos. 3–4, pp. 583–607. (2011) doi 10.1007/s00024-010-0166-1
17. Rabinowitz, N. and Steinberg, D.M.: Seismic hazard sensitivity analysis: a multi-parameter approach, *Bull. Seism. Soc. Am.*, 81, 796-817 (1991)
18. Schorlemmer D., Gerstenberger, M. C., Wiemer, S., Jackson, D.D., Rhoades, D.A.: Earthquake Likelihood Model Testing, *Seism. Res. Lett.*, Vol. 78, No. 1, 17-29 (2007)
19. Schorlemmer, D., Gerstenberger, M.C.: RELM Testing Center, *Seism. Res. Lett.*, Vol. 78, No. 1, 30-36 (2007).
20. Senior Seismic Hazard Analysis Committee – SSHAC: Recommendations for probabilistic seismic hazard analysis: guidance on uncertainties and use of experts. Report NUREG-CR-6372, in 2 volumes, Washington D.C., U.S. Nuclear Regulatory Commission (1997)
21. Stirling, M. and Petersen, M.: Comparison of the historical record of earthquake hazard with seismic hazard models for New Zealand and the Continental United States, *Bull. Seism. Soc. Am.*, 96, 1978-1994 (2006)
22. Woo G.: Kernel estimation methods for seismic hazard area source modelling. *Bull. Seim.Soc.Am.* 86 (2): 353-362 (1996)
- 23.

A proposal for the multidimensional extension of CUB models

Federico Andreis and Pier Alda Ferrari

Abstract Particular emphasis has been put, lately, on the analysis of categorical data and many proposals have appeared, ranging from pure methodological contributions to more applicative ones. Among such proposals, the CUB class of distributions, a mixture model for the analysis of ordinal data that has been successfully employed in various fields, seems of particular interest. CUB are univariate models that do not possess, at present, a multivariate version: aim of the present work is to investigate the feasibility of building a higher-dimensional version of such models and its possible applications. In order to achieve such results, we propose to employ techniques typical of the framework of copula models, that have proven to be valid tools for multivariate models construction and data analysis.

Key words: multivariate ordinal data, CUB models, copula models, dependence structures

1 Introduction

The analysis of ordinal data is nowadays a field of great interest for the vast majority of applied fields and poses interesting challenges to statisticians in the development of an adequate methodology. Diverse proposals have been introduced during the recent years for their treatment, leading to important theoretical contributions from the scholars worldwide. Among such proposals, the authors deem worth of particular consideration the CUB [5] models, a class of univariate mixture distributions that has been successfully applied in many fields such as semiotics, ability assessment,

Federico Andreis
Università degli Studi di Milano, e-mail: federico.andreis@unimi.it

Pier Alda Ferrari
Università degli Studi di Milano, e-mail: pieralda.ferrari@unimi.it

medical research and customer satisfaction; the parsimonious parameterization and the ease of estimation and interpretation make CUB models a very useful tool for ordinal data analyses. Our proposal aims at extending such modeling approach to be able to handle multivariate data, the main reason for it being the belief that multivariate data coming from the same source (such as responses to a questionnaire items) should be treated as an ensemble, rather than split up, in order to account for (possibly) existing dependence structures in the estimation procedure, analysis and final results interpretation. Such extension is sought for in the framework of copulas [4], with the awareness of the problems arising from working with categorical, rather than continuous, data; a general structure is proposed and the use of different copula models is investigated.

2 Background

This section is intended to briefly review the general framework of both CUB and copula models.

2.1 CUB models

The CUB is a class of mixture models, possibly involving covariates, developed as a new approach for modeling discrete choices processes. The most common situation in which such approach can be employed regards the analysis of questionnaire data, with items responses evaluated on Likert scales and, thus, in the presence of ordinal data. The inherent *uncertainty* component is modeled through a discrete uniform variable, whereas the latent process leading to the choice, and governed by the subjective *feeling*, is modeled using a Shifted-Binomial distribution. The probability of observing a particular response r to an item, assuming that the number of item categories m is known and fixed, is expressed as a mixture of two such components as follows:

$$P(R = r) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi) \frac{1}{m}, \quad r = 1, 2, \dots, m \quad (1)$$

with $\pi \in (0, 1]$ and $\xi \in [0, 1]$.

π define the mixture weights and as such is inversely related to the amount of *uncertainty* in the answers (the higher π , the less the uniform component contributes to the mixture), i.e. "*each respondent acts with a **propensity** to adhere to a thoughtful and to a completely uncertain choice, measured by π and $(1-\pi)$, respectively*" [3]. ξ , on the other hand, is related to personal preferences and measures the strength of *feeling or adherence, agreement* with the item (the interpretation of ξ also depends on the kind of ordering adopted for the item).

2.2 Copula models

Copulas are n -place, grounded and n -increasing real functions with the unit hypercube as domain, that can be used to link univariate distribution functions (called margins) to form multivariate distribution functions according to arbitrary dependence structure (for reference see, for example, [4]). One of the main advantages of copulas is that they allow for separate specification of margins and dependence among them, where the dependence structure can be (and usually is) characterized by one or more parameters. Sklar's Theorem [6] is central to the theory of copulas: it provides the representation through a copula of a multivariate distribution function and grants its uniqueness when dealing with continuous margins. Such fact grants many useful properties that can be exploited for estimation and inferential purposes.

Particular care is needed when working with discrete margins, as in the case of the CUB models, due to the non-uniqueness issue, as stressed in [2]; non-uniqueness stems from the fact that marginal distribution functions are not strictly monotonically increasing, rather monotonically non-decreasing, and do not possess an inverse in the usual sense, rather a pseudo-inverse (see, for example, [4]). The most severe consequence of this is that it becomes impossible to draw general conclusions on the dependence structure binding the margins based on the copula parameterization alone (every result in this sense has been shown to be margin-dependent). Nonetheless, copulas still are an easy-to-implement and interesting tool to build multivariate models, and under certain circumstances it is still possible to make assessments about dependence among margins. For example, some copula families possess the property of being ordered by Positive Quadrant Dependence (PQD, see [2]): this grants minimal requirements to be met for copula parameters to be interpretable as dependence measures. We will therefore focus on such families in this work.

3 Proposal

As said, CUB models have been developed to describe univariate discrete phenomena, e. g. the distribution of answers to a single questionnaire item. Since questionnaires are usually composed by many different questions (say k), a complete analysis with CUB would require to separately estimate the k couples $(\pi_i, \xi_i), i = 1, \dots, k$, that characterize each item. This disjoint analysis approach does not take into account the dependence (possibly) existing among items, which could be exploited to better catch further information about the phenomenon and enrich its understanding. Drawing on this, we intend to evaluate the feasibility of a multivariate joint approach to CUB modeling, through the use of copula models. We thus define a multidimensional extension of CUB models, and call it CO-CUB model, as follows:

Definition 1. A k -dimensional ($k \geq 2$) CO-CUB model with copula C is a multivariate discrete variable with margins $R_i \sim CUB(\pi_i, \xi_i), i = 1, \dots, k$, each with support $\{1, \dots, m_i\}, m_i > 3$, and joint distribution function given by:

$$\begin{aligned}\Psi(r_1, \dots, r_k; \underline{\pi}, \underline{\xi}, \underline{\theta}) &= P(R_1 \leq r_1, \dots, R_k \leq r_k; \underline{\pi}, \underline{\xi}, \underline{\theta}) = \\ &= C_{\underline{\theta}}[F_1(r_1; \pi_1, \xi_1), \dots, F_k(r_k; \pi_k, \xi_k)]\end{aligned}\quad (2)$$

where $\underline{\pi} = (\pi_1, \dots, \pi_k)'$, $\underline{\xi} = (\xi_1, \dots, \xi_k)'$ and for a particular choice of copula C , characterized by a parameter $\underline{\theta} = (\theta_1, \dots, \theta_d)'$ taking values in some real d -dimensional space Θ defining the dependence structure of its components. $F_i(r_i) = F_i(r_i; \pi_i, \xi_i)$ stands for the distribution function of the i -th margin, i.e. $F_i(r_i) = P(R_i \leq r_i)$, and the support of the CO-CUB variable is the grid $\{1, \dots, m_1\} \times \dots \times \{1, \dots, m_k\}$. The whole parameter set for a k -dimensional CO-CUB is, then, the ordered triplet $(\underline{\pi}, \underline{\xi}, \underline{\theta}) \in (0, 1]^k \times [0, 1]^k \times \Theta$.

An interesting first attempt at defining a bivariate CUB distribution using the Plackett distribution is made in [1]. Our proposal further extends this approach to a more general framework, focusing on the comparison of different choices of the copula C to define the CO-CUB models, of which [1] is shown to be a special case; while developing a general method, we specifically compare, for the sake of illustration, the well known Clayton, Frank and Plackett copulas (whose families are ordered by PQD) in the simple bivariate case, discussing from a methodological point of view estimation-related issues and parameters interpretation, as well as feasibility of extension to more than $k = 2$ dimensions.

By definition of copula, margins of (2) are all CUBs, whose parameters retain, then, the same interpretation as in the unidimensional case, while for what concerns the copula parameter θ , its interpretation as a dependence measure will be a further subject of studying: a first idea is to use it to rank couples of items by strength of dependence, when in presence of an ordered family of copulas (previously described). This might be useful for questionnaire calibration purposes and to individuate latent structures.

Acknowledgements The authors acknowledge financial support from the European Social Fund Grant (FSE), Lombardy Region.

References

1. Corduas, M.: Modelling correlated bivariate ordinal data with CUB marginals. *Quaderni di Statistica* **13**, 109–119 (2011)
2. Genest, C. and Neslehova, J.: A Primer on Copulas for Count Data. *ASTIN Bulletin* vol.37 no.2 (2007). <http://www.actuaries.org/LIBRARY/ASTIN/vol37no2/475.pdf>
3. Iannario, M. and Piccolo, D.: A program in R for CUB models inference (Version 2.0), (2009) available at: <http://www.dipstat.unina.it>
4. Nelsen, R.B.: *An Introduction to Copulas*. Springer (2010)
5. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* **5**, 85–104 (2003)
6. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231 (1959)

Data fusion in the framework of coupled matrix tensor factorization with common, partially common and unique factors

Morten Arendt Rasmussen, Evrim Acar
University of Copenhagen - Denmark

Abstract

In practical research, data are often collected from various sources with the aim of estimating the relational structure between different blocks of data. Multiblock methodology provides a framework for extraction of such underlying structures in data. Up to now the methodology has primarily focused on the extraction of common underlying patterns across all data blocks. However, typically the individual data blocks consist of *unique* variation, variation common between *some* blocks and variation common between *all* blocks. Here we devise a method based on Coupled Matrix and Tensor Factorisation (CMTF) for estimation of common, partially common and unique components with the possibility of imposing various constraints (sparsity, non negativity and orthogonality) on the individual modes. The proposed approach is not restricted to 2-way data sets and easily extends to combinations of multiple matrices and multi-way arrays. We demonstrate the usefulness of the proposed approach on a three- block data set, i.e., two 2-way and one 3-way data sets, from a clinical cohort study examining different biomarker development profiles during intervention in relation to baseline characteristics and selected genetical profiles. The data is from a well described cohort of 122 newly onset type I diabetic children and adolescents.

Estimation of Sparse Generalized Linear Models: the `dglars` package

Luigi Augugliaro and Angelo M. Mineo

Abstract `dglars` is a public available R package that implements the method proposed in Augugliaro, Mineo and Wit (2013) developed to study the sparse structure of a generalized linear model. This method, called dgLARS, is based on a differential geometrical extension of the least angle regression method (LARS). The core of the `dglars` package consists of two algorithms implemented in Fortran 90 to efficiently compute the solution curve; specifically a predictor-corrector algorithm and a cyclic coordinate descent algorithm.

Key words: generalized linear models, dgLARS, predictor-corrector algorithm, cyclic coordinate descent algorithm, sparse models, variable selection

1 Introduction

Nowadays, high-dimensional data sets, namely data sets where the number of predictors, say p , is larger than the sample size N , are becoming more and more common. Modern statistical methods developed to study this kind of data sets are usually based on the idea to use a penalty function to estimate a solution curve embedded in the parameter space and then to find the point that represents the best compromise between sparsity and predictive behaviour of the model. Recent statistical literature has a great number of contributions devoted to this problem: some important examples are the L_1 -penalty function [8], the SCAD method [5] and the MC+ penalty function [9], among others.

Luigi Augugliaro
University of Palermo, Viale delle Scienze Ed. 13 e-mail: luigi.augugliaro@unipa.it

Angelo M. Mineo
University of Palermo, Viale delle Scienze Ed. 13 e-mail: angelo.mineo@unipa.it

Differently from the methods cited above, in [3] the authors propose a new approach based on the differential geometrical representation of a GLM. The derived method, that does not require an explicit penalty function, has been called differential geometric LARS (dgLARS) method because it is defined generalizing the geometrical ideas on which the least angle regression (LARS), proposed in [4], is based. Using the differential geometric characterization of the classical signed Rao score test statistic, dgLARS gains important theoretical properties that are not shared by other methods. From a computational point of view, the dgLARS method consists essentially in the computation of the implicitly defined solution curve. In [3] this problem is satisfactorily solved by using a predictor-corrector (PC) algorithm. In [2] is proposed a much more efficient cyclic coordinate descend (ccd) algorithm to fit the dgLARS solution curve when we work with an high-dimensional data set.

In this paper we present the `dglars` package that implements both the algorithms to compute the solution curve implicitly defined by dgLARS. The object returned by these functions is a S3 class object, for which specific methods and functions have been implemented. The package `dglars` is available under general public licence (GPL-2) from the Comprehensive R Archive Network at <http://CRAN.R-project.org/package=dglars>.

2 The `dglars` package

The `dglars` package is an R [6] package containing a collection of tools related to the dgLARS method. In the following of this section we describe the main functions available with this package. For a complete description of the all functions implemented in the `dglars` package the reader is referred to the manual of the package.

The function `dglars()` is the main function of the proposed package. It can be called with the following arguments

```
dglars(X, y, family = c("binomial", "poisson"), control = list())
```

where `X` is the design matrix of dimension $n \times p$, `y` is the n -dimensional response vector and `family` is the error distribution used in the model. Finally `control` is a named list of control parameters. For a complete description of this list, the interested reader is referred to the corresponding help page.

To gain more insight on how to use the `dglars()` function we consider a simulated data set. We simulate a data set from a logistic regression model with sample size equal to 100 and $p = 5$ predictors. We also assume that only the first 2 predictors influence the response variable. The used R code is

```
R> n <- 100; p <- 4; s <- 2; X <- matrix(rnorm(n * p), n, p)
R> bs <- rep(1, s); Xs <- X[, 1 : s]
R> eta <- drop(1 + drop(Xs %*% bs))
```

```
R> mu <- binomial()$linkinv(eta); y <- rbinom(n, 1, mu)
R> out_dglasso_pc <- dglars(X = X, y = y, family = "binomial")
```

`dglars()` returns a S3 class object

```
R> class(out_dglasso_pc)
[1] "dglars"
```

As shown in the following R code, the method function `print.dglars()` can be used to print out the basic information contained in a `dglars` object.

```
R> out_dglasso_pc
```

```
Call: dglars(X = X, y = y, family = "binomial")
```

Sequence	g	Dev	%Dev	df
+x1	3.67566	134.6	0.00000	1
	3.06853	130.5	0.03080	2
+x2	3.04937	130.3	0.03171	2
	0.21800	109.0	0.19047	3
+x4	0.20859	109.0	0.19055	3
	0.05396	108.8	0.19188	4
+x3	0.03199	108.8	0.19194	4
	0.00010	108.8	0.19198	5

Algorithm pc (method = dgLASSO) with exit = 0

The column `Sequence` shows that the dgLARS method first finds the true predictors and then includes the other false predictors. The column `g` reports the value of the parameter γ used in the dgLARS method to select the trade-off between sparsity of the estimated model and prediction behaviour [3]. To be more specific, at the starting step the predictor `x1` makes the smallest angle with the tangent residual vector and then is included in the active set. The predictor `x2` is included in the active set at $\gamma^{(2)} = 3.04937$, this means that $\hat{\beta}_2(\gamma^{(2)})$ is equal to zero and then the number of non-zero estimated coefficients is equal to 2. The predictor `x4` is included at $\gamma^{(3)} = 0.20859$ and so on.

More information about the estimated sequence of models can be obtained using the method function `summary.dglars()`. The output printed out by `summary.dglars()` is divided in three different sections. The first section completes the basic information printed out by `print.default()` showing the sequence of the Akaike Information Criterion (AIC) [1] and the sequence of the Bayesian Information Criterion (BIC) [7]. The ranking of the estimated models obtained by these two criteria are also showed and the corresponding best model is identified by an arrow. The last two sections show the estimated coefficients of the two best models. For the sake of brevity, the following R code shows only the first section printed out by `summary.dglars()`:

```
R> summary(out_dglasso_pc)
```

```
Call: dglars(X = X, y = y, family = "binomial")
```

Sequence	g	Dev	Complexity	AIC	Rank.AIC	Rank.BIC	BIC
+x1	3.67566	134.6	1	136.6	8	6	139.2
	3.06853	130.5	2	134.5	7	8	139.7
+x2	3.04937	130.3	2	134.3	6	7	139.5
	0.21800	109.0	3	115.0	2	2	122.8
+x4	0.20859	109.0	3	115.0	1 <-	-> 1	122.8
	0.05396	108.8	4	116.8	4	4	127.2
+x3	0.03199	108.8	4	116.8	3	3	127.2
	0.00010	108.8	5	118.8	5	5	131.8

3 Conclusion

In this paper we have described the R package `dglars`. This package implements the differential geometric extension of the LARS method proposed in [3] and called dgLARS. The use of this package is shown by means of a simulated data set. The output of the functions are presented in a way that is easy to interpret for people familiar with standard `lm`, `glm` or `gam` output.

References

1. Akaike H.: Information Theory as an Extension of the Maximum Likelihood Principle. In BN Petrov, F Czaki (eds.), Second International Symposium on Information Theory, pp. 267–281 (1973). Akademiai Kiado, Budapest.
2. Augugliaro L, Mineo AM, Wit E.C.: Differential Geometric LARS Via Cyclic Coordinate Descent Method. In *Proceedings of COMPSTAT 2012*, pp. 67–79 (2012). Limassol, Cyprus.
3. Augugliaro L, Mineo AM, Wit E.C.: Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models. *J. R. Statist. Soc. B* (2013).
4. Efron B, Hastie T, Johnstone I, Tibshirani R.: Least Angle Regression. *Ann. Statist.* **32**(2), 407–499 (2004).
5. Fan J, Li R.: Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Statist. Ass.* **96**(456), 1348–1360 (2001).
6. R Development Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>.
7. Schwarz G.: Estimating the Dimension of a Model. *Ann. Statist.* **6**(2), 461–464 (1978).
8. Tibshirani R.: Regression Shrinkage and Selection Via the Lasso. *J. R. Statist. Soc. B* **58**(1), 267–288 (1996).
9. Zhang C.H.: Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Ann. Statist.* **38**(2), 894–942 (2010).

A comparison between two tools for data stream summarization

Antonio Balzanella and Lidia Rivoli and Elvira Romano

Abstract Data stream summarization is receiving a lot of attention due to the growth of real world applications generating huge streams of data. This paper focuses on two tools which have been recently proposed in this framework: the functional boxplot and the Quantile-functions boxplot. We at first, describes the two tools and then we propose a careful comparison in order to highlight their advantages and drawbacks.

Key words: Data stream mining, Functional Boxplot, Quantile-function boxplot

1 Introduction

Data stream mining aims at discovering knowledge from huge temporally ordered flows of data which can be read only once using limited computing and storage capabilities. They emerge in a lot of applicative fields involving the monitoring of physical quantities such as electricity consumptions, climate variables and sensor data in general. Due to the constraints imposed by the online nature of data, summarization is a key topic. The use of summaries allows to keep track of the information in data, without storing them, as well as to discover anomalous behaviors or evolutions in the monitored phenomenon. A wide literature is available on this topic [3]. Most of methods extend the computation of traditional statistics to data streams in order to support the incremental learning. A more recent trend is to develop complex

Antonio Balzanella
Second University of Naples- Caserta - Italy e-mail: antonio.balzanella@gmail.com

Lidia Rivoli
Second University of Naples - Caserta - Italy e-mail: lidia.rivoli@unina.it

Elvira Romano
Second University of Naples - Caserta - Italy e-mail: elvira.romano@unina.it

summaries which are able to collect more information than the traditional statistics mentioned above [1], [2],[4]. This paper focuses on comparing two of these complex summaries: the functional boxplot and the quantile-functions boxplot.

In the next sections we describe, shortly, the two tools and then we introduce the details of our comparison highlighting their features, advantages and drawbacks.

2 Functional Boxplot vs Quantile-function boxplot

In this section we discuss about the features of the two methods highlighting the differences, the advantages and the drawbacks.

The first topic is to analyze the data given as input to the two methods. In both the cases, it is a set of sequences S of temporally ordered observations which can come from a single data stream or from multiple streams.

Formally, for the single data stream case, let $Y_1 = [y_1, \dots, y_l, \dots]$ be an univariate data stream whose observations $y_l, l = 1, 2, \dots$ arrive continuously at fixed time stamps t_l . The data stream Y can be split into non overlapping windows $W_i = [t_j, \dots, t_{j+ws}]$ ($i = 1, 2, \dots$) of equal size ws so that for each window W_i we get a subsequence $Y_1^{W_i} = [y_j, \dots, y_{j+ws}]$. In this univariate case, the input of the two methods will be a set of temporally consecutive subsequences $S = [Y_1^{W_m}, \dots, Y_1^{W_{m+k}}]$ where k indicates the desired number of processed windows.

If the analysis is performed on multiple streams, the methods can process a set $Y = [Y_1, \dots, Y_n]$ in which the streams are still split into non overlapping windows W_i ($i = 1, 2, \dots$). Unlike to the previous case, a window frames a set of subsequences so that we can get a Functional Boxplot or a Quantile-function boxplot for each window. In this sense, the output of the methods is a set of summaries which keeps track of the streams behavior over time.

Before comparing the two tools, we shortly introduce them.

A Functional Boxplot FBP_j is a compound of five functions obtained by processing the set $S[5]$.

$$\{f_{[u]}(t), f_{[l]}(t), f_{[1]}(t), f_{[b_{min}]}(t), f_{[b_{max}]}(t)\} \quad (1)$$

where: $f_{[u]}(t)$ is the upper bound of the central region; $f_{[l]}(t)$ is the lower bound of the central region; $f_{[1]}(t)$ is the median curve; $f_{[b_{min}]}(t)$ is the upper bound of the subsequences; $f_{[b_{max}]}(t)$ is the lower bound of the subsequences

The Functional Boxplot is the analog of classical boxplot. The main difference consists in the data ordering criterion. In particular, since functions varies over a continuum, data ordering is based on the notion of band depth or modified band depth [5].

Based on the center outward ordering induced by band depth for functional data, the descriptive statistics of a functional boxplot are: the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope. The 50% central region is the analog to the "interquartile range" (IQR), it is defined by the band delimited by the 50% of deepest, or the most central observations. The border of

the 50% central region is defined as the envelope representing the box in a classical boxplot. The median is the most central observation in the box. The maximum envelope of the dataset identified by the vertical lines of the plot are the "whiskers" of the boxplot. Formally, let $f_{[i]}(t)$ denote the sample of functional subsequence associated to the i th largest band depth value. The set $f_{[1]}(t) \dots, f_{[n]}(t)$ are order statistics, with $f_{[1]}(t)$ the median curve, that is the most central curve (the deepest), and $f_{[n]}(t)$ is the most outlying curve. The central region of the boxplot is defined as

$$C_{0.5} = \left\{ (t, f(t)) : \min_{r=1, \dots, [n/2]} f_{[r]}(t) \leq f(t) \leq \max_{r=1, \dots, [n/2]} f_{[r]}(t) \right\} \quad (2)$$

where $[n/2]$ is the small integer not less than $n/2$.

As before, we here describe the Quantile-functions Boxplot. Unlike to the Functional Boxplot, the Quantile-functions Boxplot needs a preprocessing step on each sequence of S . It is the computation of a set $\{H_i\}_{i=1, \dots, N}$ of N histograms, where each $H_i = \{(I_{i1}, f_{i1}), \dots, (I_{ik}, f_{ik}), \dots, (I_{iK_i}, f_{iK_i})\}$, $i = 1, 2, \dots$, whose I_{ik} are the bins and f_{ik} are the relative frequencies associated to I_{ik} , $\forall k = 1, \dots, K_i$. Within each interval $I_{ik} = [y_{ik}, \bar{y}_{ik})$, it is assumed that the values are uniformly distributed, so the quantile function F_i^{-1} associated to each H_i is a piecewise linear function.

We want to introduce a box plot for the set of the quantile functions F_i^{-1} , $i = 1, \dots, N$. Thus, the definitions of the *Median*, the *First* and the *Third Quartile* and the *whiskers* quantile functions are necessary. Let $\mathbf{w} = \{w_1, \dots, w_l, \dots, w_m\}$, with $w_1 = 0$, $w_m = 1$ and $\max_{1 \leq i \leq N} K_i \leq m \leq \sum_{i=1}^N K_i - 1$ be a suitable set of common cumulative relative frequencies for all histograms [6]. The Median quantile function F_{Me}^{-1} is the piecewise linear function assuming as values the set of the m medians Me over the ordered set $\{F_{(1)}^{-1}(w_l), \dots, F_{(N)}^{-1}(w_l)\}$ for all $l = 1, \dots, m$, i.e. on the value $F_{(\frac{N+1}{2})}^{-1}(w_l)$ if N is odds; otherwise, on the averages of $F_{(\frac{N}{2})}^{-1}(w_l)$ and $F_{(\frac{N}{2}+1)}^{-1}(w_l)$ for all $l = 1, \dots, m$.

The First quantile function $F_{Q_1}^{-1}$ is the piecewise linear function whose values are the m first quartiles (Q_1) over the ordered set $\{F_{(1)}^{-1}(w_l), \dots, F_{(N)}^{-1}(w_l)\}$ for all $l = 1, \dots, m$. Similarly, the Third Quartile quantile function $F_{Q_3}^{-1}$ is a piecewise linear function whose values contains the m third quartiles (Q_3) over the ordered set $\{F_{(1)}^{-1}(w_l), \dots, F_{(N)}^{-1}(w_l)\}$ for all $l = 1, \dots, m$.

Thus, the *box* can be defined as the region bounded by the quantile functions associated to the First and Third Quartile-histograms and the *Inter Quartiles Range (IQR)* is the area between $F_{Q_1}^{-1}$ and $F_{Q_3}^{-1}$ computed by L^1 Wasserstein distance [6].

Finally, for the choice of the whiskers it is proposed to determine them by a point-wise method. In particular, the functions $F_{Q_1}^{-1}(w_l) - 1.5 \cdot IQR$ and $F_{Q_3}^{-1}(w_l) + 1.5 \cdot IQR$ for all $l = 1, \dots, m$ correspond to the Lower and Upper Whisker quantile functions respectively.

We, now, analyze the main differences between the described tools. A first important difference between the two methods is the way they process the sequences in S before to compute the two summaries. In Functional boxplots, the sequences are used without some preprocessing so that the it keeps the time ordering in the

computation of the five statistics. In the Quantile-functions boxplot, the temporal sequences are processed by extracting, at first, the corresponding histogram and then the associated piecewise quantile function. This approach is consistent with the most of proposals in data stream mining literature which assume that a data stream is made by concepts changing over time and which define the concept as an empirical distribution. The drawback in this case is the lost of time information inside the subsequences. A further difference concerns the criterion used for the ordering of input data. The functional boxplot is based on depth functions in which the ordering is performed by measuring the depth of each input sequence. In this sense the median is the sequence having the highest depth. The quantile-functions boxplot is based on finding an optimal solution to an optimization problem. With this aim the authors build the median (as well as the other order statistics) as a piece-wise function whose components are the medians computed on intervals of the quantile functions. We can summarize the consequences of the two approaches as follows: 1) In the functional boxplot the median, so as the other order statistics, is an observed sequence while in the quantile-functions boxplot it is build from data; 2) In the functional boxplot the median, so as the other order statistics, can intersect the other sequences having lower or higher depth while in the quantile-functions boxplot the median is always completely inside the other quantile functions without any intersection; 3) In functional boxplot, two medians could be discovered since two functions could have the same degree of depth where quantile-functions boxplot always provide a single median.

All the highlighted differences are confirmed by the application of these tools on real data.

References

1. Balzanella, A., Romano, E., Verde, R. Summarizing and Mining Streaming Data via a Functional Data Approach. In *Classification and Multivariate Analysis for Complex Data Structures*, 409–416. Ed. Springer (2011).
2. Balzanella, A. and Romano, E.: A Clustream strategy for Functional Boxplots on multiple streaming time series. In *Proceedings 46th Scientific Meeting of the Italian Statistical Society, Roma (2012)*. CLEUP. ISBN 978-88-6129-882-8 (2012). (extended version: <http://arxiv.org/abs/1212.2784>)
3. Gama, J., Gaber, M. M.: *Learning from Data Stream. Techniques in Sensor Networks*. Ed. Springer Verlag (2007).
4. Rivoli L., Irpino A., Verde R.: The median of a set of histogram data. In: *XLVI Riunione Scientifica della Societ Italiana di Statistica*, CLEUP. ISBN 978-88-6129-882-8 (2012).
5. Sun Y., Genton M.G.: Functional boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316-334. (2011).
6. Verde, R., Irpino, A.: Dynamic clustering of histogram data: using the right metric. *Studies in Classification, Data Analysis, and Knowledge Organization 2007 Part I*, **12**, 123–134, (2007) doi: 10.1007/978-3-540-73560-1.

Gini Index Estimation in Randomized Response Surveys

Lucio Barabesi and Giancarlo Diana and Pier Francesco Perri

Abstract We address the problem of estimating the Gini index when data are assumed to be collected through the randomized response method proposed by Greenberg et al. (1971). In the design-based framework, we treat the Gini index as a population functional and follow the approach proposed by Deville (1999) to obtain the corresponding estimator. Variance estimation is also considered.

Key words: Income, influence function, population functional estimation, sensitive questions

1 Introduction

Surveys on sensitive or highly personal issues such as drug taking, tax evasion, illegal income and so on, are likely to meet with refusal to cooperate (*unit-non-response*), refusal to answer specific questions (*item-non-response*) or untruthful answers (*measurement error*), especially when direct questions are posed to participants. These sources of nonsampling errors can deteriorate the data quality and thus jeopardize the usefulness of the data for both researchers and policy makers. Although these errors cannot be totally avoided, they may be limited by using questioning techniques which increase respondent cooperation. The Randomized Response (RR) technique introduced by Warner (1965) can be used to achieve this. In fact, the mechanism is based on a randomization device which determines the

Lucio Barabesi
Department of Economics and Statistics, University of Siena, e-mail: lucio.barabesi@unisi.it

Giancarlo Diana
Department of Statistical Sciences, University of Padova, e-mail: diana@stat.unipd.it

Pier Francesco Perri
Department of Economics, Statistics and Finance, University of Calabria,
e-mail: pierfrancesco.perri@unical.it

question to answer to. Since the respondent does not reveal to anyone the outcome of the device, his/her true status remains uncertain and privacy is protected.

In this note we attempt to investigate how the RR approach could be used to alleviate problems arising from collecting data on income which is notoriously considered a sensitive character to be surveyed in the sense that people are reluctant to disclose it, mostly in the case of income from self-employment, property and financial assets. In fact, it usually happens that the rich tend to understate it so as to reduce tax liability, while it may happen that the poor tend to overstate their income for sense of shame or, to a lesser extent, that taxpayers declare slightly more than they earn to avoid controls. Consequently, this may result in seriously biased analyses and estimates of welfare indicators such as the Gini index.

In short, we shall give an idea on how respondent privacy can be protected by using a simple randomization device, hence we analyze how the device can affect the estimation of the Gini index. In so doing, we first define the estimator of the index and of the corresponding variance under a randomization step, then analyze the accuracy of the estimators through a limited simulation study carried out on income data from the *Survey of Household Income and Wealth* (SHIW) conducted by Bank of Italy (2010).

2 Gini Index Estimation

Let $U = \{1, \dots, N\}$ be a fixed population of N identifiable units and let us suppose that a quantitative variable assumes value y_i on the i th unit. Let $\theta = \theta(y_1, \dots, y_N)$ be a population parameter to be estimated on the basis of a random sample s of fixed size n selected from U in accordance with a given sampling design. Let $\pi_i > 0$ denote the first-order inclusion probability.

Finding estimators of θ and obtaining the corresponding variance estimator is a crucial issue in sampling practice. In this paper, we consider the approach proposed by Deville (1999) which is based on the concept of statistical functionals and may be summarized in three steps: (1) the definition of a measure induced by a non-decreasing and non-negative function; (2) the introduction of a corresponding empirical measure; (3) the definition of a suitable plug-in estimator.

Let us consider the population Gini index usually expressed in the form (see, e.g., Berger, 2008)

$$G = \frac{2 \sum_{i \in U} \sum_{j \in U} y_i I_{[y_j, \infty)}(y_i)}{N \sum_{i \in U} y_i} - 1,$$

with I denoting the indicator function. Let us suppose that the randomization procedure proposed by Greenberg et al. (1971) is adopted to perturb the true response y_i in such a way to ensure confidentiality to the respondents. Under this protocol, the i th individual reports his/her true income value y_i with probability q , or he/she generates a random variate from a suitable distribution function H with probability $(1 - q)$. Note that $q = 1$ leads to direct questioning. Obviously, the true values y_i

are unrecognizable on the sampled individuals once the randomization procedure is performed. As a consequence, standard estimation procedures for G are precluded. In order to achieve an estimator of G when the randomization stage is added, let the random variable Z_i represent the answer of the i th individual and let us consider the measure induced by the non-decreasing and non-negative function $M_R : \mathbb{R} \mapsto \mathbb{R}^+$ given by

$$M_R(y) = \sum_{i \in U} F_{Z_i}(y) = qM(y) + (1 - q)NH(y),$$

where $M(y) = \sum_{i \in U} I_{[y_i, \infty)}(y)$, while

$$F_{Z_i}(z) = qI_{[y_i, \infty)}(z) + (1 - q)H(z)$$

is the distribution function of Z_i . Hence, it is possible to reformulate G as a functional of $M_R(y)$ and then to achieve the plug-in estimator \widehat{G} by considering the empirical counterpart of M_R given by $\widehat{M}_R(y) = \sum_{i \in s} I_{[Z_i, \infty)}(y) / \pi_i$. Variance estimation can be obtained by following Deville's (1999) generalized linearization method based on the influence function and extending the Result 3 in Langel and Tillé (2013).

Technical aspects for both Gini index estimation and variance estimation are intentionally skipped since too detailed to be adequately discussed in this short note.

3 A Simulation Study

In order to evaluate the performance of \widehat{G} and of the corresponding variance estimator in a RR set-up, we have carried out a Monte Carlo study based on real data from the SHIW 2010. We assume the 7951 households of the survey as the target population and consider the ideal situation in which each family participates in the survey and declares its *household net disposal income* truthfully. Hence, for this situation we have $G = 0.347$. We assume this value as the benchmark throughout the study. Beside this situation, we suppose that all the households take part in the survey though some of them may release untruthful data. Hence, families are instructed to carry out Greenberg et al.'s (1971) method and, without breaking the rules laid down by the inquirer, provide the value of the random variable Z_i with distribution function H . With regard to the choice of H , we basically considered two alternatives in order to show how an inappropriate choice can have a negative impact on the reliability of the estimation process: (i) the Uniform distribution in the interval $(0, \max(y_1, \dots, y_{7,951}))$; (ii) the Dagum model whose parameters have been estimated on the income data from SHIW 2008.

For the study, $K = 5000$ samples of size $n = 250, 1000$ are drawn from the population according to SRSWOR. The bias of \widehat{G} has been evaluated by means of $\text{PRB}_1 = 100 \times (G_{\text{sim}} - G) / G$ where $G_{\text{sim}} = \sum_{k=1}^K \widehat{G}_{(k)} / K$, $\widehat{G}_{(k)}$ being the estimate of G in the k th sample. To assess the accuracy of the variance estimator of \widehat{G} , we first computed $\text{var}_{\text{in}}[\widehat{G}] = \sum_{k=1}^K \widehat{\text{var}}_{(k)}[\widehat{G}] / K$, where $\widehat{\text{var}}_{(k)}[\widehat{G}]$ is the lin-

Table 1 Simulation results under the Uniform and Dagum distribution H

n	q	Uniform					Dagum				
		G_{sim}	var_{sim}	var_{lin}	PRB ₁	PRB ₂	G_{sim}	var_{sim}	var_{lin}	PRB ₁	PRB ₂
250	1.0	0.349	3.6E-04	3.4E-04	0.57	-4.1E+00	0.349	3.6E-04	3.4E-04	0.57	-4.09
	0.8	0.327	3.0E-02	7.3E-02	-5.55	1.4E+02	0.348	5.8E-04	5.3E-04	0.32	-8.05
	0.6	0.403	5.0E-02	3.9E+01	16.26	7.7E+04	0.345	9.4E-04	9.0E-04	-0.47	-3.96
1000	1.0	0.347	8.2E-05	8.2E-05	0.14	-2.2E-01	0.347	8.2E-05	8.2E-05	0.14	-0.22
	0.8	0.356	7.9E-03	8.1E-03	2.59	3.1E+00	0.348	1.1E-04	1.1E-04	0.28	1.54
	0.6	0.274	2.6E-02	6.9E-02	-21.04	1.7E+02	0.345	2.4E-04	2.3E-04	-0.62	-1.94

earization variance estimate computed in the k th sample. Hence, we compared $\text{var}_{\text{lin}}[\widehat{G}]$ with the Monte Carlo variance of \widehat{G} , say $\text{var}_{\text{sim}}[\widehat{G}]$, by means of $\text{PRB}_2 = 100 \times (\text{var}_{\text{lin}}[\widehat{G}] - \text{var}_{\text{sim}}[\widehat{G}]) / \text{var}_{\text{sim}}[\widehat{G}]$.

The simulation results, partially shown in Table 1 for $q = 0.6, 0.8, 1.0$, point out that estimation carried out by assuming the Uniform distribution is highly unstable and unsatisfactory. Moreover, even if not reported here, we have found many realizations of \widehat{G} lying outside the admissible interval. This is hardly surprising since the Uniform distribution does not fit income data at all. On the contrary, satisfactory results may be obtained when the randomizing variable is chosen wisely as for the Dagum distribution. In this case, the results seem to be even competitive for modest sample sizes when compared with those from direct questioning ($q = 1$). This fact seems worth remarking since it can represent a response to some of the criticisms directed at the RR theory and can contribute to its reappraisal as a tool for data collection. Obviously, our study does not claim to change the standard way to collect income data. It is rather an attempt to estimate a complex population parameter when privacy disclosure may become a serious problem for the researcher. Nonetheless, a question we would raise in conclusion: Keeping in mind the *gigo principle* (garbage in, garbage out), is it better to perform studies on flawed data or on “true” data which are perturbed following a rule prescribed by the researcher?

References

1. Bank of Italy (2010) Survey of Household Income and Wealth. Available at <http://www.bancaditalia.it/statistiche/indcamp/bilfait/dismicro>
2. Berger, Y.G.: A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, **24**, 541–555 (2008)
3. Deville, J.C.: Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203 (1999)
4. Greenberg, B.G., Kubler, R.R., Abernathy, J.R., Horvitz, D.G.: Applications of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, **66**, 243–250 (1971)
5. Langel, M., Tillé, Y.: Variance estimation of the Gini index: revisiting results several times published. *Journal of the Royal Statistical Society A*, **176**, 521–540 (2013)
6. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63–69 (1965)

A test for time-invariant individual effects in generalized linear models for panel data

Francesco Bartolucci, Federico Belotti and Franco Peracchi

Abstract We propose a computationally convenient test for the null hypothesis of time-invariant individual effects in generalized linear models for panel data. The proposed test is based on a Hausman-like statistic comparing fixed-effects estimators defined as the maximand of full and pairwise conditional likelihood functions. Thus, no assumptions are required on the distribution of the individual effects and their correlation with the covariates. We investigate the finite sample properties of the test through a set of Monte Carlo experiments. Finally, an example from the health economics literature is used to illustrate the proposed test.

Key words: Generalized linear models; Hausman-type test; Health and Retirement Study; Individual effects; Longitudinal data; Self-reported health

1 Introduction

A distinctive feature of panel data modeling is the treatment of unobserved heterogeneity, which is typically interpreted as the effect of unobservable factors on the outcome of interest. The simplest way of dealing with this form of heterogeneity is to include in the model time-invariant individual effects; for a detailed treatment see [1], [6], and [9]. However, assuming that these effects are constant over time may be difficult to justify, especially in the case of long panels.

A few studies have recently tried to relax the assumption of time-invariant individual effects by modeling unobserved heterogeneity as a unit-specific time-series process; see, for instance, [3] and [5]. Because these approaches are valid only under strong assumptions and may be computationally demanding, practitioners may find

Francesco Bartolucci
University of Perugia, e-mail: bart@stat.unipg.it

Federico Belotti
University of Rome Tor Vergata, e-mail: federico.belotti@uniroma2.it

Franco Peracchi
University of Rome Tor Vergata and EIEF, e-mail: franco.peracchi@uniroma2.it

it useful to carry out a preliminary test for the presence of time-invariant unobserved heterogeneity before estimating this type of model.

In order to test for the null hypothesis of time-invariant individual effects in generalized linear models (GLMs) for panel data, we propose to compare alternative estimators obtained by maximizing full and pairwise conditional likelihood functions. Unlike the standard version of the Hausman test [4], we compare estimators that are both inconsistent under the alternative. It is worth emphasizing that the proposed test does not require assumptions on the distribution of unobserved heterogeneity; also, it does not require assumptions on how time-invariant regressors enter the model. Moreover, it can be easily implemented using standard statistical software, as the computation of the test statistic only requires a quadratic form which involves the difference of the parameter estimates and a consistent estimator of its asymptotic variance matrix.

In the following we first outline the statistical framework and the proposed test (Section 2). We then summarize the results of a simulation study aimed at investigating its finite sample properties and an empirical illustration based on data from the Health and Retirement Study (Section 3). A detailed description of the proposed approach can be found in [2].

2 The proposed approach

We assume that the data consist of a balanced panel where n units, randomly drawn from a given population, are observed for T periods each. For every unit $i = 1, \dots, n$, we denote by $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ the vector of observed outcomes and by \mathbf{X}_i the $T \times k$ matrix of observed covariates, with t th row denoted by \mathbf{x}_{it} .

Under the null hypothesis of time-invariant unobserved heterogeneity, our model is a GLM with canonical link function [7]; using a standard notation, the conditional density of the model is of the form

$$f(y_{it} | \alpha_i, \mathbf{X}_i) = f(y_{it} | \alpha_i, \mathbf{x}_{it}) = \exp \left[\frac{y_{it} \eta_{it} - b(\eta_{it})}{\gamma} + c(y_{it}, \gamma) \right],$$

with $\eta_{it} = \alpha_i + \mathbf{x}_{it}' \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the $k \times 1$ vector of parameters of interest and α_i is an individual effect, γ is a known dispersion parameter, and $b(\cdot)$ and $c(\cdot, \gamma)$ are known functions.

If α_i is time-invariant and the covariates are strictly exogenous conditional on α_i , then inference about $\boldsymbol{\beta}$ may be based on the conditional density of the data given a sufficient statistic for the time-invariant individual effect, such as $y_{i+} = \sum_{t=1}^T y_{it}$. This conditional density depends only on $\boldsymbol{\beta}$, not on α_i . The resulting maximum likelihood estimator of $\boldsymbol{\beta}$, called the full conditional maximum likelihood estimator and denoted by $\hat{\boldsymbol{\beta}}_1$, maximizes the conditional log-likelihood function $L_1(\boldsymbol{\beta}) = \sum_{i=1}^n \ln f(\mathbf{y}_i | \mathbf{X}_i, y_{i+})$. In order to construct a Hausman-like specification test, we need an alternative estimator of $\boldsymbol{\beta}$ that is also consistent if the

unit-specific effects are time-invariant but has different convergence properties if they are time-varying. For this aim, we consider the pairwise conditional maximum likelihood estimator, denoted by $\hat{\boldsymbol{\beta}}_2$, which is obtained from the maximization of $L_2(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{t=2}^T \log f(y_{i,t-1}, y_{it} | \mathbf{x}_{i,t-1}, \mathbf{x}_{it}, y_{i,t-1} + y_{it})$.

It is easily shown that, under the null hypothesis of time-invariant individual effects, $\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_2 \xrightarrow{p} \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$. Thus, under the null hypothesis, both estimators are consistent and have a joint asymptotically normal distribution with variance matrix \mathbf{W}_0 . Under the alternative hypothesis of time-varying individual effects, neither $\hat{\boldsymbol{\beta}}_1$ nor $\hat{\boldsymbol{\beta}}_2$ are consistent for $\boldsymbol{\beta}$ in general. Further, being based on different functions of the data when $T > 2$, they will generally converge to different points in the parameter space, being $\hat{\boldsymbol{\beta}}_2$ more robust to violations of the assumption of time-invariant unobserved heterogeneity than $\hat{\boldsymbol{\beta}}_1$ (see [8] and [10]).

The above results suggest a test that rejects the null hypothesis of time-invariant unobserved heterogeneity for large values of the statistic

$$\hat{\xi} = n(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)' \widehat{\mathbf{V}}_0^{-1} (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2), \quad (1)$$

where $\widehat{\mathbf{V}}_0$ is a suitable transformation of a consistent estimate $\widehat{\mathbf{W}}_0$ of \mathbf{W}_0 . If the asymptotic variance matrix \mathbf{V}_0 is nonsingular, then the test statistic $\hat{\xi}$ exists with probability approaching one for large values of n , and its asymptotic null distribution is χ^2 with number of degrees of freedom equal to k . If the asymptotic variance matrix is singular and one replaces the inverse of $\widehat{\mathbf{V}}_0$ in (1) with a generalized inverse, then the asymptotic null distribution of $\hat{\xi}$ is still χ^2 with number of degrees of freedom equal to the rank of \mathbf{V}_0 , which in this case is less than k .

In [2], this approach is studied in detail for four commonly used GLMs: (i) logit regression, (ii) ordered logit regression, (iii) Poisson regression, and (iv) the Gaussian linear model. Note that the ordered logit model is estimated by considering all possible dichotomizations $y_{it}^{(j)}$ of the ordinal outcome y_{it} for each unit in the sample, where $y_{it}^{(j)} = 1\{y_{it} > j - 1\}$, $j = 1, \dots, J - 1$. For each of these dichotomizations, the conditional log-likelihood of a logit model is defined and all these functions are summed up to define the overall log-likelihood that is maximized by a standard Newton-Raphson algorithm.

3 Monte Carlo study and application

To investigate the finite sample properties of the proposed test, we performed a series of simulations, the results of which are discussed in detail in [2]. For each of the considered GLMs (i)–(iv), we adopted a data generating process in which individual-specific effects follow an AR(1) process, parameterized as in [5], with different values of the autocorrelation parameter ρ . It is worth noting that, under the chosen parametrization, $\rho = 1$ corresponds to the null hypothesis of time-invariant

unobserved heterogeneity, whereas values of ρ smaller than 1 corresponds to violations of this hypothesis. We also studied the case of correlation between these individual effects and the observable covariate. To evaluate the effect of an increase of the sample size and of the number of time occasions, different values of n and T have been considered.

The simulation results confirm that the proposed test attains the nominal size even in small samples. As for its power, we find that it rapidly increases with both n and T . Not surprisingly, the test loses power for values of ρ close to zero, as in this case the individual effects are confounded with the idiosyncratic error.

The proposed testing procedure is illustrated in [2] through an empirical application to self-rated health status (SRHS) of the elderly U.S. population using data from the Health and Retirement Study, a longitudinal survey that interviews every two years a representative sample of over 26,000 people aged 50 and older. We focus our attention on the subset of $n = 4,094$ individuals who responded to all waves. SRHS is measured on a 5-point ordered scale (poor, fair, good, very good, excellent). The covariates include a set of socio-demographic characteristics (gender, age, education and ethnicity), the number of doctor visits, and the body mass index (BMI).

We consider two model specifications. The first includes as regressors a constant term, an age spline, BMI and the number of doctor visits. The second adds to the first a set of wave dummies. Our key result is that, regardless of the model type (we estimate both logit and ordered logit models) and specification, we strongly reject the null hypothesis of time-invariant unobserved heterogeneity. Given this result, we estimate the latent AR(1) random-effects logit and ordered logit models proposed by [5], obtaining a statistically significant estimate of the autocorrelation coefficients ρ (close to 0.95).

References

1. Arellano, M., Bonhomme, S.: Nonlinear panel data analysis. *Annual Review of Economics* **3**, 395–424 (2011)
2. Bartolucci, F., Belotti, F., Peracchi, F.: Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data. EIEF Working paper **12** (2013)
3. Bartolucci, F., Farcomeni, A.: A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association* **104**, 816–831 (2009)
4. Hausman, J.A.: Specification tests in econometrics. *Econometrica* **46**, 1251–1271 (1978)
5. Heiss, F.: Sequential numerical integration in nonlinear state space models for microeconomic panel data. *Journal of Applied Econometrics*, **23**, 373–389 (2008)
6. Hsiao, C.: *Analysis of Panel Data*. Cambridge University Press (2005)
7. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*, 2nd Edition. Chapman and Hall (1989)
8. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42 (2011)
9. Wooldridge, J.M.: *Econometric Analysis of Cross Section and Panel Data*. MIT press (2010)
10. Xu, X., Reid, N.: On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, **141**, 3047–3054 (2012)

Identification of the distribution of the causal effect of an intervention using a generalised factor model

Erich Battistin, Università di Padova

Carlos Lamarche, University of Kentucky

Enrico Rettore, Università di Padova

Abstract.

The literature on the identification of the causal effect of a policy intervention on an outcome of interest took off since the early 70's. A common feature of many empirical contributions to this literature is the stress placed on the role of heterogeneous returns to participation in the intervention. Despite this emphasis on heterogeneity causal inference has been mostly concerned with identification of the *average* effect of the intervention. In this setting, the role played by heterogeneity is at best investigated by comparing average returns for different groups in the population, just ignoring within group variability in returns. Since the seminal work by Heckman *et al.* (1997), many authors have discussed the theoretical and empirical relevance of understanding the conditions required to learn about the *distribution* of the causal effect. The approach considered in the literature we take in what follows postulates a factor structure that relates potential outcomes (see, for example, Carneiro *et al.*, 2003, Aakvik *et al.*, 2005, Heckman *et al.*, 2006, Cunha and Heckman, 2007, and Heckman *et al.*, 2011). In its bare essentials, the key assumption required for identification is that the dependence across potential outcomes is solely generated by a low dimensional set of variables. We show that moving from a generalized factor model allowing for heteroskedastic uniqueness the implied distribution of the causal effect is considerably more flexible than that implied by the standard factor model. A simple GMM type estimator allows to obtain estimates for the parametric component of the model.

Internal effectiveness of educational offer and students' satisfaction: a SEM approach

Matilde Bini and Lucio Masserini

Abstract The aim of this paper is to explain the relationship between university students' satisfaction with their study experience career and the perception of a poor effectiveness of the organizational aspects of the teaching activities. This relationship can be modelled with a structural equation approach. The analysis is performed using a data set collected by a CATI survey on students enrolled at the University of Pisa in the academic year 2010-'11.

1 Introduction

In Italy, the need to improve the performance of the universities' course programs is a relevant issue. To satisfy this requirement, it is important to modify and make more efficient the organization and contents of the teaching activities, as well as to offer adequate services to students.

Deep changes occurred over the last twenty years introduced new evaluation systems in the Italian universities useful to improve performance of the teaching and research activities. A better performance could make students more satisfied about their study experiences, thus improving the acquired knowledge and their university careers as a whole. Hence, more efficient degree courses may attract motivated students and even public and private financial resources.

People involved in university activities interpret the term "quality" of educational process in different ways and this meaning changes depending to the context where each activity is carried out. It is well know that university system provides a service that can be judged not only from potential students but also from other stake-holders (such as academic and administrative personnel, etc..).

¹Matilde Bini, Department of Human Sciences, European University of Rome; email: mbini@unier.it

Lucio Masserini, Statistical Observatory, University of Pisa; email: l.masserini@adm.unipi.it

Two important features of the university performance are the assessment of how resources are employed to get expected results (efficiency analysis) and the qualitative assessment of the results and the level of achievement of objectives (effectiveness analysis). Internal effectiveness of the degree courses is strictly linked to the quality of teaching activity and organizational aspects, that may be measured by the students' satisfaction. In this paper a structural equation model is applied to explain the relationship between students' satisfaction and the perception of a poor effectiveness of the organizational aspects of the teaching activities.

The analysis is performed using a data set collected by a survey on the students' career for students enrolled at the University of Pisa in the academic year 2010-11. A stratified simple random sample of 1945 students was selected from the target population of 51758 enrolled students in the academic year 2010-'11 (Masserini and Pratesi, 2013). The allocation of the students into the strata was proportional to the population size and data were collected using a Computer Assisted Telephone Interviewing system. The performed analysis is limited to the 1,371 students enrolled in the first cycle degree courses.

2 Measuring the relationship between internal effectiveness and students' satisfaction

Students's satisfaction (S) is a latent variable measured with three observed indicators concerning overall satisfaction about university experience (y_1), satisfaction with academic achievements in career (y_2) and satisfaction with expectations at the time of the enrollment (y_3). Internal effectiveness of the educational offer can be evaluated for the quality of teaching, for the services and for the organization. Here, the analysis is focused on the organizational aspects measured with a latent variable using the following observed indicators: poor organization of the teaching activities (y_4), difficulties in getting learning materials (y_5), difficulties in getting reception hours (y_6) and difficulties in getting information about classes (y_7). The previous observed indicators define a measure of a *poor internal effectiveness* (E).

The relationship between students' satisfaction and poor internal effectiveness is not unique: satisfaction for the university experience can also be affected by how much students are capable to organize their studies, to undertake networks with other students and to have abilities to take information about courses. More specifically, the former characteristics define a latent variable named *student organization* (O) which is measured by the following three observed indicators: ability to organize time for attending classes (y_8), ability to prepare exams (y_9) and ability to plan studies (y_{10}). The latter characteristics define a latent variable named *students' relationships* (R) which is measured by the following indicators: ability to undertake relations with students (y_{11}), ability to study with other students (y_{12}), ability to get class materials (y_{13}), ability to acquire information about characteristics of degree courses (y_{14}) and ability to know administrative aspects (y_{15}). All the observed indicators are binary responses and represent students' assessments about several aspects about their university experiences, where 0= poor/low assessment and 1 = good/very good assessment.

Some other observed variables are included into the model as explanatory of the latent variables, defining a more complex system of relationships between students'

satisfaction and internal effectiveness: enrolment motivated by interests for courses (x_1), inactivity status (x_2), internships experience (x_3), repeat of years during high school (x_4) and having a job (x_5).

The system of relationships depicted by the previous latent and observed variables are modeled with a structural equation model (SEM). SEM is a multivariate technique used to test complex relationships among observed (measured) and unobserved (latent) variables as well as relationships between two or more latent variables. A SEM model is characterized by two components: a *structural* model, designed to explain the relationships among the latent variables and among the latent and the observed variables, and a *measurement* model, to explain the relationships among the latent variables and the observed indicators. The structural model can be expressed by the following equation (Muthén, 1984):

$$\eta = \beta\eta + \Gamma x + \zeta,$$

where η is an $m \times 1$ vector of endogenous latent variables; β is an $m \times m$ matrix for the endogenous latent variables; Γ is an $m \times n$ matrix of regression coefficients among the latent variables and the observed variables; ξ is an $n \times 1$ vector of exogenous latent variables; ζ is an $m \times 1$ vector of errors.

The measurement model is defined as follows:

$$y = \Lambda\eta + \varepsilon,$$

where y is a $p \times 1$ vector of the observed indicators; Λ is a $p \times m$ matrix of factor loadings and ε is $p \times 1$ vector of residuals. In presence of observed binary or categorical indicators, the conventional measurement model for continuous indicators is constructed as specified in Muthén (1984), by defining an underlying normally distributed latent variable for the corresponding observed variable. Here, the latent responses are linked to observed categorical responses via threshold models, yielding probit measurement models. The estimation of the structural parameters is performed with a three-stage limited-information procedure as described in Muthén (1984) and in Muthén and Satorra (1996), using a weighted least-squares fit function.

3 Main results

The analysis has been accomplished using the package *Mplus* 5.21 (Muthén, 2004). The main results from model estimation are shown in Figure 1. Poor internal effectiveness (E) has a direct effect (-0.346) on students' satisfaction (S) and it means that in situations characterized by a lower effectiveness of the educational offer, the level of satisfaction decreases. But the perception of a poor internal effectiveness can be affected by several characteristics of students, as evidenced in the figure by the latent variable student relationships (R). The results show that students who are capable to establish networks and other related abilities or skills, have a better perception of the internal effectiveness (-0.661). Also, having networks produce a positive effect on the students' organization (0.651), as could be expected.

Moreover, a direct effect on students' satisfaction (0.685) is exercised by Student organization (O).

As regards observed variables, only enrolment motivated by interests for course and inactivity status affect students' satisfaction (S); in fact, inactive students are less satisfied (-0.210) while students enrolled with interests on subjects pertaining the undertaken degree courses are more satisfied (0.134). Moreover, being inactive implies less networks (-0.096). Surprisingly, the internships experience has a negative effect (0.275) on poor internal effectiveness (E) and probably it means that these activities are perceived as inadequate or poorly organized; on the contrary, having an internships experience gives a positive effect on student organization (0.302). Finally, having a job (-0.226) and repeating years during high school (-0.214) negatively affect the student organization (O).

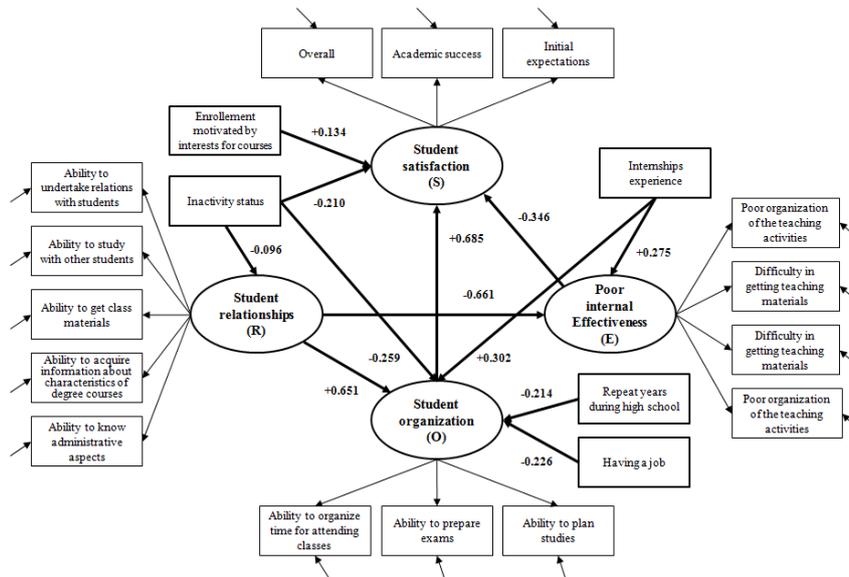


Figure 1: Relationships among internal effectiveness, students' satisfaction, students' organization and students' relationships

References

1. Masserini, L. and Pratesi, M. (2013). A sample survey on inactive students: weighting issues in modeling the inactivity status. *Advances in Theoretical and Applied Statistics*, forthcoming.
2. Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*. 49, 115-132.
3. Muthén, B. and Satorra, A. (1996). Technical aspects of Muthén's LISOCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*. 60, 489-503.
4. Muthén, B. (2004). *MPLUS Statistical Analysis With Latent Variables. Technical Appendices*, <http://www.statmodel.com/download/techappen.pdf>.

Groups heterogeneity and sectors concentration: a structural equation modeling for micro level analysis of firms

Matilde Bini*, Leopoldo Nascia*, Alessandro Zeli[□]

Abstract This paper is focused on the study of firms' growth strategy, with particular aim at verifying the relationship between external and internal factors affecting firm's growth strategy. The analysis refers to the Italian firms belonging to a group having a specific internal skill mix structure and different market structures in terms of market concentration and a heterogeneity of the economic sectors covered by the group distribution. Because of the difficulty in finding empirical growth indicator, the firm's growth strategy is considered as a measure of performance not directly observable but defined through economic concepts. To carry out it, the use of a structural equation approach is proposed.

1 Introduction

Competitiveness and growth strategy of firms are characteristics of firms' performance which cannot be measured using a straightforward and direct indicator. In fact, in the literature there is not a unique definition of them, but indeed they are associated to several aspects of firms' activities such as costs, degree of quality, level of differentiation, share of market or market segmentation and scale economy, or firm's

* Corresponding author Matilde Bini, Università Europea di Roma; email: mbini@unier.it

* Leopoldo Nascia, Università Europea di Roma; email: nascia@istat.it

[□] Alessandro Zeli, Servizio Studi econometrici e previsioni economiche, ISTAT; email: zeli@istat.it

behaviors in different contexts. The multidimensionality of these characteristics makes also problematic the detection of an empirical measurement. The most widely used indicators of competitiveness and growth strategy are the efficiency, productivity and efficacy. The analysis of the competitiveness and growth strategy involves the study of several factors which affect them, and above all, it is important to consider the relationship between the firm and the outside environment and the skills embodied in the internal firm governance. In fact, they are not disjointed: the rates and direction of learning are shaped by the internal structure and internal norms of behavior of individual organizations (Dosi, 1992). These ones can be viewed as interacting in a dynamic system where competences and internal governance structure co-evolve with the external environment. Nonetheless, organizational change, as well as skill and competencies mix, are highly dependent by firm history, and a study of the market structure where firms operate have to be considered, in particular the level of concentration of the market, i.e. if firms belong to a group and to the economic and productive heterogeneity within groups of firms. Moreover, it is necessary to consider the evolution of firms' labour structure which is defined for example by skill mix, average cost of personnel. Some authors (Onida, 2003) detect several causes of the increasing weakness of the large industry in Italy, that makes it less capable to provide dynamic and rewarding employment to young high-school and graduates, especially those having technical and scientific degrees, to offer higher education and training of new managerial and supervisory cadres, as well as to overcome the lag of the multinational expansion accumulated in the 1990's. This paper is focused on the study of firms' growth strategy, with particular aim at verifying the relationship between external and internal factors affecting it. The performed analysis refers to the Italian firms belonging to a group having a specific internal skill mix structure and different market structures in terms of market concentration and a heterogeneity of the economic sectors covered by the group distribution. Because of the difficulty in finding empirical growth strategy indicator, we consider the firms' growth strategy as a measure of the performance not directly observable, but rather defined through economic concepts. To carry out it the use of a structural equation approach is proposed.

2 The database and the SEM approach to the analysis

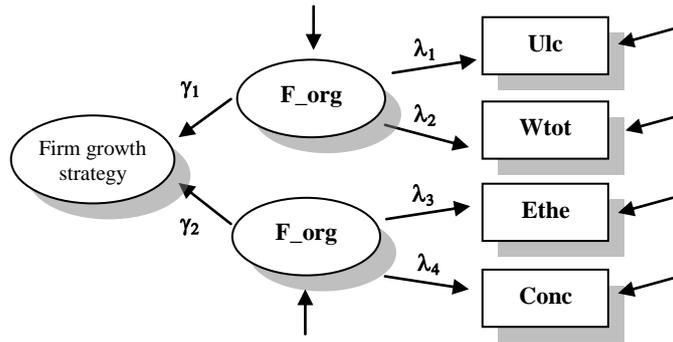
In this paper indirect indicators of firms' growth strategy are detected and modeled using a structural equation model following Linear Structural Relationship (LISREL) approach (Jöreskog and Sörbom, 1996). A theory describing the behavior of economic entities consists in two groups of relationships: the first connecting the constructs that represent the theory, and the second linking the constructs to measurement variables. The relationship between constructs and indicators (measurement variables) is central in validating constructs (latent variables). The measure of a construct validity can be focused on relationship between construct and measures (indicators) without considering the relationship with other constructs (trait validity) or it can be focused on theoretical relationships within the leading theory describing these interactions between different constructs (nomological validity). Hence, the theory can be represented by a

structural equation model. It is well known that LISREL model is characterized by two components: a structural model, designed to explain the relationships between latent variables, and a measurement model for evaluating the variations detected on the variables observed. The usual notation of LISREL model without endogenous variables is expressed by:

$$\eta = \Gamma \xi + \zeta \quad \text{structural model}$$

$$x = \Lambda_x \xi + \varepsilon \quad \text{exogenous measurement model}$$

where y are exogenous vectors of observed variables, η is the vector of the underlying latent variables, Γ , Λ_x are the matrices of the coefficients and ε and ζ are the terms of error of the measurement section (for further details, see Bollen, 1989). In this study three latent variables are considered in the structural model: the external market structure (labeled F_env) and internal firm organization (labeled F_org) as influencing the third variable, the firms' growth strategy. The variables to estimate external and organizational factors come from the archive of survey on firms' economic accounts combined with the Business group register of ISTAT, and they are defined respectively as follows: concentration index (Conc) which is the turnover share of the first three enterprises in the Nace sector; group heterogeneity according to economic activity (Ethe); white collar on total employees ratio (Wtot); personnel costs on employees (Ulc). The model we consider can be presented in the following scheme:



The parameters γ are related to the structural part of the model and estimate the influence of external and internal factors mix on firms behavior. The latent variables are linked to manifest variables by means the parameters λ (measurement model parameters). If they are significant imply that the correspondent exogenous variable is a measure of the construct (trait validity).

3 Main results

Because of lack of space, main results from model estimation are showed. The goodness of fit statistics are displayed below (Table 1). Chi-Square Test of Model Fit

yielded a value of 5.8202, with 3 degree of freedom and a p-value of 0.12 that is not significant at 0.1, this result demonstrates that the model has a good fit.

Table 1: Fit Summary

Chi-Square	5.820
Chi-Square DF	3
Pr > Chi-Square	0.120
Goodness of Fit Index (GFI)	0.995
Adjusted GFI (AGFI)	0.985
RMSEA Estimate	0.037
Standardized Root Mean Square Residual (SRMSR)	0.024
Bentler Comparative Fit Index	0.959

The other tests show a good fit of the model: CFI and AGFI indices are respectively 0.96 and 0.985; RMSEA estimate is 0.038 and SRMR index is 0.025.

Table 2 shows the parameters estimation, all parameter estimates are high significant and, as predicted, a positive relationship between the measurement variable and the exogenous ones. The trait validity is so verified.

Table 3 presents the regression parameters estimates between exogenous variables. They are significant, so it well represents the relationship between the internal and external factors. The firm strategy and behavior depends on two others latent variables and, anyway, to a greater extent on internal organization.

Table 2 : Regression Coefficients between the latent and the observed variables

Latent variables and indicators	Est	Std err	t Value	R ²
Internal Organization (F_{org}) respect to				
Personnel costs per capita	0.427	0.041	10.41	0.18
White collar share on total employment	0.6298	0.0258	24.38	0.40
External environment(F_{env}) respect to				
Group heterogeneity	0.3257	0.0605	5.39	0.11
Economic sector concentration	0.3089	0.0616	5.01	0.10

Table 3 : Standardized Results for Linear Equation (dependent=Firm strategy)

Latent variables	Est	Std err	t Value
Internal Organization	0.6078	0.093	6.54
External environment	0.3469	0.0311	11.16

References

1. Bini, M., Bertaccini, B., Masserini, L.: Italian PHD courses System: Evaluating External Effectiveness by Structural Equation Models, forthcoming *IJAS* (2013)
2. Bollen, K.A.: *Structural equations with latent variables*, Wiley, New York (1989)
3. Dosi, G.: Industrial organization, competitiveness and growth, *Revue d'économie industrielle*. Vol. 59, 1^{er} trimestre, 27-45 (1992)
4. Jöreskog, K.G., Sörbom, D.: *LISREL 8: User's reference guide*. Chicago: Scientific Software (1996)
5. Onida, F.: *Growth, competitiveness and firm size: factors shaping the role of Italy's productive system in the world arena*, CESPRI WP n.144 (July) (2003)

Use of Relevant Principal Components to Define a Simplified Multivariate Test Procedure of Optimal Clustering

Giuseppe Boari, Marta Nai Ruscone

Abstract Clustering is the problem of partitioning data into a finite number, k , of homogeneous and separate groups, called clusters. A good choice of k is essential for obtaining meaningful clusters. The intraclass correlation coefficient ρ is frequently used to measure the degree of intragroup resemblance (for example of characteristics such as blood pressure, weight and height). The theory concerning ρ is well established for single variables analysis (Sheffè, 1959; Rao, 1973). In this paper, this task is addressed by means of a multiple test procedure defining the optimal cluster solution under normality assumption of the involved variables. Relevant principal components are used to define a simplified multivariate test of null intraclass correlation procedure and the proposal of a new statistical stopping rule is evaluated.

Key words: Principal components, cluster analysis, intra class correlation, union intersection principle

1 Introduction

Clustering is the problem of dividing data into a finite number of relevant classes, so that items in the same group are as similar as possible, and items in different groups are dissimilar as possible (Duda et al., 2000). To this purpose the unsupervised learning technique, called k -means clustering method, has been widely used in a variety of areas. For an integer $k \geq 1$ the k -means consists in partitioning a random sample X in k groups by minimizing a sort of empirical distortion. In this context,

Giuseppe Boari

Department of Statistics, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milan ,
e-mail: giuseppe.boari@unicatt.it

Marta Nai Ruscone

Department of Statistics, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milan
e-mail: marta.nairuscone@unicatt.it

an essential problem is the selection of the right number, k , of clusters. Indeed, if in some situations the choice of k may be motivated by the applications, it is in general unknown. A presentation of various procedures for choosing k can be found in Milligan and Cooper (1985) and Hardy (1996) while Gordon (1999) compares the performances of the best five rules exposed in Milligan and Cooper (1985). These methods can be divided in two main types: global or local. Global procedures consist in performing clustering for different values of k and then retaining the value minimizing or maximizing some target function. In local procedures, it must be decided at each step whether a cluster should be partitioned. In our work we consider a local procedure; the resulting groups should exhibit high internal (within clusters) homogeneity and high external (between clusters) heterogeneity. However, the standard clustering algorithms do not suggest a definitely optimal solution and alternative heuristic criteria are, in general, available. For example, one may consider a sort of loss deriving from the internal dissimilarity of the identified groups. As usually underlined, the observations within a group are differently correlated with respect to the other clusters and variable could be correlated. However, a typical clustering problem has a multivariate reference structure, derived by a multivariate data set, generally assumed to be normally distributed. In many studies, especially in business and economics, the primary objective is to reproduce underlying structure by applying a clustering strategy. However the methodology is conceptually straight-forward; computers algorithms are readily available and they purportedly find good, but not necessary optimal partitioning.

2 The Choice of k : Model, Hypothesis Test and Stopping Rules

First recall the case of a single variable normally distributed, describing one specific character X used for the classification of the units. The observation are assumed to be scattered into J groups, of size n_j and is also assumed that each observation x_{ij} ($i = 1, \dots, n_j; j = 1, \dots, J$) consists of a sum of two independent random contributions:

$$x_{ij} = v_j + u_{ij} , \quad (1)$$

where u_{ij} represents independent observations from normally distributed variates U_{ij} , such that $E(U_{ij}) = 0$ and $V(U_{ij}) = \sigma_e^2$, and the v_j are random deviations from normally distributed variates with zero mean and variance $V(V_j) = \sigma_a^2$. Under the previous assumptions the ICC is defined as:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} , \quad (2)$$

where σ_a^2 corresponds to the between groups variance and σ_e^2 the within groups variance. The intraclass correlation is a measure of the effect of data clustering: if $\rho = 0$, then $\sigma_a^2 = 0$, and there is no clustering; if $\rho = 1$, then $\sigma_e^2 = 0$, and there is "complete clustering" in the sense that there is no within-cluster variability. The

object of the inferential approach is to test the hypothesis $\rho > 0$. It can be shown (Fisher, 1954) that the following statistic is appropriate:

$$F = \frac{N-J}{J-1} \frac{(1-\rho)}{\{1+(n-1)\rho\}} \frac{n \sum (x_{.j} - x_{..})^2}{\sum \sum (x_{ij} - x_{.j})^2} = \frac{(1-\rho)}{1+(n-1)\rho} F^*, \quad (3)$$

where the usual notation $N = \sum_j n_j$, $x_{.j} = \frac{\sum_i x_{ij}}{n_j}$ and $x_{..} = \frac{\sum \sum x_{ij}}{N}$ is used and F^* is distributed as the usual F distribution with $(J-1)$ and $(N-J)$ degrees of freedom (Commenges, Jacquin, 1994). In order to introduce the multivariate counterpart, we consider a classification procedure based on K variates, and propose a new procedure based on the union-intersection principle, suggested by Roy (1953) as a heuristic method of test construction for the multivariate problem. As previously, we assume that the observed units are scattered into J groups, but we further assume that is defined by the finest common group classification obtained by each considered character $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K$. For each specific number J of groups, the n_j observations of the generic group j ($j = 1, \dots, J$) are assumed from a K dimensional random variable $N(\mu_j, \Sigma)$, i. e. with the same variance matrix among the groups and possibly different mean vector. Considering the observations on a single character they are so scattered in the same way described by (1). Some of these variables could be correlated. In cluster analysis, the correlation between variables may mask the true group structure. A common objective in exploratory multivariate analysis is to identify a subset of the variables which conveys the main features of the whole sample. For the selection of the best uncorrelated subset of variables, we suggest the use of principal component analysis. Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of uncorrelated variables. Furthermore, this transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible, under the constraint of incorrelation with the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed. The union intercept principle methods, suggested by Roy (1957), may be involved to join the single tests (3), performed on the single \mathbf{X}_k of the form $H_{k0} : \rho = 0$. Let θ be the (vector valued) parameter for which the following test is proposed

$$H_{K0} : \theta \in \bigcup_{k=1}^K \Theta_k \text{ versus } H_{K\alpha} : \theta \in \bigcap_{k=1}^K \Theta_k^c. \quad (4)$$

The associated union intersection test considers the statistic:

$$\bar{p}_J = 1 - (1 - P(F_1 < F_1^*, F_2 < F_2^*, \dots, F_k < F_k^* | H_{k0} : \rho = 0)) = 1 - \prod_{k=1}^K (1 - p_k), \quad (5)$$

where p_J is the associated p-value of the single test on the $k - th$ variable and \bar{p}_J the significance level at which the null hypothesis can be rejected at least for one variable.

The smaller \bar{p}_J , the more strongly the test rejects the null hypothesis. We compare all \bar{p}_J (obtained by performing the test for different values of J): the optimal number of clusters correspond to the classification linked to the minimum \bar{p}_J , among the range of grouping previously selected.

However, like most of the clustering procedures based on assigning unit to the closest group, the previously described procedure tends to overestimate the optimal number of clusters. This is clearly due to the fact that units are always allocated in separated groups. In order to overcome this drawback we suggest to adopt a statistical stopping rule based on the Jarque-Bera test (Jarque, Bera, 1987, Mineo et al., 2004) of the null hypothesis that the sample comes from a single normal distribution with unknown mean and variance, against the alternative that it comes from a mixture of normal distributions. The test statistic is the following:

$$JB = \frac{n}{6} \left(s^2 + \frac{(k-3)^2}{4} \right), \quad (6)$$

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. A simulation study was also performed in order to evaluate the validity of our procedure being the real number of clusters J known in advance.

References

1. Commenges, D., Jacquin, H.: The Intraclass Correlation Coefficient: Distribution-Free Definition and Test. *Biometrics*. **50**, 517–526 (1994)
2. Duda, R. O., Hart, P. E., Stork, D. G.: *Pattern classification*. Wiley-Interscience, New York (2000)
3. Gordon, A. D.: *Classification. Monographs on Statistics and Applied Probability vol 82*. Chapman and Hall/CRC, Boca Raton (1999)
4. Hardy, A.: On the number of clusters. *Computational Data Analysis* **23**, 83–96 (1996)
5. Jarque, C. M., Bera, A. K.: A test for normality of observations and regression residuals. *International Statistical Review*. **55**, 163–172 (1987)
6. Milligan, G. W., Cooper, M. C.: An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. **50**, 159–179 (1985)
7. Mineo, A., Chiodi, M., Bock, H.H.: *Advances in Multivariate Data Analysis*. Springer, 2004.
8. Rao, C. R.: *Linear Statistical Inference and its Application*. Wiley, New York (1973)
9. Sheffè, H.: *The Analysis of Variance*. Wiley, New York (1959)

Some Distance Proposals for Cluster Analysis in Presence of Ordinal Variables

Giuseppe Boari, Gabriele Cantaluppi, Angelo Zanella

Abstract Cluster Analysis is a well established methodology in Marketing research to perform market segmentation. Data from questionnaire surveys are often measured on ordinal scales, which are usually analyzed as they were of the interval type or by having recourse to non parametric distances like the γ index or distances based on the Spearman rank correlation. New dissimilarity distances dealing with variables of the ordinal type are presented.

Key words: Ordinal Variables, Cluster Analysis, Thurstone Transformation

1 Introduction

Cluster Analysis is a well established methodology in Marketing research to perform market segmentation.

Furthermore, it often happens that data are measured on ordinal scales; a typical example concerns customer satisfaction surveys, where responses given to a questionnaire are on Likert type scales, usually assuming a unique common finite set of possible categories.

We propose a transformation based on the traditional psychometric approach, by applying a method for treating ordinal measures according to the well-known Thurstone scaling procedure [7], assuming the presence of a continuous underlying variable for each ordinal indicator [6].

Giuseppe Boari · Gabriele Cantaluppi · Angelo Zanella
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, e-mail: {giuseppe.boari, gabriele.cantaluppi, angelo.zanella}@unicatt.it

2 Assumptions on the Genesis of Ordinal Observed Variables

The set of responses are assumed to be expressed on a conventional ordinal scale. This type of scale requires, according to the classical psychometric approach, appropriate scaling methods to be applied. In this regard we propose to adopt the traditional procedure, by considering the existence, for each observed ordinal manifest variable, of an underlying continuous unobservable latent variable.

Let us denote with $\mathbf{X} = (X_1, \dots, X_K)'$ the K -dimensional categorical random variable giving rise to the set of responses and assume, for simplicity, that the components of \mathbf{X} are defined on the same I ordered categories, coded into the conventional integer values $i = 1, \dots, I$.

Let $P(X_k = i) = p(x_{ki})$, with $\sum_{i=1}^I p(x_{ki}) = 1, \forall k$, be the corresponding marginal probabilities and let

$$F_k(i) = \sum_{j \leq i} p(x_{kj}) \quad (1)$$

be the probability of observing a conventional value x_{kj} for X_k not greater than i .

Furthermore assume that to each categorical variable X_k there corresponds an unobservable latent variable X_k^* , which is represented on an interval scale with a continuous distribution function $\Phi_k(x_k^*)$.

The latent variables X_k^* are usually called in the psychometric literature instrumental latent variables [1, 2]. The marginal distributions of the continuous K -dimensional latent random variable $\mathbf{X}^* = (X_1^*, \dots, X_K^*)$ are usually assumed to be standard normal.

Each observed ordinal indicator $X_k, k = 1, \dots, K$, is related to the corresponding latent continuous X_k^* by means of a non linear monotone function, see [5], of the type

$$X_k = \begin{cases} 1 & \text{if } X_k^* \leq a_{k,1} \\ 2 & \text{if } a_{k,1} < X_k^* \leq a_{k,2} \\ \vdots & \\ I_k - 1 & \text{if } a_{k,I_k-2} < X_k^* \leq a_{k,I_k-1} \\ I_k & \text{if } a_{k,I_k-1} < X_k^* \end{cases} \quad (2)$$

where $a_{k,1}, \dots, a_{k,I_k-1}$ are marginal threshold values defined as $a_{k,i} = \Phi^{-1}(F_k(i)), i = 1, \dots, I_k - 1$, being $\Phi(\cdot)$ the cumulative distribution function of a specific random variable, we will consider the standard Normal; $I_k \leq I$ denotes the number of categories effectively used by the respondents; $I_k = I$ when each category has been chosen by at least one respondent.

Later on the $a_{k,i}$ will be estimated by means of the empirical distribution function.

Thus a point estimate of the underlying continuous X_k^* cannot be uniquely determined in presence of ordinal variables for the generic subject: we can only establish an interval of possible values between the threshold values pertaining to the latent variable X_k^* underlying each ordinal manifest variable.

Later on we will replace threshold values equal $\pm\infty$ with ± 4 , that is we define $a_{k,0} = -4$ and consider $a_{k,I} = 4$.

3 Score Evaluation

Classical distance functions need a value to be computed for each subject, so one of the following proposals can be adopted in order to assign a value to the latent variable X_k^* :

- **Median estimation.** Compute the median of the variable X_k^* over the pair of thresholds giving the category expressed by the respondent

$$\text{Median}(X_k^* | a_{k,i-1} < X_k^* \leq a_{k,i}) = \Phi^{-1} \left(\frac{\Phi(a_{k,i-1}) + \Phi(a_{k,i})}{2} \right). \quad (3)$$

- **Mean estimation.** Compute the mean of the variable X_k^* over the pair of thresholds giving the category expressed by the respondent

$$E(X_k^* | a_{k,i-1} < X_k^* \leq a_{k,i}) = \frac{\phi(a_{k,i-1}) - \phi(a_{k,i})}{\Phi(a_{k,i}) - \Phi(a_{k,i-1})}, \quad (4)$$

being $\phi(\cdot)$ the density function corresponding to $\Phi(\cdot)$. The relationship can be obtained by deriving with respect to t the Moment Generating Function of $X_k^* | (a_{k,i-1} < x_k^* \leq a_{k,i})$, $M(t) = e^{\frac{1}{2}t^2} \cdot \frac{\Phi(a_{k,i} - t) - \Phi(a_{k,i-1} - t)}{\Phi(a_{k,i}) - \Phi(a_{k,i-1})}$, and setting $t = 0$.

Clustering algorithms can then be applied by means of classical distances computed on transformed values. Furthermore new alternative distances can also be introduced, which use all the information given by the knowledge of the distributions of X_k^* conditional on pair of adjacent thresholds.

4 Use of All Information Provided by the Latent Instrumental Variables

With reference to the previous approach, we can now define a family of new distances derived from the generic Minkowski distance hereafter recalled. Let x_{kr}, x_{ks} be the realizations of X_k , $k = 1, \dots, K$, for two subjects r and s . The Minkowski distance of order p between (x_{1r}, \dots, x_{Kr}) and (x_{1s}, \dots, x_{Ks}) in presence of interval variables is

$$\left(\sum_{k=1}^K |x_{kr} - x_{ks}|^p \right)^{1/p} = \left(\sum_{k=1}^K h_p(x_{kr}, x_{ks}) \right)^{1/p}. \quad (5)$$

For $p = 2$ we have the Euclidean distance

$$\sqrt{\sum_{k=1}^K (x_{kr} - x_{ks})^2} = \sqrt{\sum_{k=1}^K h_2(x_{kr}, x_{ks})}, \quad (6)$$

while for $p = 1$ we have the Manhattan or City-Block distance

$$\sum_{k=1}^K |x_{kr} - x_{ks}| = \sum_{k=1}^K h_1(x_{kr}, x_{ks}). \quad (7)$$

To apply the Minkowski distances when some variables are of the ordinal type we need to define $h_p(\cdot)$ for this specific case, in order to use all the information provided by the underlying latent variable X_k^* when transformations (3) or (4) are not adopted and we may refer to generic latent values obtained by using relationship (2).

For example in case of a Euclidean-like distance let us assume $\phi_r(x_{kr}^*)$ and $\phi_s(x_{ks}^*)$ be the density functions of $X_{kr}^* | (a_{k,i-1} < x_{kr}^* \leq a_{k,i})$ and $X_{ks}^* | (a_{k,j-1} < x_{ks}^* \leq a_{k,j})$

$$\phi_r(x_{kr}^*) = \frac{\phi(x_{kr}^*)}{\Phi(a_{k,i}) - \Phi(a_{k,i-1})} \quad \phi_s(x_{ks}^*) = \frac{\phi(x_{ks}^*)}{\Phi(a_{k,j}) - \Phi(a_{k,j-1})}.$$

The generic element $h_2(x_{kr}, x_{ks})$ pertaining to the k th variable in the distance function (6) can be redefined by having recourse to the underlying non observable variables X_{kr}^* and X_{ks}^* and considering the distribution of the random variables giving the difference of two generic realizations of X_k^* for two generic subjects, r and s , who chose, respectively, the categories i and j :

$$h_2(x_{kr} = i, x_{ks} = j) = \int_{a_{k,i-1}}^{a_{k,i}} \int_{a_{k,j-1}}^{a_{k,j}} (x_{kr}^* - x_{ks}^*)^2 \phi_r(x_{kr}^*) \phi_s(x_{ks}^*) dx_{kr}^* dx_{ks}^*, \quad (8)$$

$i, j \in \{1, 2, \dots, I\}$, having assumed the independence across subjects.

We have obtained in this way the average square difference on the two intervals defined by the thresholds. The Euclidean distance between subjects r and s is then obtained by applying relationship (6).

We foresee to analyse the properties of these new distances and compare their clustering performance with procedures based on the optimal scaling transformation of data [4].

References

1. Bollen, K.A.: Structural Equations with Latent Variables, John Wiley, New York (1989)
2. Bollen, K.A., Maydeu-Olivares, A.: A Polychoric Instrumental Variable (PIV) Estimator for Structural Equation Models with Categorical Variables. *Psychometrika* **72**, 309–326 (2007)
3. Heagerty, P.J., Zeger, S.L.: Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024–1036 (1996)
4. Mair, P., de Leeuw, J.: A General Framework for Multivariate Analysis with Optimal Scaling: The R Package aspect. *Journal of Statistical Software* **32**(9) 1–23. <http://www.jstatsoft.org/v32/i09/> (2010)
5. Muthén, B.O.: A general structural equation model with dichotomous, ordered, categorical, and continuous latent variable indicators. *Psychometrika* **49**, 115–132 (1984)
6. Pearson, K.: Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **195** 1–47 (1900)
7. Thurstone, L.L.: The measurement of Values, University of Chicago Press, Chicago (1959)

A general model for INDCLUS with external information

Laura Bocci and Donatella Vicari

Abstract This paper presents a model for partitioning two modes of three-way proximity data which generalizes INDCLUS by incorporating possible external information on objects and/or subjects. Specifically, subjects are partitioned into homogeneous classes, where class-conditional groups of objects are determined. The classifications of both objects and subjects are assumed to be related to possible external variables to better account for the meaning and the determinant of the groups. The model is fitted in a least-squares framework and an efficient ALS algorithm is given. An illustrative application to a benchmark data set is presented.

1 Introduction

Typical two-mode three-way proximity data consist of several symmetric matrices \mathbf{S}_h ($h=1, \dots, H$), whose entries represent pairwise proximities between N objects coming from an individual subject (occasion, experimental condition, or other source of data). Such three-way data generally contain a wide range of information, which is, however, usually complex and hard to comprehend. Specific methodologies are needed for extracting relevant features and a way to reach this aim is to synthesize data by reducing one or more of the modes to a small number of homogeneous classes. Here, we focus on the simultaneous reduction of the two modes (objects and subjects) of the proximity data set.

Carroll and Arabie (1983) introduced the INDCLUS (INDividual Differences CLUStering) model to extract overlapping clusters of the N objects from the set of H proximity matrices. Specifically, the model assumes that each group of objects is differently weighted by each subject (occasion). Therefore, all subjects in the

¹ Laura Bocci, Dipartimento di Comunicazione e Ricerca Sociale, Sapienza Università di Roma; email: laura.bocci@uniroma1.it

Donatella Vicari, Dipartimento di Scienze Statistiche, Sapienza Università di Roma; email: donatella.vicari@uniroma1.it

INDCLUS model are assumed to employ the same partition of objects, but with different patterns of weights. A fuzzy variant of the INDCLUS model, called FINDCLUS, has been proposed by Giordani and Kiers (2012), which assigns a fuzzy membership matrix, still common to all subjects, to the classification of objects.

In the similar context of the INDSCAL model, when the number of subjects is large, different sets of weights are rarely interpreted for individual subjects as pointed out by Winsberg and De Soete (1993). They have proposed to classify the subjects into a small set of classes in order to obtain a more parsimonious model in a mixture model approach. Furthermore, in order to handle possible systematic differences in the judgements, a different model, K -INDSCAL, has been proposed in the least squares framework (Bocci and Vichi, 2011).

Moreover, the possible availability of external information on the objects may be used to better interpret the results. For example, the procedures of cluster analysis (CA) or Multidimensional Scaling (MDS) and regression relating proximities to external variables are usually carried out independently and sequentially, although combinations of two of the three procedures (CA and MDS, or MDS and regression, as in Bock, 1997, Heiser, 1993) or all of them (Kiers et al., 2005) have been proposed in the literature.

The present paper aims at modelling both the subject and object heterogeneity of the three-way data and simultaneously incorporating the possible external information available. Specifically, we assume that a number of unobserved classes of subjects does exist, each having a different weight structure for the common groups of objects. To better capture the common behaviour of the classes of subjects in evaluating the pairwise proximities, the class-conditional set of weights is in turn linearly related to external variables on objects and/or subjects.

2 The Model

The INDCLUS model (Carroll and Arabie, 1983) can be written

$$\mathbf{S}_h = \mathbf{P}\mathbf{W}_h\mathbf{P}' + c_h\mathbf{1}_N\mathbf{1}_N' + \mathbf{E}_h, \quad h=1,\dots,H, \quad (1)$$

where $\mathbf{P}=[p_{ij}]$ ($p_{ij} = \{0,1\}$ for $i=1,\dots,N$ and $j=1,\dots,J$) is a $N \times J$ binary matrix defining the possibly overlapping clustering of N objects, \mathbf{W}_h is a non-negative diagonal weight matrix of order J for subject h , c_h is a real-valued additive constant for subject h , $\mathbf{1}_N$ denotes the column vector with N ones and \mathbf{E}_h is the square matrix of error components which explains for the part of \mathbf{S}_h not accounted for by the model.

When subjects present systematic differences in their judgements, they can be partitioned into K homogeneous classes and a different set of weights associated to the J groups of objects may be assumed. This specifies a new model

$$\mathbf{S}_h = \sum_{k=1}^K u_{hk}\mathbf{P}\mathbf{W}_k\mathbf{P}' + c_h\mathbf{1}_N\mathbf{1}_N' + \mathbf{E}_h, \quad h=1,\dots,H, \quad (2)$$

where $u_{hk} \in \{0, 1\}$, ($h=1,\dots,H$ and $k=1,\dots,K$), $\sum_{k=1}^K u_{hk} = 1$ ($h=1,\dots,H$) are the entries of a membership matrix \mathbf{U} , identifying the partition of the H subjects into K classes and \mathbf{W}_k is a $J \times J$ class-conditional diagonal matrix of non-negative weights associated to the groups of objects.

We suppose that an $N \times M$ matrix \mathbf{X} of the scores of the N objects on M variables and an $H \times V$ matrix \mathbf{Z} of the scores of the H subjects on V variables are measured and the set of class-conditional non-negative weights \mathbf{W}_k can be modelled as follows

$$\mathbf{W}_k = g_k \mathbf{B}_k = g_k \text{diag}(\mathbf{b}_k) \quad (k=1, \dots, K) \quad (3)$$

where \mathbf{B}_k is a diagonal matrix of order J having vector \mathbf{b}_k as main diagonal and g_k is a non-negative weight.

The underlying assumption is that vectors \mathbf{b}_k and $\mathbf{g} = [g_1, \dots, g_k, \dots, g_K]'$ depend on the \mathbf{X} and \mathbf{Z} variables, respectively, and can be modelled as follows

$$\mathbf{b}_k = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}\boldsymbol{\beta}_k, \quad \mathbf{b}_k \geq \mathbf{0}, \quad (k=1, \dots, K) \quad (4)$$

and

$$\mathbf{g} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Z}\boldsymbol{\gamma}, \quad \mathbf{g} \geq \mathbf{0}. \quad (5)$$

Moreover, in order to solve the identifiability problem for \mathbf{W}_k in (3), vector \mathbf{b}_k is constrained to sum to 1.

By including (3), (4) and (5) in model (2), we get

$$\mathbf{S}_h = \sum_{k=1}^K u_{hk} g_k \mathbf{P} \text{diag}((\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}\boldsymbol{\beta}_k)\mathbf{P}' + c_h \mathbf{1}_N \mathbf{1}_N' + \mathbf{E}_h, \quad (h=1, \dots, H). \quad (6)$$

In model (6), the classification matrices \mathbf{U} and \mathbf{P} of subjects and objects, respectively, the class-conditional vectors of coefficients $\boldsymbol{\beta}_k$ ($k=1, \dots, K$) and $\boldsymbol{\gamma}$, and the individual constant c_h ($h=1, \dots, H$) can be estimated according to the following least-squares fitting problem

$$\min F(\mathbf{U}, \mathbf{P}, \boldsymbol{\beta}_k, \boldsymbol{\gamma}, c_h) = \sum_{h=1}^H \|\mathbf{S}_h - \sum_{k=1}^K u_{hk} g_k \mathbf{P} \text{diag}((\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}\boldsymbol{\beta}_k)\mathbf{P}' - c_h \mathbf{1}_N \mathbf{1}_N'\| \quad [\text{P1}]$$

subject to

$$u_{hk} = \{0, 1\} \quad (h=1, \dots, H; k=1, \dots, K) \quad \text{and} \quad \sum_{k=1}^K u_{hk} = 1 \quad (h=1, \dots, H),$$

$$p_{ij} = \{0, 1\} \quad (i=1, \dots, N; j=1, \dots, J),$$

$$(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}\boldsymbol{\beta}_k \geq \mathbf{0} \quad \text{and} \quad ((\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}\boldsymbol{\beta}_k)' \mathbf{1}_j = 1$$

$$g_k \geq 0 \quad (k=1, \dots, K).$$

Problem [P1] can be solved using an appropriate coordinate descent algorithm also known as Alternating Least-Squares (ALS) algorithm, which alternates the estimation of a set of parameters when all the other are fixed. The algorithm here proposed detects a covering \mathbf{P} of objects and a partition \mathbf{U} of subjects. After these two steps, within each class of subjects the vectors of coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_k$ are estimated so that \mathbf{g} and \mathbf{b}_k are updated straightforwardly. Finally, the individual additive constant c_h is estimated by successive residualizations of the three-way data matrix. The five main steps are alternated and iterated until convergence and the best solution over different random starting classification matrices is retained to prevent from local minima.

3 Illustrative application

We have consider the Cola data published by Schiffman, Reynolds, and Young (1981, pp. 33-34). In a sensory experiment, 10 subjects (nonsmokers, aged 18-21 years) tasted ten different brands of cola: Diet Pepsi (DiP), RC Cola (RCC), Yukon (Yuk), Dr. Pepper (DRP), Shasta (Sha), Coca Cola (CoC), Diet Dr. Pepper (DDP), Tab (Tab), Pepsi Cola (PCo), Diet Rite (DiR). Each subject provided 45 dissimilarity judgments which were transcribed on a scale from 0-100 representing same (near 0) and different (near 100). Some external information on subjects and objects (colas) are available. We know that five subjects (A, D, E, F, I) have the ability to taste the chemical compound PTC which is bitter to all of them and is tasteless to all the other five subjects (B, C, G, H, J) who do not have this inheritable characteristic. About objects, it is possible to distinguish between diet (DiP, DDP, Tab, DiR) – nondiet (RCC, Yuk, DRP, Sha, CoC, PCo) colas.

We have applied the proposed model after the transformation of the dissimilarities into similarities, considering $K=2$ classes of subjects and $J=2$ groups of colas and assuming that $\beta_1 = \beta_2 = \beta$ for parsimony. The best solution retained over 100 starting runs corresponds to a partition which distinguishes between cherry (DRP, DDP) versus non cherry (DiP, Tab, DiR, RCC, Yuk, Sha, CoC, PCo) colas, while the first class C_1 of the partition of the 10 judges consists of subjects (A, D, E, F, G, I), i.e. all the PTC-taster plus a non PTC-taster, and the second class C_2 of the remaining four subjects (B, C, H, J). Class C_1 shows a class-weight smaller than class C_2 ($g_1 = 94.7738$ and $g_2 = 123.8915$) due to a positive coefficient ($\gamma = 14.3960$) of the external variable on subjects. All subjects weigh more the group of cherry colas (its weight is 0.5714 vs 0.4286 of the non-cherry group of colas), in fact the binary variable diet-nondiet has a positive coefficient ($\beta = 1.1429$).

References

1. Bocci, L., Vichi, M.: The K -INDSCAL model for heterogeneous three-way dissimilarity data. *Psychometrika* **76**, 691-714 (2011).
2. Bock, H.H.: On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: Bozdogan, H., Gupta, A.K. (eds.), *Multivariate Statistical Modeling and Data Analysis*, pp. 17–34. New York:Reidel (1987).
3. Carroll, J.D., Arabie, P.: INDCLUS: an Individual Differences Generalization of ADCLUS model and the MAPCLUS algorithm. *Psychometrika* **48**, 157-169 (1983).
4. Giordani, P., Kiers, H.A.L: FINDCLUS: Fuzzy INdividual Differences CLUstering. *J. Classif.* **29**, 170-198 (2012).
5. Heiser, W.J.: Clustering in low-dimensional space. In: Opitz, O., Lausen, , Klar, R. (eds.), *Information and Classification: Concepts, Methods and Applications*, pp 162–173. Berlin Heidelberg NewYork: Springer (1993).
6. Kiers, H.A.L, Vicari, V., Vichi, M.: Simultaneous classification and multidimensional scaling with external information, *Psychometrika* **70**, 433-460 (2005).
7. Schiffman, S.S., Reynolds, M.L., Young, F.W.: *Introduction to multidimensional Scaling*, London: Academic Press (1981).
8. Winsberg, S., De Soete, G.: A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika* **58**, 315-330 (1993).

The financial literacy and the undergraduates

Paola Bongini, Paolo Trivellato, Mariangela Zenga

Abstract It is acknowledged that financial literacy among the young is influenced by socio-demographic characteristics. The paper investigates whether differing profiles of students measured according to standard socio-demographic characteristics, in particular gender- show different levels of financial literacy at the beginning of their university careers. The study takes in account 366 Business Studies freshmen during their first few weeks at a large Italian university. The non-randomly chosen sample is composed of freshmen with no prior educational exposure to financial matters except, in some cases, a high school diploma in commercial studies or a personal interest in financial issues.

Key words: financial literacy; freshmen; latent regression Rasch model; gender differences.

1 Introduction

In recent years, financial literacy has gained the attention of a wide range of organizations, both at the national level - policymakers and financial regulatory authorities - and at the international level, with the OECD leading the way with its International. Moreover, widespread financial illiteracy among young people is of particular concern for two main reasons. First, as they enter adulthood, a number of important financial decisions are to be undertaken (such as financing college studies; moving away from home; purchasing their first car; using credit cards;), for which they might not be adequately prepared. It is well known that financial literacy of the young is influenced by socio-demographic characteristics, in particular gender; work and financial experience; parents behavior and background, such as educational attainment; personal education. Indeed, participation in courses on personal

Paola Bongini, Paolo Trivellato, Mariangela Zenga
Milano-Bicocca University, via B. degli Arcimboldi, e-mail: mariangela.zenga@unimib.it

finance or choosing a business major enhances students financial literacy: individuals receiving a financial education at high school or college tend to show proper financial behavior and attitudes. The financial literacy of college students has recently attracted a growing number of studies, especially in those countries (as the US, UK and Australia) where, on the one side, students borrow heavily to finance their higher education through student loans and, on the other, the use of credit cards is widespread. The issue has received less attention in continental Europe, where parents are the main source of finance and young people live at home longer. However, a first wave of research is being done with interesting results, providing insight both as regards the domain of measurement issues and our understanding of the determinants of financial literacy. These studies have similar findings, i.e., that the level of financial literacy of college students seems wanting and is associated with gender, ethnicity, education, work experience, social origins, interaction with peers and peer behavior. If the gender gap problem is considered in financial knowledge, several studies suggest that, on average, female college students are less financially knowledgeable than their male counterparts. Chen and Volpe (2002) find that college women generally have less enthusiasm for, lower confidence in and less willingness to learn about personal finance topics than men. Similarly, Ford and Kent (2010) find higher intimidation and lower market awareness among collegiate females. Outside the US, on the other hand, recent surveys do not confirm gender differences: Wagland and Taylor (2009), surveying business students at the University of Western Sydney (UWS) suggested that gender was not a significant factor among Australian collegiate; analogously, Koshal et al. (2008) reported that gender differences were not significant among a sample of Indian MBA Students. As far as financial behavior is concerned, there is mixed support for gender differences in financial practices. Hayhoe et al (2000) find that female students were more likely to have a written budget, plan their spending, keep bills and receipts and save regularly; on the contrary, females are more likely to engage in risky behavior (Lyons, 2004) or to engage in problematic financial behaviors (Worthy, Jonkman and Blinn-Pike, 2010).

2 The survey

Our study tests whether differing profiles of students measured according to gender and other standard socio-demographic characteristics show different levels of financial literacy at the beginning of their university careers. We surveyed 366 Business Studies freshmen during their first few weeks at a large Italian state university (Milan-Bicocca) in 2009: in other words, our sample comprised freshmen without prior educational exposures to financial matters unless they had a high school diploma in commercial studies or had a personal interest in finance. The survey instrument consisted of 39 questions, 13 of which selected from the JumpStart Coalition test of financial literacy. The multiple choice test used in the 2008 JumpStart Coalition Survey of College Students [7], and aimed at assessing the financial lit-

eracy of Young American Adults, was translated and adapted to the Italian context. According to the JumpStart Coalition, such questions can be grouped into three specific areas: a) money management; b) saving and investing; c) spending and credit. They are meant to express the concepts underlying basic financial transactions, financial planning, day-to-day financial decision-making or functioning of the banking system up to more complex issues, such as risk and returns of different asset classes or retirement planning. A second group of items is related to perceived knowledge. Students were asked to rate on a four-point scale their level of knowledge about specific financial topics, related to our three areas of savings and investments, spending and credit, and money management. In order to explain differences among students with respect to financial literacy, we use the latent regression Rasch model. Such a model, described in [12] is particularly helpful when sub-populations can be identified in the sample. For Rasch model, the probability of a given student answering an item correctly is a logistic function of the difference between the p -th persons level of knowledge of financial literacy (θ_p) and the level of financial literacy expressed by the i -th item (β_i), that is :

$$\eta_{pi} = \ln \frac{P(X_{pi} = 1 | \theta_p, \beta_i)}{[1 - P(X_{pi} = 1 | \theta_p, \beta_i)]} = \theta_p - \beta_i \quad (1)$$

The model proposed by Zwinderman differs from Rasch model in that θ_p is replaced with a linear regression equation in equation (1):

$$\theta_p = \sum_{j=1}^J \vartheta_j Z_{pj} + \varepsilon_p \quad (2)$$

so that:

$$\eta_{pi} = \sum_{j=1}^J \vartheta_j Z_{pj} - \beta_i + \varepsilon_p \quad (3)$$

where Z_{pj} is the value of the student p on student property (covariate) j ($j = 1, \dots, J$), ϑ_j is the regression weight of the student property j , ε_p is the effect remaining after the effect of the person properties is accounted for (with $\sigma_p \sim N(0, \varepsilon_p^2)$).

We constructed a latent regression Rasch model¹ using as items both the 13 multiple-choice items and the 8 self-assessment items. The person properties used in our analysis are Gender, Major, Nationality, High School background, Parents Social background, Parents schooling, Work experience and Financial experience .

Regarding the gender gap, we obtained a result similar to the results in literature. In fact, female students were substantially less financially literate than their male counterparts. Male students, with an estimated effect equal to 0.164, were 1.18 times more financially literate than women. Our sample of freshmen displayed different financial literacy levels according to their gender, even after controlling for item

¹ We first built measures of tested financial knowledge and of self-assessed financial knowledge, subsequently we analyze the results of the explanatory item response models [3]. Parameters estimations and other data manipulation were obtained using R software (lme4 package), as in [4].

characteristics and other socio-demographic factors. Our evidence confirms in the American literature, where gender is an important predictor of financial literacy.

3 Conclusions

Our paper investigates whether gender gap is an important factor explaining university students financial literacy at the beginning of their university careers. For this reason, first-year Business students were targeted. In spite of some shortcomings (non-random sample, first year students enrolled in only one institution, limited number of cases) our research obtained the following results. Our findings confirm the results of previous research undertaken since the 1990s in America and Australia into the socio-economic variables associated with financial literacy. Not surprisingly, we found that in Italy too, financial literacy depends on gender, but also on family background, previous high school experience, financial experience, and nationality.

References

1. Bongini, P., Trivellato, P., Zenga, M.: Measuring financial literacy among students: an application of Rasch analysis. *Electronic Journal Of Applied Statistical Analysis*, **5**(3), 425–430 (2012).
2. Chen, H., Volpe, R.P.: Gender differences in personal financial literacy among College students. *Financial Services Review*, **11**, 289–307 (2002).
3. De Boeck, P., Wilson, M.: *Explanatory Item Response Models: A Generalized Linear and Non Linear Approach*. Springer, New York. (2004)
4. De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., Partchev, I.: The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, **39**(12), 1–28 (2011).
5. Ford, M.W., Kent, D.W.: Gender differences in student financial market attitudes and awareness: an exploratory study. *Journal of Education for Business*, **85**(1), 7–12 (2010).
6. Hayhoe, C.R., Leach, L.J., Turner, P.R., Bruin M.R., Lawrence, F.C.: Differences in Spending Habits and Credit Use of College Students. *The Journal of Consumer Affairs*, **34**(1), 113–133 (2000).
7. JumpStart Coalition for Personal Financial Literacy. *The Financial Literacy of Young American Adults*, <http://www.jumpstart.org/assets/files/2008SurveyBook.pdf> (2008)
8. Koshal, R.K., Gupta, A.K., Goyal A., Choudhary V.N.: Assessing Economic Literacy of Indian MBA Students. *American Journal of Business*, **23**(2), 43–52 (2008).
9. Lyons, A.C.: Risky Credit Card Behavior of College Students. In: Xiao, J.J. (eds.) *Handbook of Consumer Finance Research*, pp. 185–207. Springer (2008).
10. Wagland, S.P., Taylor, S.T.: When it comes to financial literacy, is gender really an issue?. *The Australasian Accounting Business and Finance Journal*, **3**(1), 13–23 (2009).
11. Worthy, S.L., Jonkman, J., Blinn-Pike, L.: Sensation-Seeking, Risk-Taking, and Problematic Financial Behaviors of College Students. *Journal of Family and Economic Issues*, **31**, 161–171 (2010).
12. Zwinderman, A.H.: A generalized Rasch model for manifest predictors. *Psychometrika*, **56**(4), 589–600 (1991).

Credit risk measurement and ethical issue: some evidences from the italian banks

Riccardo Bramante, Marta Nai Ruscone, Pasquale Spani

Abstract This paper gives a contribution in variable identification within credit scoring models using Random Forest. Specifically, we provide some insights about the behavior of the variable importance index based on random forests, focusing on the differences between “for-profit” and “not-for-profit” enterprises. We investigate two classical issues of variable selection: the first one is variable extraction for bankruptcy interpretation, whereas the second one is more restrictive and tries to design a good prediction model. Finally we provide an application to a real data set provided by Banca Popolare Etica.

Key words: Random Forest, Variable Importance Measure, Ethical Dimension

1 Introduction

Rating the “not-for-profit” sector is a complex issue due to its specific purposes and form of activity. Moreover, the challenge for financial institutions, working with these type of organizations, of assigning them to the appropriate rating scale is a relevant topic nowadays due to the their sharp increase over the last two decades.

Riccardo Bramante
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milan
e-mail: riccardo.bramante@unicatt.it

Marta Nai Ruscone
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milan
e-mail: marta.nairuscone@unicatt.it

Pasquale Spani
Banca Popolare Etica Scrl, Via Niccolò Tommaseo 7, 35131 Padova
e-mail: pasquale.spani@bancaetica.com

The main discriminating criterion, in comparison to those of conventional profit oriented companies, relies on a different interpretation of the classical indicators, which contribute to the organization debt service capacity, combined with qualitative factors.

The main contribution of this work is twofold: investigate the behavior of the variable importance index based on Random Forests (*RF*) and compare selection criteria in terms of bankruptcy prediction. The strategy involves a ranking of explanatory variables, separately for “not-for-profit” and “for-profit” enterprises, using the *RF* score of importance and a stepwise ascending variable introduction strategy. The standard algorithm is modified in the selection procedure in order to assess the impact of the two grouping variables (“not-for-profit” - “for-profit”; bankrupt - non bankrupt) on the variable importance measure.

The proposed selection method is empirically tested using a real data set provided by Banca Popolare Etica, an Institution focused on the “not-for-profit” sector and a pioneer in credit risk measurement techniques which integrate traditional quantitative factors and qualitative ones, related to the social environment of costumers.

2 Random Forest

Random Forest is a classification method that combines results from an ensemble of many, say M , decision trees built on bootstrap samples drawn with replacement from the original training sample. Each bootstrap sample size is the same, say N , as the original sample. Drawing with replacement guarantees that roughly $1/3$ of elements from the original sample are not used in each bootstrap sample (indeed, note that the probability of not drawing a particular element is $(1 - \frac{1}{N})^N \approx e^{-1}$). For each tree in the forest, elements of the original sample not used to grow a particular tree are called out-of-bag or *oob* elements for the tree. Assume that each element (object) in the training sample is given as a vector of P variables. At each stage of tree building, i.e. for each node of any particular tree in the forest, p variables out of all P are randomly selected, where $p \ll P$ (say, $p = \sqrt{P}$), and the best set of these p variables is used to split the data in the node. Each tree is grown to the largest extent possible, i.e. there is no pruning. In this way, *RF* consisting of M trees is constructed. Classification of each (new) object is made by simple voting of all trees.

3 Estimation of attribute importance

As data sets grow in their size and complexity, effective and efficient techniques are needed to target important features in the variable space. In this paper a new variable selection technique is introduced in supervised classification, based on the *RF* approach. Data sets are often described with far too many variables for practical

model building. Usually most of these variables are irrelevant to the classification matter, and obviously their relevance is not known in advance. There are several disadvantages of dealing with overlarge feature sets. One is purely technical: dealing with large feature sets slows down algorithms, takes too many resources and is simply inconvenient since many machine learning algorithms exhibit a decrease of accuracy when the number of variables is significantly higher than optimal (Kohavi, John, 1997). Therefore selection of the small (possibly minimal) feature set giving best possible classification results is desirable for practical reasons. This problem, known as minimal – optimal problem (Nilsson, Pena, Bjorkegren, Tegner, 2007), has been intensively studied and there are plenty of algorithms which are available to reduce feature set to a manageable size. Nevertheless, this very practical goal shadows another very interesting problem, i.e. the identification of all variables which are in some circumstances relevant for classification, the so called “all relevant problem. Finding all relevant attributes, instead of only the non-redundant ones, may be very useful in itself. In particular, this is necessary when one is interested in understanding mechanisms related to the subject of interest, instead of merely building a black box predictive model. For example, when dealing with results of credit risk, identification of all variables – related to the two considered groups – is necessary to understand completely the data generating process, whereas a minimal – optimal set of variables might be more useful to track bankruptcy risk.

The algorithm is a wrapper built around the *RF* classification algorithm implemented in the *R* package *randomForest* (Liaw, Wiener, 2002). The *RF* classification algorithm is relatively quick, can usually be run without tuning of the parameters and gives a numerical estimate of the feature importance. It is an ensemble method in which classification is performed by voting of multiple unbiased weak classifiers decision trees. These trees are independently developed on different bagging samples of the training set. The importance measure of a variable is defined as the loss of classification accuracy caused by a random permutation of attribute values between objects. It is computed separately for all trees in the forest which use a given attribute for classification. Then the average and standard deviation of the accuracy loss are computed. We want to take into account the fluctuations of the mean accuracy loss among trees in the forest. For each variable we create a corresponding “shadow” variable, whose values are obtained by shuffling values of the original attribute across objects. To this end we have extended the information system with variables that are random by design. We then perform a classification using all variables of this extended system, previously decorrelated to the response, and compute the importance of all variables. Thus, the set of “importances” of shadow variables is used as a reference for deciding which attribute is truly important. The algorithm iteratively compares importances of attributes with importances of shadow attributes, created by shuffling original ones. Attributes that have significantly worst importance than shadow ones are being consecutively dropped. On the other hand, attributes that are significantly better than shadows are admitted to be confirmed. The process of determining whether a set of variables contributes significantly to the final prediction is based on multiple application of *RF* along with the utilization of this set in other classification algorithms.

4 Empirical Evidence

The sample is composed of approximately 42 thousands firms – divided into “not-for-profit” and “for-profit” – from a database of Cooperative Banks including Banca Etica. Bankruptcy predictors are assumed to be a set of commonly used balance ratios, such as Return on Assets – Sales to Asset Ratio and Debt to Equity, along with specific “not-for-profit” indicators, for instance Length of the Operating Cycle and Working Capital Turnover Ratio. All the indicators are used to set up the model in the random forest case, separately for the two group of firms, and compare it with other tools, specifically the traditional classification scheme, i.e. Linear Discriminant Analysis (*LDA*). As for *RF*, variable extraction was performed on the basis of the proposed selection criterion; *LDA* was applied using stepwise regression in variable selection.

In our experiments we used classification accuracy and Receiver Operating Characteristics (*ROC*) / Area Under Curve (*AUC*) as model performance measures. *LDA* is compared to *RF* by means of the accuracy measures when *RF* selected variables are used in *LDA* modeling: *RF* slightly improve accuracy of default prediction in all the performed analysis, whilst results obtained in the two considered groups of companies (“not-for-profit” – “for-profit”) are substantially equivalent. Moreover, in most of the cases classification error is low, and in all cases it is significantly lower than percent error of the random classifier. In detail, we reported a 4% improvement (from 78% of *LDA*) in global success rate of discrimination when using *RF* selected variables; if only the bankrupt group is considered the improvement was slightly higher (9%). As for the *ROC* measure, we obtained a global 2% improvement, similar for the two considered groups. Regarding the variable selection algorithm, “not-for-profit” distinguishing features are confirmed by *RF* variable extraction since specific “not-for-profit” ratios are extracted only within this group of enterprises.

References

1. Breiman, L.: Random Forests. *Machine Learning*. **45**, 5–32 (2001)
2. Hastie, T, Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference and prediction. Springer, New York (2001)
3. Kohavi, R., John G. H.: Wrappers for feature subset selection. *Artificial intelligence*. **97**, 273–324 (1997)
4. Liaw, A., Wiener, M.: Classification and Regression by randomForest. *R News*, **2**, 3, 18–22 (1997)
5. Nilsson, R., Pena, J., Bjorkegren, J., Tegner, J.: Feature Selection for Pattern Recognition in Polynomial Time. *Journal of Machine Learning Research*. **8**, 612 (1997)

On the Stylometric Authorship of Ovid's Double Heroides: An Ensemble Clustering Approach

Pierpaolo Brutti, Lucio Ceccarelli, Fulvio De Santis, Stefania Gubbiotti

Abstract *Double Heroides* are six elegies traditionally attributed to Ovid, whose authenticity have been repeatedly questioned. As a contribution to establish the period of composition of these elegies, this article proposes a statistical analysis based on consensus clustering of mixed data composed of standard (i.e. scalar) and compositional variables derived by an extensive metrical study of the Ovid's poetic production.

Key words: Consensus clustering, compositional data, Latin metric.

1 Introduction: Ovid's double letters

A relevant open problem in Latin literature is the authorship and the date of Ovids *Heroides XVI-XXI (double Heroides)*. The *Heroides* are a collection of 21 elegies written in *elegiac couplets*. The authenticity of the collection has been questioned since Lackmann [1] who, on the basis of certain metrical anomalies with respect to genuine Ovidian poems as well as of non metrical considerations, claimed that some of the *Heroides* (including the double Heroides) were not composed by Ovid. In this article we propose a statistical approach to provide further elements of discussion on the dating problem. In particular our attempt is to find statistical support, based on metrical features, to the hypotheses that *double Heroides* belong to one of the three main phases of Ovid's poetic production: the early Roman period, the mature Roman period and the exile period. The stylometric methodology we adopt is based on consensus clustering techniques for mixed data composed of standard (i.e. scalar) and compositional variables.

Lucio Ceccarelli
Università degli Studi dell'Aquila e-mail: lucio.ceccarelli@univaq.it

Pierpaolo Brutti, Fulvio De Santis, Stefania Gubbiotti
Sapienza Università di Roma, e-mail: stefania.gubbiotti@uniroma1.it

2 Basics of metric: the elegiac couplet

Meter is the basic rhythmic structure of a verse. The study of meters and forms of versification is known as metric. We here focus on the so called *elegiac couplet* or *distich*, a poetic form initially introduced in Greek lyric, later on adopted by Roman poets, and in particular by Ovid. To illustrate the “anatomy” of the elegiac couplet, let us consider one of the 406 distichs from Ovid’s *Fasti VI*: The main steps in the

Tempora labuntur, tacitisque senescimus annis;
et fugiunt freno non remorante dies.

metrical analysis are summarized as follows (see Figure 2).

- Each syllable of the words of a verse is categorized as *long* (–) or *short* (U) according to its *weight*. (e.g. in *Tempora*, Tem- is long, -po- and -ra are short).
- Specific sequences of syllables define a *foot* (delimited by |...|), for instance

Dactyl (D)	formed by 3 syllables	– U U
Spondee (S)	formed by 2 syllables	– –
Trochee (T)	formed by 2 syllables	– U

- Each verse is formed by a specific number of *feet*. The first verse of the elegiac distich is called *hexameter*; the second one is called *pentameter*.
- The metrical pattern of feet in a verse is then summarized by a sequence of letters. For instance, in the couplet above the scheme is DSDDDS · DD–DD–

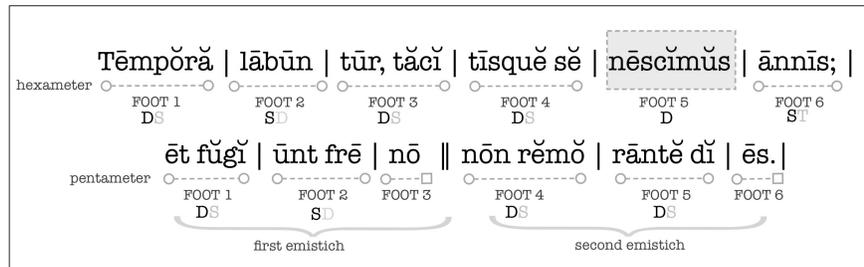


Fig. 1 “Anatomy” of the elegiac couplet; (–) and (U) denote long and short syllables respectively.

In general, in the hexameter each of the first four feet can be alternatively a *dactyl* or a *spondee*. The fifth foot is almost always a *dactyl*. The sixth foot is either a *spondee* or a *trochee*. Conversely, the pentameter is made up of two equal parts containing two dactyls followed by a long syllable (*hemistich*). Spondees replace dactyls in the first half, but never in the second. The choice of a particular scheme for the verses, together with many other metrical features that will be mentioned in the following section, yield a great variability in the realization of the elegiac distich, that strongly characterizes the style of each single author. For instance, if we consider the first four feet of the hexameter only, there are $2^4 = 16$ possible alternative choices of dactyls and spondees. In summary, the metrical technique is

a very personal ability of the poet and it reflects not only his skills, but also his sensitivity. In this sense, the stylometric analysis can be helpful in the attempt of attributing a poem to a specific author.

3 Stylometric features

The goal of a metric study is to identify characterizing stylistic features of a poet with respect to the metric language of the tradition, as well as to detect deviations of metric features of some parts of his own poetic production with respect to his entire work. This kind of study requires a translation of metrical phenomena into quantitative data, whose relevance can be pointed out uniquely by appropriate statistical analysis of poetic *corpora* (see [2]). All the poetic production in elegiac couplets attributed to Ovid has been examined from a metrical point of view in [3]. In this metrical study many quantitative informations has been extracted and collected on each section of every poem, such as the total frequency of the dactyls and the distribution of dactyls both in the first four feet of the hexameter and in the first half of the pentameter, the presence of short syllables in the couplet, the occurrence of some particular forms (e.g. *synalepha*, *clausula*), and so on. Consequently, our dataset consists of 15 distinct metrical features measured on each of the 27 poems considered as statistical units. Most of these 15 variables are actually frequency distributions summarizing the occurrence of a specific metrical phenomenon over a whole poem, and therefore they will be treated as *compositional data* (see [1] and [8]). For example, in Figure 3 are represented three variables observed on two of

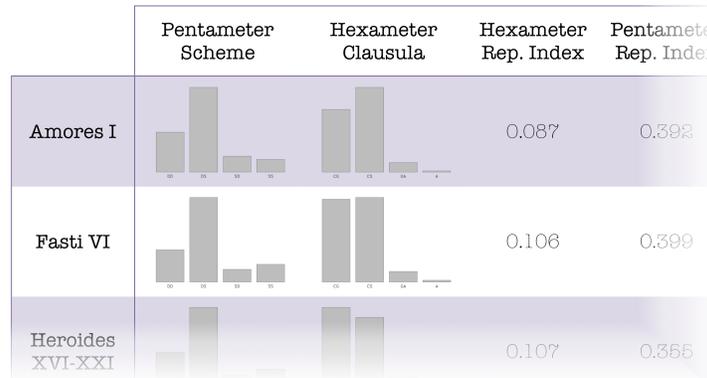


Fig. 2 A snippet of the stylometric dataset.

the poems: the first and the second one consist of distributions associated to the realizations of the pentameter scheme, and to the clausula in the hexameter respectively, whereas the third scalar one is a particular stylometric quantity usually called *replication index*.

4 Methods and results

Consensus clustering, emerged as an important elaboration of the classical clustering problem. These methods are commonly used to establish consensus among multiple clustering algorithms, or multiple realizations of the same clustering algorithm on a single dataset, or even to integrate multi-source data (for a survey see [7]). In this work we adapt to the peculiarities of the data at hand, two recently proposed techniques, namely the *Multiple Dataset Integration* method developed in [4] and the Bayesian consensus clustering based on finite Dirichlet mixture models described in [6]. Both approaches use a statistical framework to cluster each data source separately while simultaneously modeling dependence between the clusterings in order to borrow strength across data sources. Aggregation mechanisms of this type make the resulting overall clustering more robust and stable of other unsupervised classification solutions, while still allowing to better interpret the contribution of each data-source to the final partition.

In summary, the consensus clustering shows a quite remarkable distinction between three groups, characterized by different metric profiles. It can be noticed that in each group one of the three periods prevails. Finally, the *double Heroides* turn out to be compatible with the (stylo)metric features of the latest period of Ovid's poetry.

References

1. Aitchison J.: *The Statistical Analysis of Compositional Data*. The Blackburn Press, (2003)
2. Ceccarelli L.: *Contributi sulla storia dell'esametro latino*. Herder, Roma (2008)
3. Ceccarelli L.: L'evoluzione del distico elegiaco tra Catullo e Ovidio. In: Cristofoli R., Santini C., Santucci F. *Properzio tra tradizione e innovazione*, Atti del convegno internazionale (Assisi-Spello, 21-23 maggio 2010), Assisi 2012, pp. 47–97.
4. Kirk P., Griffin J.E., Savage R.S., Ghahramani, Z., Wild, D.L.: Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**(24), 3290-3297, (2012)
5. Lachmann K.: *De Ovidi epistulis*. Progr. Univers. Berolinensis, (1848)
6. Lock E.F., Dunson D.B.: Bayesian consensus clustering. *arXiv:1302.7280*, (2013)
7. Nguyen, N., Caruana, R. Consensus clusterings. In: *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, October 28-31, 2007, Omaha, Nebraska, USA, pp. 607-612. IEEE Computer Society, 2007
8. van den Boogaart K.G., Tolosana-Delgado R.: *Analyzing Compositional Data with R*. Springer, (2013)

Causal Inference in Gender Discrimination in China: Nutrition, Health, Care

Silvia Caligaris, Fulvia Mecatti and Patrizia Farina

Abstract The increased demand of gender-sensitive statistical information does not meet nevertheless the development of statistical tools and *ad hoc* models: most of gender statistical measures proposed in these decades are indeed composite indicators involving several arbitrary choices and affecting both indexes transparency and interpretation. The aim of this work is to show how graphical models and in particular causal graphs may represent useful tools to depict the complex relationships pattern among variables involved in gender disparity processes, to deep causal mechanisms originating gender gaps as well as to explore the effects of gender tailored policies. We show the potential of such models through an application to real data from China Health and Nutrition Survey 2011 ; in particular we explore the eventual existence of gender discrimination in children' nutrition and health as possible indicator of preference for sons. The analysis takes in exam socio-demographic, economical as well as biological variables. Resorting to the PC algorithm and the IDA algorithm, we aim to learn the underlying causal structure and to estimate causal effect of siblings on children' nutrition from observational data.

Key words: Causal Graph, China, Gender Gap Measurement, PC algorithm, IDA algorithm

Silvia Caligaris
University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1 - 20126 Milano
e-mail: s.caligaris@campus.unimib.it

Fulvia Mecatti
University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1 - 20126 Milano
e-mail: fulvia.mecatti@unimib.it

Patrizia Farina
University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1 - 20126 Milano
e-mail: patrizia.farina@unimib.it

1 Gender Statistics: what, why, how

Gender statistics appears as an independent field of statistics with its typical products such as indexes, tables and graphs that cuts across traditional applications in social, economical, human and life science. In a broader sense [7] it implies a forward-looking perspective inspired by the increasing demand of gender-sensitive statistical information arising from society, official agencies and economy.

Main aims of Gender Statistics includes:

- to understand gaps and disparities based on either gender—as a social structure—or sex—as a biological factor—;
- to explore the role of women *and* men in the society, economy and family;
- to formulate, monitor and evaluate policies;
- to monitor changes;
- to disseminate evidence-based information to the public.

Since the 80's several gender statistical indicators are being released annually by international agencies in order to compare and rank countries' performances with respect to selected macro-themes such as as economics, politics, education and health, under a gender perspective. Most of them are composite indicators provided by (usually linear) aggregation of a set of simple indicators such as female/male ratios, each singled out to measure one particular component of the gender gap viewed as the underlining latent variable. However the many arbitrary choices involved in the aggregation coupled with the incapability of a unique indicator to grasp the complexity of gender-related issues, lead to final indexes lacking of transparency and ultimately producing objectionable country rankings [1].

2 Causal Graphs as a Gender Statistics Tool

Main purpose of this work is to explore the potential of graphical models as a language to untangle the complex relationship among variables selected for statistically assessing gender disparities. We will focus on causal graphs which allows for deepening and interpreting the causal mechanism that may have originated a gender gap as well as estimating believes under changing conditions of the causal structure.

We now give a synthetic idea of how a pattern of relationships can be described by means of a graph. Let $G = (V, E)$ define a graph with set of edges E and of vertices V where the vertices represent variables and the edges will denote a relationship between the variables. The edges can either be directed (single arrowheaded) or undirected (unmarked link) or possibly bidirected (duble arrowheaded) denoting the existence of an unobserved common causes not included in the graph. Kinship terminology, e.g. parents, children, siblings etc. is common and effective language to denote relationships in a graph. For example in a direct edge $X_i \rightarrow X_j$ where X_i represents a direct cause of X_j , we will call X_i the parent and X_j a descendant [6]. A directed acyclic graph (henceforth DAG) is a graph with all edges directed and

no cycles, i.e. feedback processes. We will focus on DAGs causally interpreted, as known as *Causal Graphs*. Any missing arrow between two variables means that they have no causal relationship and the two are causally independent [8]; instead, conditional independence can be derived through the so called Markov Condition, namely when each variable is independent on its non-descendant in the graph given the state of its parents.

Conditional relationships encoded in a DAG can be easily read off by applying the graphical criterion called *d-separation* [8] which represents the translation device between the language of probability distributions and the language of causality.

DAGs can be learned by conditional independence information if we assume *faithfulness*: with it we can say that the graph that generated the data implies, by d-separation, exactly the independence relations that are present in the population. However, the same conditional independence structure can be described by several DAG which leads to a *Markov equivalence class* [8]. Markov equivalence classes of DAGs can be described uniquely by a completed partially directed acyclic graph (CPDAG) [2].

Since the DAGs encode conditional independences, information on the latter helps to infer aspects of the former. This intuition forms the basis of the PC algorithm (from its inventors Peter Spirtes and Clark Glymour) which is able to reconstruct the structure of the underlying DAG model given a conditional independence oracle up to its Markov equivalence class [9]. PC algorithm is efficiently implemented in the R-package *pcalg* [5].

Often the interest is also on estimating the size of the causal effect between pairs of variables: one can do it via Pearl's *do*-calculus. The operator $do(X_i = x_i)$ simulates physical intervention by removing certain functions from the model and by replacing them with the constant $X_i = x_i$, while keeping everything else unchanged [8]. Particularly this is accomplished by the IDA algorithm (Intervention when the DAG is Absent) [5] providing estimated bounds on causal effects from observational data generated from an unknown causal DAG).

3 Application to China Health and Nutrition Survey

Causal Inference is traditionally applied in genetic and epidemiology with the aim of exploring the effect of a treatment variable upon an outcome variable through randomized controlled experiments.

In many settings these experiments can be expensive, violating ethical standards or even not applying as for social studies. However, re- interpreting in a broader sense the concept of "treatment" and "outcome", we can efficiently extend the causal framework to gender statistics. Indeed the presence or absence of a gender discrimination in a specific gender-related field has a direct influence on the specific outcome measured as equal access to resources. Suppose for instance we are interested on education: the presence of an educational gender gap will influence the outcome measured as number of boys and girls at that specific grade . Conversely in absence

of gender discrimination the treatment variable "gender" should result independent by the outcome as it does not produce any effect on it.

Causal graphs as tools for gender statistics allow for a) exploring the causal structure underlying gender gap, b) identifying the main factors of gender gaps, c) highlighting both direct and indirect factors influencing gender unbalance persistence, d) exploring the potential causal effects of policies on gender disparities through the do-operator as it is possible to simulate physical intervention on a certain variable and therefore e) monitoring changes over time.

With the purpose of showing the potential of such models, an application of the causal framework to real data from China Health and Nutrition Survey (CHNS) 2011 [3] is performed.

In China a higher sex ratio at birth than the physiological level and an excess female child and infant mortality deviating from the worldwide norm, would represent outcomes of various kinds of discrimination against girls [4]. We explore in particular the existence of gender discrimination in children' and adolescents' (0-18 years old) nutrition and health patterns. Our aim is to verify if the treatment variable "gender" plays any influence on the outcome "child's health and nutrition". We measure the outcome through child's weight and height as indicators of long-term nutritional status and exposure to diseases both influenced by household decision-making and, presumably, by a preference for sons. The analysis also considers potential confounders [8], in particular parents' socio-demographic variables as education, job, health, diet knowledge, pregnancy history and fertility preferences as well as physical measurements as Body Mass Index (kg/m^2).

Through the PC algorithm we intend to inquire the structure of the underlying DAG model while through the IDA algorithm we explore the causal effect of siblings' presence on child's nutrition.

References

1. Caligaris, S., Mecatti, F., Crippa, F.: A Narrower Perspective? From a Global to a Developed-Countries Gender Gap Index. *Statistica*, special issue on gender studies in press (2013).
2. Chickering, D. M.: Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2**, 445-498 (2002).
3. China Health and Nutrition Survey. <http://www.cpc.unc.edu/projects/china>
4. Choe, M.K., Hao H., Feng, W.: Effects of gender, birth order, and other correlates on childhood mortality in China. *Biodemography and Social Biology*, **42**, 50-64 (1995).
5. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47**, 1-26 (2011).
6. Lauritzen, S.: *Graphical models*. Oxford University Press, USA (1996).
7. Mecatti, F., Crippa, F., Farina, P.: A Special Gen(d)er of Statistics: Roots, Development and Methodological prospects of a Gender Statistics. *International Statistical Review*, **80**, 452-467 (2012).
8. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge (2000).
9. Spirtes, P., Glymour, C.N., Scheines, R.: *Causation Prediction & Search 2e*, MIT press (2000).

Self-Selection and Direct Estimation of Across-Regime Correlation Parameter

Giorgio Calzolari and Antonino Di Pino

Abstract A direct Full Information Maximum Likelihood (*FIML*) procedure to estimate the “generally unidentified” across-regime correlation parameter in a two-regime endogenous switching model is here provided. The results of a Monte Carlo experiment, assuming normally distributed error terms, confirm consistency and relative efficiency of our direct *FIML* estimation.

Key words: Endogenous switching models, Across-Regime correlation parameter

1 Introduction

We consider a simultaneous Two-Equation model (generalized Roy model with two regimes), in which a non null correlation between the error terms occurs as a consequence of the joint influence of latent factors on both the choice of regime and the outcome gained by the subject in the chosen regime. However, this correlation (or covariance) across regimes is not empirically identifiable as a result of the selection criterion, involving that both dependent variables cannot be jointly observed. Nevertheless, the partial knowledge of this parameter is considered relevant to provide information about the agents behaviour in a two-regime switching model (Vijverberg, 1993) and to obtain predicted out-of-sample distribution of the outcome gains (Poirier and Tobias, 2003).

Giorgio Calzolari
Università di Firenze, DISIA. “G. Parenti”, Viale Morgagni, 59, Firenze, email:
calzolar@disia.unifi.it

Antonino Di Pino
Università di Messina, Dipartimento S.E.A.M. Via T. Cannizzaro, 278, Messina, email:
dipino@unime.it
(We thank the financial support of the project MIUR PRIN MISURA - Multivariate Models for Risk Assessment)

“Indirect” estimates of this parameter are possible applying the relationships among the errors’ second-order moments estimated by *FIML* or *Two-Stage* procedures (Maddala, 1983 pp. 223-228). Another estimation approach is based on the assumption that across-correlation is determined by a latent factor (such as, for instance, the unobserved individual ability) common to both outcome equations and a third selection equation. In this case, the estimation procedure applies Factor Analysis methods (Carneiro et al., 2003; Aakvik et al. 2005, among others). However, since individuals cannot be observed jointly in both regimes, to identify a latent factor common to all the equations is not simple. The solution provided by analysts is to generate “counterfactuals” to match with the observed cases. As a consequence, to apply a matching procedure, some variables that generate the conditional independence, usually assumed in matching, should be introduced in the model as exclusion restrictions. Thus, identification of an across-regime parameter depends on variables conditioning matching results, not included in the outcome equations. Unlike these approaches, the aim of this study is to suggest a simple selection criterion in a two equations switching regression model that permits identification and “direct *FIML* estimation also” of the across-regime correlation. With respect to other *FIML* approaches using an additional selection equation (c.f. Poirier and Ruud, 1981; and Maddala, 1983 and 1986), our model specifies the two outcome equations only. With respect to methods based on a latent-factor identification, no exclusion restrictions are imposed applying our “Two-Equation *FIML*” procedure, and no matching procedure to generate counterfactuals are adopted. In our model, the selected regime is the one that produces the larger outcome.

Assuming that outcome can always be observed in one of the two regimes, our model can be theoretically specified as a switching regression model with “sample separation known” (cf. Maddala, 1986 for a survey). The agent is assumed to compare the outcomes of the two equations, and to choose the larger; thus, the larger between the two dependent variables is observed, the smaller is latent, but its value is “upper bounded” by the observed one. The model is therefore a sort of “two simultaneous censored equations” with endogenous censoring. For each individual, the contribution to the likelihood is given by the probability density of the observed variable (the larger) and by the (conditional) probability that the other variable has a smaller value: besides coefficients and variances, the likelihood includes therefore also the cross equations correlation. With respect to other *FIML* or Two-Stage (*T-S*) methods (such as the *T-S* Heckman or Control Function), there is no additional stochastic equation to select between the two regimes.

Using simulated data, we compare the performance of our “Two-Equation *FIML*” estimator with the traditional “Three-Equation *FIML*” and Two-Stage (*T-S Heckman*) methods in which the across regime correlation is not estimated directly. Monte Carlo experiments evidence the “good” performances of our estimator as far as the cross-equation correlation parameter is concerned. In the next section, the specification of our *FIML* model is provided. In the third section results of simulations, based on the use of Three-Equation *FIML* (Poirier and Ruud, cit.), *T-S* Heckman and our Two-Equation *FIML* procedures, respectively, are discussed.

2 Model Specification and Two-Equation *FIML* Estimator

Our model can be specified as follows:

$$W_{1i} = \mathbf{x}'_{1i} \boldsymbol{\beta}_1 + u_{1i} \quad W_{1i} >= W_{2i} \quad \text{then } W_{1i} \text{ is observed and } W_{2i} \text{ is latent.} \quad (1)$$

$$W_{2i} = \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + u_{2i} \quad W_{1i} <= W_{2i} \quad \text{then } W_{2i} \text{ is observed and } W_{1i} \text{ is latent} \quad (2)$$

With respect to the traditional Roy model, based on the assumption that the same regressors are included in each regime, in our approach the vectors \mathbf{x}'_{1i} and \mathbf{x}'_{2i} can include different regressors as well as the same regressors in each regime. The error terms u_{1i} and u_{2i} are normally distributed with zero mean and variances equal to σ_1^2 and σ_2^2 . We assume that the across-regime covariance of errors, σ_{12} , may be different from zero. The specification of our likelihood function is based on the probability of a subject to gain the outcome of the two regimes; it is the probability density of the observed variable, multiplied by the conditional probability that the other variable (latent) is smaller than the observed variable. Considering also the error terms, $u_{2i} = W_{2i} - \mathbf{x}'_{2i} \boldsymbol{\beta}_2$ and $u_{1i} = W_{1i} - \mathbf{x}'_{1i} \boldsymbol{\beta}_1$, as normally distributed, we obtain from the censoring rule in Eqs. (1) and (2) the following log-likelihood function:

$$\begin{aligned} \ln \mathbf{L} = & -\frac{(W_{1i} - \mathbf{x}'_{1i} \boldsymbol{\beta}_1)^2}{2\sigma_1^2} - \frac{1}{2} \ln \sigma_1^2 + \ln \Phi \left(\frac{(W_{1i} - \mathbf{x}'_{2i} \boldsymbol{\beta}_2) - \frac{\sigma_{12}}{\sigma_1^2} (W_{1i} - \mathbf{x}'_{1i} \boldsymbol{\beta}_1)}{\sqrt{\sigma_2^2 - \sigma_{12}^2 / \sigma_1^2}} \right) \\ & - \frac{(W_{2i} - \mathbf{x}'_{2i} \boldsymbol{\beta}_2)^2}{2\sigma_2^2} - \frac{1}{2} \ln \sigma_2^2 + \ln \Phi \left(\frac{(W_{2i} - \mathbf{x}'_{1i} \boldsymbol{\beta}_1) - \frac{\sigma_{12}}{\sigma_2^2} (W_{2i} - \mathbf{x}'_{2i} \boldsymbol{\beta}_2)}{\sqrt{\sigma_1^2 - \sigma_{12}^2 / \sigma_2^2}} \right) \end{aligned} \quad (3)$$

where $\Phi()$ is the standard normal cumulative distribution function utilized to specify the contribution to the likelihood of censoring of W_1 or W_2 .

3 Monte Carlo Results

Investigation by a Monte Carlo experiment is motivated by a desire to compare inferential properties of our “Two-Equation *FIML*” estimator, with the “indirect” estimators of the across-regime covariance, such as the Poirier-Ruud (1981) “Three – Equation *FIML*” and the *T-S Heckman* methods. The design here considered is generated by the model equations (1), (2), characterized by the inclusion in their right side of a single regressor variable (plus the constant). Slope and intercept coefficients are the same for both Eqs. (1) and (2). Then, in order to simulate the presence of cross correlation, we set the across-regime correlation, alternatively, with positive ($\rho_{12} = 90\%$) and negative sign ($\rho_{12} = -90\%$). To estimate with Poirier-Ruud and T-S Heckman methods, we include a third equation: a Probit selection equation where the explanatory variables are the same of eqs. (1) and (2). Consequently, taking into account the meaning of the relationships between the errors’ second-order moments of outcome and selection equations (cf. Heckman and Honoré, 1990; and Vijverberg, 1993), $\sigma_1^2 = \sigma_2^2$

implies null correlation between outcome and selection equation (absence of "endogenous selection"). Instead, when σ_1^2 and σ_2^2 are different (our results are obtained setting $\sigma_1^2 = 4\sigma_2^2$) a nonzero correlation between the outcome and the selection equation occurs (thus we may talk of "endogenous selection"). For the sake of brevity, we report in Table 1 the results of the estimated correlation parameters only. We utilize mean bias and root mean-square error (*RMSE*) to compare the performance of the estimators. The percentage of cases observed in each regime on the total of cases is symmetrically equal to 50%. With respect to other methods, the "Two-Equation *FIML*" procedure generally provides much more efficient estimates for both regression coefficients (not reported here) and across-regime correlation parameter (cf. Table 1). In addition, the "direct" estimate of the across-regime correlation appears to be always consistent, while the indirect estimates of ρ_{12} performed by *T-S* Heckman procedure, in absence of endogeneous selection and with $\rho_{12} = 90\%$ (strong correlation of outcomes across regimes), is biased (equal to 79,98%).

Table 1: Simulation results - Errors Distribution: Bivariate Normal; sample: 10000; NREP: 1000

	<i>Two-Equation FIML</i>		<i>T-S Heckman</i>		<i>Three-Equation FIML (Poirier-Ruud)</i>	
	coef	<i>RMSE</i>	coef	<i>RMSE</i>	coef	<i>RMSE</i>
$\rho_{12} = 90\%$ (positive)						
$\sigma_1^2 = \sigma_2^2$ Absence of endogenous selection	90.00%	0.00489	79.98%	0.007	89.99%	0.0049
$\sigma_1^2 = 4\sigma_2^2$ Endogenous selection	90.00%	0.00002	89.93%	0.0003	90.07%	0.0002
$\rho_{12} = -90\%$ (negative)						
$\sigma_1^2 = \sigma_2^2$ Absence of endogenous selection	-90.02%	0.0068	-89.93%	0.0597	-89.85%	0.0392
$\sigma_1^2 = 4\sigma_2^2$ Endogenous selection	-90.03%	0.00005	-90.16%	0.005	-89.90%	0.002

References

1. Aakvik, A., Heckman J. J., Vytlacil E. J.: Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *J. Econometrics* **125**, 15–51 (2005)
2. Carneiro, P., Hansen, K. T., Heckman, J. J.: Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *Int. Econ. Rev.* **44** (2), 361-422 (2003)
3. Heckman, J. J., Honoré, B. E.: The Empirical content of the Roy Model. *Econometrica* **58** (5), 1121-1149 (1990)
4. Maddala, G.S.: Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press. Cambridge (UK) (1983)
5. Maddala, G.S.: Disequilibrium, Self-Selection, and Switching Models. In Griliches, Z., Intriligator, M.D. (eds.) *Handbook of Econometrics*, Vol. III, pp. 1633-1688. Elsevier Science, North-Holland (1986)
6. Poirier, D.J. and P.A. Ruud: On the appropriateness of endogenous switching. *J. Econometrics* **16**, 249-256 (1981)
7. Poirier D. J., Tobias J. L.: On the predictive distributions of outcome gains in the presence of an unidentified parameter. *J Bus Econ Stat.* Vol.24 No 2. 258 -- 268 (2003)
8. Vijverberg W. P .M.: Measuring the unidentified parameter of the extended Roy Model of selectivity. *J Econometrics* **57**, 69-89 (1993)

Modern Vs. Traditional: A cluster-based specification of gender and familistic attitudes and their influence on the division of labour of Italian couples

Maria Gabriella Campolo, Antonino Di Pino and Ester Lucia Rizzi

Abstract The effect of transition to parenthood on the labour division of Italian couples is estimated adopting a Difference-in-Differences specification of simultaneous equations of paid and unpaid work. In addition, a cluster-based classification of couples is provided to identify the influence of gender and familistic beliefs on partners' division of labour. For the empirical analysis we use data on Italian couples provided by the Istat Multipurpose Panel Survey in years 2003 and 2007.

Key words: Gender attitudes, Difference-in-Differences, Clustering, Simultaneous Equations.

1 Introduction

Several studies found that the partners' allocation of paid and domestic work is affected by life-course events, and in particular, by the birth of a child ([2], i.a.). When trying to measure the impact of the transition to parenthood on the partners' allocation of time, one has to be aware that individual attitudes that concern gender roles and the family life play an important role. If attitudes are neglected, the consequence may be a bias in

Maria Gabriella Campolo and Antonino Di Pino
Department of Economics, Management, Environmental and Quantitative Methods , University of Messina, Via T. Cannizzaro, 278, Messina, Italy, e-mail: mgcampolo@unime.it, dipino@unime.it
Ester Lucia Rizzi
Research Centre on Population and Societies, Université Catholique de Louvain, Place Montesquieu 1, Louvain-la-Neuve, Belgium, e-mail: ester.rizzi@uclouvain.be

the estimation of the effect of transition to parenthood on paid work and domestic work. However, individual attitudes are difficult to observe, and their influence on partners' allocation of time is often misspecified [4].

In this study, we try to identify the impact of the gender and familistic attitudes using a dummy resulting from a cluster procedure that allows us to classify the couples in homogeneous groups according to their answer to normative statements. This dummy signals if a couple belongs to a more traditional or to a modern group in terms of gender and familistic attitudes. The dummy is inserted as a regressor in the paid and unpaid work equations estimated for both partners (four equations in total). In this way, the impact of the transition to parenthood on the partners' division of labour is estimated in association with attitudes through an interaction term.

In our study we use longitudinal data from the Istat Multipurpose Panel Survey for the years 2003 and 2007. A sample of Italian couples married or cohabiting is considered. Our results show that traditional attitudes, as to gender and the family life, negatively influence women paid work supply, while they are weakly associated to domestic activity.

2 Methodological Issues and Model Specification

In this analysis, we use an "index score" measuring the agreement of the subjects to statements regarding gender roles and the family life [4]. This index is obtained considering the individual level of disagreement/agreement with the following five statements: 1) marriage is an outdated institution; 2) a couple can live together without planning to marry; 3) a woman can have child alone even if she doesn't want a relation; 4) children aged 18-20 years old should leave parents' home; 5) it is right that unhappy spouses divorce, even if they have had children. Responses ranged from 1=strongly disagree to 5=strongly agree. We create a scalar index measuring individual agreement with the whole of statements by computing the mean of the scores for each subject. A Cronbach's alpha test (equal to .76) is used to test the reliability of the items scores [3]. We perform a cluster procedure applying the Ward algorithm, and taking into account the attitudes score index and a categorical variable signaling the level of agreement of partners with labour division as classification variables. The Calinski-Harabasz test is applied to choose the number of groups. In this way, a partition of the sample in two groups of couples is provided: one of "traditional" couples (who disagree with the five statements), and another of "modern" couples (who show a higher agreement with these). According to this bipartition, we obtain a dummy variable signaling if a couple belongs to a traditional or to a modern group. This dummy is introduced in the model as a regressor to identify the impact of attitudes on the partners' division of labour.

In order to correct for the endogeneity of fertility, we use the sex of previous children as the most relevant instrument in a reduced-form Poisson regression of the woman fertility (as number of children ever born). The validity of instruments is confirmed by the significance of estimated coefficients of Poisson regression (Girl as first child: $\beta=.24$, $p<0.01$; Girl as second child: $\beta=.31$, $p<0.001$; Boy as second child: $\beta=.53$, $p<0.001$ (the other coefficients are not reported here for the sake of brevity). At a second stage, the predicted fertility is included as an *IV* in each equation of the simultaneous-equation model (cf. Eq.1), where paid and unpaid work equations of both

partners are estimated simultaneously using an iterative *GLS* procedure. We apply an iterative *GLS* estimation procedure of our *DID* simultaneous equation model using residual-based estimation of the error covariance matrix to correct, respectively, for cross-sectional unobserved heterogeneity and for over-time bias effect. We specify our model by considering the time spent in paid and unpaid work (logarithm of weekly working hours, $\ln H$) by each partner as the dependent variable of four simultaneous equations. Let us denote with H_{kji} the time spent in paid work ($k=1$) and domestic work ($k=2$) by a subject i (a woman, if the subscript $j=w$, or a man, if $j=m$) at time t , with $t=0$ (2003) or $t=1$ (2007):

$$\ln H_{kji} = \mathbf{s}'_i \boldsymbol{\alpha}_{kj} + \lambda_{kj} t + t \mathbf{s}'_i \boldsymbol{\delta}_{kj} + \mathbf{x}'_i \boldsymbol{\beta}_{kj} + \mathbf{z}'_{it} \boldsymbol{\gamma}_{kj} + \varphi_{kj} \hat{C}h_i + u_{kji} \quad (1)$$

$\boldsymbol{\alpha}$ is a vector of coefficients measuring the impact of life-course events dummies included in the row vector \mathbf{s}' . Each (time invariant) dummy signals the status of the subject as a consequence of a specific life course event. In this study the event considered is the transition to parenthood. Consequently, $t\mathbf{s}'$ is a row vector whose elements are dummies that signal if status has changed over time, t . The scalar product between the vector $t\mathbf{s}'$ and the vector of coefficients $\boldsymbol{\delta}$, given by $t\mathbf{s}'\boldsymbol{\delta}$, measures the interaction effect of both status and time. The vector of the explanatory variables \mathbf{x} , whose corresponding parameters are $\boldsymbol{\beta}$, includes also the dummy variables measuring familistic and gender attitudes. The vector \mathbf{z} , whose parameters are $\boldsymbol{\gamma}$, indicates time-invariant and time-varying control variables. The control variables, referred to the subject or to the couple, are education, the area of residence, age, residence in an urban area (dummy) and hourly wage.

3 Estimation Results and Discussion

We use a two waves balanced panel sample of 956 couples drawn from the Istat Multipurpose Panel Survey in 2003 ($t=0$) and 2007 ($t=1$). Partners are married or cohabiting, employed or unemployed. In our sub-sample women are aged 18-45 and their partners are aged 18-60 in 2003.

The results of clustering procedure, discussed above, allow us to provide a dummy variable (indicated as *Attitude* in the model), equal to one for couples who express their agreement to traditional norms concerning gender role and the family life. This dummy is inserted as a regressor in paid and unpaid work equations. To better specify the interaction effect of gender attitudes and parenthood on both paid and unpaid work, we add the dummies *Birth*Attitude* and *t*Birth*Attitude* in the regressors set, where "*Birth*" is the dummy of transition to parenthood.

In Table 1 the *DID* estimated coefficients are presented for three different models: model 1 considers all orders of birth, model 2 refers to the first order of birth, and model 3 to the second order of birth. For the first and second birth (model 2 and 3), estimates of men's paid and unpaid work are non significant; for this reason we only report results referred to paid and unpaid work of women. Since our dependent variables are the natural logarithm of paid and unpaid work hours, the estimated coefficients measure the impact of each regressor on the dependent variable in percentage terms. This allows us to better evaluate the joint influence of two or more regressors by adding together the respective estimated coefficients. Considering the

effect of the birth of a child (whatever the order), we can evaluate the reduction of woman paid work by the sum of t and t^*Birth parameters, equal to -55% (-0.34 -0.21). In addition, the influence of parenthood transition on domestic work is positive and equal to 32% (0.46-0.14). Analogously, considering the transition to the first birth, traditional gender and familistic attitudes have a negative effect on paid work (-0.41+0.52-0.28= -17%), while the reduction of paid work due to the birth of the first child is equal to -21% (0.87-1.08). Instead, the impact of gender and familistic attitudes on woman domestic work is weak. For men, contrary to what happens for women, market and domestic working activities seem to be less sensitive to transition to fatherhood (all orders). With respect to the previous studies on intra-household labour division [1], our estimates confirm that Italian women are disadvantaged in both market and domestic work by the birth of a child; while attitudes affect the woman's paid work but not the domestic work.

Table 1: A GLS-DID procedure to estimate the partners' allocation of time by order of the new born. Only coefficients for time, birth, previous fertility and attitudes are shown.

RESULTS ALL ORDERS OF BIRTH		Model 1			
Dependent variable (log):		Women		Men	
Weekly working hours		Paid	Unpaid	Paid	Unpaid
<i>Transition to parenthood (all orders, no.: 956 couples; 1912 subjects; 165 transitions)</i>					
<i>Fertility</i> (instrumental variable)		-0.23 ***	0.20 ***	0.04	0.14 **
t (Dummy: 0=2003;1=2007)		-0.20 ***	-0.05	-0.01	0.35 ***
<i>Birth</i> (parenthood transition)		-0.21 *	-0.14	0.08	0.24
t^*Birth		-0.34 **	0.46 ***	-0.17	-0.38
<i>Attitude</i> (Dummy: traditional=1; modern=0)		-0.58 ***	0.09 **	0.05	-0.03
<i>Birth*Attitude</i>		0.29 **	-0.01	-0.07	-0.05
$t^* Birth *Attitude$		0.18	-0.06	0.11	0.62 **
RESULTS ONLY FOR WOMEN		Model 2 (112 couples)		Model 3 (302 couples)	
Dependent variable (log):		<i>Transit. to a 1st birth^a</i>		<i>Transit. to a 2nd birth^b</i>	
Weekly working hours		Paid	Unpaid	Paid	Unpaid
<i>Fertility</i> (instrumental variable)		-2.10 ***	0.68 *	-0.79 ***	-0.21
t (Dummy: 0=2003;1=2007)		-0.19	0.11	-0.06	-0.09
<i>Birth</i> (parenthood transition)		0.87 ***	-0.70 ***	-0.64 ***	0.06
t^*Birth		-1.08 ***	1.09 ***	-0.21	0.27
<i>Attitude</i> (Dummy: traditional=1;modern=0)		-0.41 ***	-0.01	-0.75 ***	0.10
<i>Birth*Attitude</i>		-0.28	0.54 **	0.68 ***	-0.15
$t^* Birth *Attitude$		0.52 *	-0.58 *	0.07	0.14

Note: p -value = * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ^a53 transitions; ^b77 transitions

References

1. Anxo D., Flood L., Mencarini L., Pailhé A., Solaz A., Tanturri M.L.: Gender differences in time use over the life course in France, Italy, Sweden, and the US. *Fem. Econ.* **17**(3), 159-195 (2011)
2. Baxter, J., Hewitt, B., Haynes, M.: Life course transitions and housework: marriage, parenthood, and time on housework. *J. Marriage Fam.* **70**, 259-272 (2008)
3. Cronbach, L.J., Shavelson, R.J.: My current thoughts on Coefficient Alpha and successor procedures. *Educ. Psychol. Meas.* **64**(3), 391-418 (2004) doi:10.1177/0013164404266386
4. Vella, F.: Gender roles and human capital investment: the relationship between traditional attitudes and female labour market performance. *Economica.* **61**(242), 191-211 (1994)

Clustering the Four Gospels in the Greek, Latin, Gothic and Old Church Slavonic Translations

Gabriele Cantaluppi, Marco Passarotti

Abstract We present an unsupervised and data-driven comparison of the four Gospels in their versions in four Indoeuropean languages: Greek, Latin, Old Church Slavonic and Gothic. Our method is based on lemmatised and syntactically annotated full texts and provides a lexical- and syntax-based comparison of the texts, by applying clustering and factor analysis techniques. Our results show the high degree of similarity of the different versions on the basis of the regularity of the translations.

Key words: Gospels, Hierarchical Clustering Analysis, Contribution Biplots

1 Introduction

It is a well known fact that three of the four Gospels in the New Testament (namely, those of Luke, Mark and Matthew) are similar to each other and differ from the Gospel of John. This is the reason why they are called the 'synoptic' Gospels, as they report events in a parallel fashion, while John deviates from the narration provided by the others.

This paper wants to provide a lexical- and syntax-based comparison of the four Gospels, by applying clustering and factor analysis techniques on their versions in Greek, Latin, Old Church Slavonic and Gothic. In particular, the paper aims to evaluate the degree of similarity of the different versions on the basis of the regularity of the translations.

Gabriele Cantaluppi

Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, e-mail: gabriele.cantaluppi@unicatt.it

Marco Passarotti

Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione, Università Cattolica del Sacro Cuore, Milano, e-mail: marco.passarotti@unicatt.it

2 Data and Method

The lemmatised and syntactically annotated texts of the four Gospels in their versions in Greek, Latin, Old Church Slavonic and Gothic were made available recently by the PROIEL corpus¹. We apply clustering and factor analysis techniques to each version of the Gospels in the different languages and, then, we compare the results, in order to evaluate the degree of regularity of the translations².

All the experiments were performed with the R statistical software. In particular, we apply the 'tm' package [2] to build and analyse the document-term matrices.

Clustering We use clustering in order to study the relations of similarity/dissimilarity between the four Gospels in each of the translations considered.

In particular, we perform two kinds of experiments, as we compare the texts by computing their distance in terms of similarity of the relative frequency of (a) the shared lemmas and (b) the syntactic functions³.

This means that texts that share a high number of lemmas with similar distribution and texts that feature a similar distribution of syntactic functions are considered to have a high degree of similarity and, thus, fall into the same or related clusters. For each translation, the analysis starts from the document-term matrix, \mathbf{X} , and builds clusters from the four rows of \mathbf{X} (each row corresponds to one Gospel).

Each row was first divided by its total to obtain relative frequencies and, thus, overcome the different size of the four texts concerned while clustering. A distance based on correlations was considered, by using a complete linkage hierarchical clustering method.

Factor Analysis While the clustering task computes and represents the similarity/dissimilarity value between the texts by clusters, it does not inform about which features distinguish one text from the other. These features are those properties that make two texts similar or dissimilar to each other.

Accordingly to the two kinds of experiments mentioned above, the features that we considered are lemmas and syntactic functions. In order to know which lemmas and syntactic functions distinguish one (or more) Gospel(s) from the other(s) in the different translations, we apply a factor analysis technique.

We compare the distribution of the features in the different translations by using a vector-based representation that results from factor analysis. The degree of similarity/dissimilarity between such representations corresponds to the degree of regularity of the translations.

¹ <http://www.hf.uio.no/ifikk/english/research/projects/proiel>.

² Among the huge amount of publications on (and approaches to) computing textual similarity, here we just report a few citations taken from the Proceedings of the recently held conference SEM-2012 and refer to them for further bibliography [1, 4].

³ While comparing the texts by lemmas, we excluded the function words by exploiting the PoS tagging provided by the morphological annotation of PROIEL (adverbs, articles, conjunctions, prepositions, pronouns). This reduced the size of the data of 50% on average in all the versions of the texts concerned. The resulting number of terms was respectively 2,506 for Latin, 2,807 for Greek, 2,167 for Gothic and 2,693 for Old Church Slavonic.

We follow the Principal Component Analysis biplots presentation by [3], p. 67, and produce 'contribution biplots'. Starting from an $I \times J$ term-document matrix, \mathbf{Y} (whose values had been previously standardised by column, in order to give the same importance to all documents), a reduction of the column space can be achieved by using Factor Analysis and considering factors which relate documents that feature high correlation in their standardised term distributions. The highest correlation occurs when two documents with the same profiles are concerned ('conditional' distributions of terms): in this case, the standardised distributions would result the same. In the opposite case, if two texts do not share any term, the corresponding standardised columns will have a negative correlation.

A weighted singular value decomposition (SVD) of $\mathbf{Y}/(IJ)^{1/2}$ is then performed

$$\mathbf{S} = \mathbf{Y}/(IJ)^{1/2} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}'$$

where \mathbf{U} and \mathbf{V} are matrices containing respectively the left and right singular vectors and \mathbf{D}_β is a diagonal matrix containing the singular values in decreasing order. The SVD allows the calculation of coordinates \mathbf{U} for terms and $\mathbf{G} = J^{1/2}\mathbf{V}\mathbf{D}_\beta$ for documents. By considering the first two columns of \mathbf{U} and \mathbf{G} , we have the coordinates with respect to the first two principal components (orthogonal factors).

The square of the elements in \mathbf{D}_β divided by their total inform about the amount of variance explained by the principal components. In all situations, we observed that the first component alone is able to explain over 90% of the variance.

By considering the squared values of the coordinates of terms we obtain their contribution to the principal axes.

3 Results and Discussion

Clustering It is not surprising that the Gospel of John is clustered apart from the three synoptic ones by both the lexical- and syntax-based analyses applied on all the translations concerned. What differs from one language to the others is the way the synoptic Gospels are clustered. Indeed, two of them are always clustered together against the third, but the most similar couple of Gospels does not remain the same in all the translations.

The most frequent configuration of clusters is such that the couple formed by the Gospels of Luke and Matthew belongs to one cluster separated from that of the Gospel of Mark. This configuration results from both the lexical-based and the syntax-based analysis of the Greek and Latin translations, and from the syntax-based analysis of the Gothic and Old Church Slavonic ones.

A few exceptions to this mainstream configuration do hold. Namely, they are the lexical-based analysis of (a) the Gothic translation, where the couple is formed by the Gospels of Luke and Mark, and (b) the Old Church Slavonic translation, according to which the texts of Mark and Matthew are separated from that of Luke. However, among these two exception, the latter is the only really meaningful one,

as the Gothic translation of the Gospel of Matthew is incomplete and this may influence the analysis.

The high degree of similarity between Luke and Matthew that we observed confirms the standard theory about the problem of the source relationships between the synoptic Gospels. This is a two-source hypothesis, according to which the origin of the Greek text of both Matthew and Luke (written between 80 and 90 CE) are the Gospel of Mark (~ 70 CE) and one hypothetical Sayings Source, usually referred to as the Q source (from German *Quelle*, i.e. 'source').

Factor Analysis According to the above results, only one factor should be extracted in all instances, as the amount of variance explained by the first dimension is always higher than 90%. Nevertheless, we decided to consider also a second factor, as it is able to characterise small (but meaningful) differences among the Gospels, although its amount of explained variance is always lower than 4%.

The percentage values of variability explained by the two factors are almost the same when the texts in Greek (0.943, 0.032), Latin (0.959, 0.023) and Old Church Slavonic (0.955, 0.021) are concerned. These values are slightly different for the Gothic translation (0.925, 0.037).

We observed that the overall distribution of lemmas is much similar in all the languages. In particular, the lemmas that mean *father*, *Jesus*, *jew*, *man*, *to believe*, *to know* and *world* occur in similar positions in all the figures.

However, there is a number of differences, which we resume in the following:

- the lemmas that mean *to make/to do* occur in pretty different positions in Greek and Latin;
- the position of the lemmas that mean *to be* and *to say* is much similar in Greek and Gothic and it is different from that in Latin and Old Church Slavonic;
- the lemma *to come* occurs in the same position in Greek, Latin and Old Church Slavonic, while it is slightly displaced in Gothic;
- the lemmas that mean *house*, *to see* and *to say* in Latin (*domus*, *video*, *aio*) and Old Church Slavonic (домъ, видѣти, рещи) appear in similar position, but they are missing in the figures of Greek and Gothic. The same holds for the lemmas that mean *day* in Gothic (dags) and Old Church Slavonic (дѣнь).

The distribution of syntactic functions is similar in all the versions of the Gospels.

References

1. Dinu G., Thater S.: Saarland: Vector-based models of semantic textual similarity. In: SEM 2012: The First Joint Conference on Lexical and Computational Semantics, pp. 603-607, Association for Computational Linguistics, Montréal, Canada (2012)
2. Feinerer, I., Hornik, K., Meyer, D.: Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5) 1–54. <http://www.jstatsoft.org/v25/i05/> (2008)
3. Greenacre, M.: *Biplots in Practice*, Fundación BBVA, (2010)
4. Yeh E., Agirre E.: SRIUBC: Simple Similarity Features for Semantic Textual Similarity. In: SEM 2012: The First Joint Conference on Lexical and Computational Semantics, pp. 617-623, Association for Computational Linguistics, Montréal, Canada (2012)

Regression Trees for change point analysis: methods, applications and recent developments

Carmela Cappelli and Francesca Di Iorio

Abstract The detection of change points in time series is a popular subject of research and the most challenging task is to identify multiple changes occurring at unknown date. At this aim [2] proposed a method called ART (Atheoretical Regression Trees) that employs least squares regression trees to detect multiple breaks in the mean. In this paper we present a review of the method as well as two recent extensions meant to deal either with changes in the coefficients of a parametric model or with changes in time series imprecisely observed i.e. time ordered observations whose values are not known exactly, such as interval or ordinal time series. An empirical example concerning a continuous-valued imprecise time series is presented and discussed.

Key words: Change point analysis, Regression trees, Imprecise time series

1 Section Heading

Change point analysis comprises various statistical tools which are employed for determining if and when a change in a data set has occurred. In the last two decades it has emerged as a relevant research topic both in the statistics and econometric literature. Indeed, the detection of change points is relevant from several points of view. First it can reveal a behavior of the time series that could otherwise be misunderstood and modeled inadequately, a well known example is the confusion between long memory and occasional breaks in mean that may lead to an erroneous identification of a fractionally integrated process. Second, in the context of forecasting detecting change points allows to improve the quality of the forecasting especially in case of long series covering extended periods of time. Eventually the identifi-

Dipartimento di Scienze Politiche, Università Federico II di Napoli , e-mail: carcappe@unina.it
e-mail: fdiorio@unina.it

cation of breaks might isolate shorter periods between longer ones, revealing the presence of outliers and thus the need for adjusting the data.

In the last decade most contributions have focused on detecting multiple breaks occurring at unknown dates. In case of multiple changes in mean [2] have proposed a method called Atheoretical Regression Trees (ART) that employs Least Square Regression Trees to estimate the number and location of multiple change points. Extensive simulation studies, comparison with current methods and applications to various real time series have provided evidence of the usefulness of the approach (see [4]).

In this paper we present two recent extensions of ART. The first one ([1]), called Theoretical Regression Trees (TRT) has been introduced for locating changes in the coefficient of a parametric model considering the general framework of the linear model. The second one ([3]) addresses the case of time series which are imprecisely or vaguely observed (so forth denoted imprecise time series) where the imprecision is assumed to be represented by means of fuzzy sets giving rise to fuzzy time series. In both cases the recursive partitioning principle of regression trees is exploited considering either model residuals or alternative measures of deviation. The applicative effectiveness of the proposed approach is illustrated by an application to an imprecise continuous-valued time series.

2 Regression trees for change points analysis

The problem of detecting multiple changes occurring at unknown dates can be synthesized as follows. Let y_t be an observed time series of length T characterized by m change points (T_1, \dots, T_m) . A common estimation method of the set of unknown break dates is that based on the least square principle

$$(\hat{T}_1, \dots, \hat{T}_m) = \operatorname{argmin}_{(T_1, \dots, T_m)} SSR(T_1, \dots, T_m) \quad (1)$$

where $SSR(T_1, \dots, T_m)$ denotes the sum of squared residuals of the partition.

As shown in Cappelli *et al.* (2008) the set of unknown change points can be recursively estimated by means of regression trees where each node h i.e. a subsample of observations, is split into its left and right descendants h_l and h_r to minimize the sum of squared residuals:

$$SSR(h_l) + SSR(h_r) \quad (2)$$

Thus the splitting criterion (2) corresponds to objective function (1) computed for a binary partition. Upon a proper specification of the sum of squared residuals different types of changes can be estimated. In particular we distinguish three cases: changes in mean where $SSR(h_l) + SSR(h_r) = \sum_{g \in \{l, r\}} \sum_{y \in h_g} (y - \hat{\mu}(h_g))^2$, changes in the coefficient of a parametric model where $SSR(h_l) + SSR(h_r) = \sum_{g \in \{l, r\}} \sum_{y_t \in h_g} (y_t - x_t' \hat{\beta}(h_g))^2$ and eventually changes in an imprecise time series where

$$\begin{aligned}
SSR(h_l) + SSR(h_r) = & \sum_{g \in \{l,r\}} [3 \sum_{t=1}^{T(h_g)} (c_t - \bar{c}(h_g))^2 + \\
& - 2\lambda \sum_{t=1}^{T(h_g)} (c_t - \bar{c}(h_g))(l_t - \bar{l}(h_g)) + \lambda^2 \sum_{t=1}^{T(h_g)} (l_t - \bar{l}(h_g))^2 \\
& + 2\rho \sum_{t=1}^{T(h_g)} (c_t - \bar{c}(h_g))(u_t - \bar{u}(h_g)) + \rho^2 \sum_{t=1}^{T(h_g)} (u_t - \bar{u}(h_g))^2] \quad (3)
\end{aligned}$$

This latter case relies on the parametrization of the imprecise time series in the form of a LR *fuzzy time series* $\tilde{y}_t \equiv (c_t, l_t, u_t)_{LR}$ with a proper membership function, where c_t denotes the center at time t , l_t and u_t the left and right *spreads* and λ and ρ are parameters that control the form of the membership function. The deviation measure (3) is based on the metric of Yang-Ko ([5]) that satisfies the decomposition property.

3 Empirical example

In order to illustrate the applicative usefulness of the proposed approach we present an empirical application considering the case of a real-valued imprecise time series. We have analyzed the time series of the temperatures collected in Rome during the year 1999, focusing on the subperiod 01:July-30:October, thus the length of the series is $T = 122$. The original data set provides hourly values of the temperature whose conversion into a single value for each day entails loss of information and inaccuracy. In order to overcome these drawbacks, following [6], we have generated a fuzzy variable with LR membership function defining the centers as the daily mean temperatures whereas the left and right spreads are obtained by averaging the hourly deviations of the values lower and higher than the mean, respectively. The time series of the centers and upper and lower bounds based on the right and left spreads are depicted in panel (a) of Figure 1.

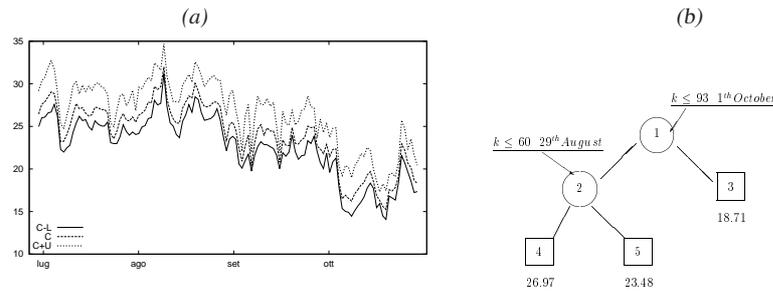


Fig. 1 (a): Fuzzy time series of the temperatures in Rome, 1 July - 30 October 1999. (b): Tree diagram of the temperature change points

The graph suggests the presence of two breaks; setting a minimum segment length of 15 observation we have detected two changes at dates 1-th of October and 29-th of August, a third change (identified but not chosen) is located at date 4-th of August. Panel (b) of Figure 1 provides the graphical representation of the tree partition corresponding to the change points whereas in Table 1 are reported some stylized facts of the entire series and the sub-periods defined by the change points.

Table 1 Stylized facts of the entire series and of the subperiods identified by the change points

	\bar{c}	\bar{l}	\bar{u}	min	max
<i>Entire series</i>					
1:July-30:October	24.0	1.45	2.60	9.2	40.4
<i>Regimes</i>					
01:July-29:August	26.9	1.50	2.60	18.7	40.4
30:August-01:October	23.5	1.47	2.73	16.6	36.3
02:October- 30:October	18.7	1.33	2.48	9.2	31.2

The tree diagram and the values in Table 1 show that the first change point identified by the procedure that separates July, August and September from October is the strongest one and it is associated with decreasing temperature and lower variability of both temperatures above and below the center (mean). The second change separates the hottest months July and August from September that is milder than the previous ones but characterized by higher variability of the temperatures above the mean. The results confirm that the proposed approach represents a useful tool to investigate the presence of change points also in imprecise time series.

Acknowledgements The authors wish to thank Dipartimento di Scienze Politiche for financial support.

References

1. Cappelli C., Di Iorio F. : Theoretical Regression Trees: a tool for multiple structural-change models analysis. In: Grigoletto M., Lisi F., Petrone S.(eds), *Complex Models and Computational Methods in Statistics*, Springer, 63–76 (2013).
2. Cappelli C., Penny R., Rea W. , Reale M. : Detecting multiple mean breaks at unknown points in official statistic. *Math Comput. Simulation*, **78**, 351–356 (2008).
3. C. Cappelli, D’Urso P., Di Iorio F.: Change point analysis of imprecises time series. *Fuzzy Sets Syst.*, **225**, 23–38 (2013).
4. Rea W., Reale M., Cappelli C., Brown J.A.: Identification of changes in mean with regression trees: an application to market research. *Econometric Rev.*, **29**, 754–777 (2010).
5. Yang M.S., Ko, C.H.: On a class of fuzzy c-numbers clustering procedures for fuzzy data. *Fuzzy Sets Syst.* **84**, 49–60 (1996).
6. Coppi, R., D’Urso P., Giordani, P., Santoro, A. : Least Square estimation of a linear regression model with LR fuzzy response, *Computational statistics & Data Analysis*, **51** 267–286 (2006).

Bayesian Stochastic Correlation Models

Roberto Casarin and Marco Tronzano and Domenico Sartore

Abstract This paper applies a Bayesian multivariate stochastic correlation model to the detection of correlation regimes in exchange rates. We follow a MCMC approach to parameter and latent variable estimation and provide evidence of significant differences between volatility and correlation dynamics.

Key words: Bayesian Inference, Multivariate Stochastic Volatility, Markov-switching, Stochastic Correlation

1 Introduction

Modelling and forecasting contagion between financial markets are crucial and challenging issues in financial management. The time-variations in the financial return volatilities and in the correlations between returns are two of the most relevant features of the contagion dynamics. In dynamic volatility modelling there are two main streams of literature: GARCH models and Stochastic Volatility (SV) models (e.g., [5]). Earlier contributions in these areas focused on univariate time series modelling. Subsequently, the attention has shifted to multivariate models with dynamic covariances (e.g., see [4]).

The aim of this paper is twofold. First, we provide a joint estimation of the return mean, volatility and correlation of exchange rates. Secondly, we

Roberto Casarin
University Ca' Foscari of Venice, e-mail: r.casarin@unive.it

Marco Tronzano
University of Genova, e-mail: m.tronzano@mclink.it

Domenico Sartore
University Ca' Foscari of Venice, e-mail: sartore@unive.it

investigate the presence of independent shifts in the volatility and correlation dynamics. In this sense we extend the empirical findings for exchange rates due to [1].

The structure of the paper is as follows. Section 2 introduces the stochastic correlation model. Section 3 describes briefly the Bayesian inference approach used. Section 4 presents the results for three daily exchange rate series. Section 5 concludes.

2 A Markov-switching Stochastic Correlation Model

Let $\mathbf{y}_t = (y_{1t}, \dots, y_{mt})' \in \mathbb{R}^m$ be a vector-valued time series, representing the log-differences in the spot exchange rates, $\mathbf{h}_t = (h_{1t}, \dots, h_{mt})' \in \mathbb{R}^m$ the log-volatility process, $\Sigma_t \in \mathbb{R}^m \times \mathbb{R}^m$ the time-varying covariance matrix, and $s_{1,t} \in \{0, 1\}$ a two-states Markov chain. We consider here a special case of the stochastic correlation model (*MSSC*) given in [3]

$$\mathbf{y}_t = \mathbf{a}_{00} + \mathbf{a}_{01}s_{1,t} + (A_{10} + A_{11}s_{1,t})\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma_t) \quad (1)$$

$$\mathbf{h}_t = \mathbf{b}_{00} + \mathbf{b}_{01}s_{1,t} + (B_{10} + B_{11}s_{1,t})\mathbf{h}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}_m(\mathbf{0}, \Sigma_\eta) \quad (2)$$

with $\boldsymbol{\varepsilon}_t \perp \boldsymbol{\eta}_s \forall s, t$, and $\mathcal{N}_m(\boldsymbol{\mu}, \Sigma)$ the m -variate normal distribution, with mean $\boldsymbol{\mu}$ and covariance matrix Σ , and $\mathbf{a}_{00}, \mathbf{a}_{01}, A_{10}, A_{11}, \mathbf{b}_{00}, \mathbf{b}_{01}, \mathbf{b}_{10}$ and \mathbf{b}_{11} parameters to be estimated. The probability law governing $s_{1,t}$ is $s_{1,t} \sim \mathbb{P}(s_{1,t} = j | s_{1,t-1} = i) = p_{1,ij}$, with $p_{1,ij}, ij \in \{0, 1\}$. As regards the conditional covariance matrix Σ_t , we use the decomposition (see [2]):

$$\Sigma_t = \Lambda_t \Omega_t \Lambda_t, \quad (3)$$

with $\Lambda_t = \text{diag}\{\exp\{h_{1t}/2\}, \dots, \exp\{h_{kt}/2\}\}$, a diagonal matrix with the log-volatilities on the main diagonal and $\Omega_t = \tilde{Q}_t^{-1} Q_t \tilde{Q}_t^{-1}$ the stochastic correlation matrix with $\tilde{Q}_t = (\text{diag}\{\text{vecd } Q_t\})^{1/2}$ and $Q_t^{-1} \sim \mathcal{W}_m(\nu, S_{t-1})$ where:

$$S_{t-1} = \frac{1}{\nu} Q_{t-1}^{-d/2} \bar{Q}_t Q_{t-1}^{-d/2}, \quad \bar{Q}_t = [\lambda_{s_{2,t}} \bar{D}_{s_{1,t}} + (1 - \lambda_{s_{2,t}}) I_m],$$

$$\bar{D}_{s_{1,t}} = \sum_{k=0,1} \mathbb{I}_{\{k\}}(s_{1,t}) \bar{D}_k$$

and $\bar{D}_k, k \in \{0, 1\}$, is a sequence of positive definite matrices, which capture the long-term dependence structure between series in the different regimes, and d is a scalar parameter. The correlation-switching process $s_{2,t} \in \{0, 1\}$ has transition probability: $s_{2,t} \sim \mathbb{P}(s_{2,t} = j | s_{2,t-1} = i) = p_{2,ij}$.

3 Bayesian Inference

Define $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_T)'$, and $\mathbf{z} = (\mathbf{h}, \mathbf{q}, \mathbf{s}_1, \mathbf{s}_2)$, with $\mathbf{h} = (\mathbf{h}'_0, \dots, \mathbf{h}'_T)'$, $\mathbf{s}_k = (s'_{k,0}, \dots, s'_{k,T})'$, $k = 1, 2$, and $\mathbf{q} = (\text{vech}(Q_0)', \dots, \text{vech}(Q_T)')$. The complete-data likelihood function of the MSSC model is:

$$\begin{aligned} \mathcal{L}(\mathbf{y}, \mathbf{z} | \boldsymbol{\theta}) = & \quad (4) \\ & \prod_{t=1}^T \left(\frac{1}{(2\pi)^{m/2} |\Sigma_t|^{1/2}} e^{-\frac{1}{2} \boldsymbol{\varepsilon}'_t \Sigma_t^{-1} \boldsymbol{\varepsilon}_t} \frac{1}{(2\pi)^{m/2} |\Sigma_\eta|^{1/2}} e^{-\frac{1}{2} \boldsymbol{\eta}'_t \Sigma_\eta^{-1} \boldsymbol{\eta}_t} \right. \\ & \cdot 2^{-\frac{m\nu}{2}} \Gamma_m(\nu/2)^{-1} |S_{t-1}|^{-\frac{\nu}{2}} e^{-\text{tr}(\frac{1}{2} S_{t-1}^{-1} Q_t^{-1})} |Q_t^{-1}|^{\frac{\nu-m-1}{2}} \\ & \cdot \left. \prod_{k=1,2} \left(p_{k,00}^{1-s_{k,t}} (1-p_{k,00})^{s_{k,t}} \right)^{1-s_{k,t-1}} \left(p_{k,01}^{s_{k,t}} (1-p_{k,01})^{1-s_{k,t}} \right)^{s_{k,t-1}} \right), \end{aligned}$$

where $\Gamma_m(\nu/2)$ is the m -variate gamma function and $\boldsymbol{\theta} = (\mathbf{a}'_{00}, \mathbf{a}'_{01}, \text{vec}(A_{10})', \text{vec}(A_{11})', \mathbf{b}'_{00}, \mathbf{b}'_{01}, \text{vec}(B_{10})', \text{vec}(B_{11})', \text{vech}(\Sigma_\eta)', \nu, d, \lambda_1, \text{vech}(\bar{D}_0), \text{vech}(\bar{D}_1), p_{1,11}, p_{1,22}, p_{2,11}, p_{2,22})'$ is the parameter vector. We arrange $\boldsymbol{\theta}$ in four sub-vectors: $\boldsymbol{\theta}_1 = \text{vec}(\Psi)$, with $\Psi = (\psi_1, \dots, \psi_m)$, which has in the columns the vectors $\psi_j = (a_{00,j}, a_{01,j}, (A_{10,j1}, \dots, A_{10,jm}), (A_{11,j1}, \dots, A_{11,jm}))'$, $j = 1, \dots, m$; $\boldsymbol{\theta}_2 = (\boldsymbol{\phi}', \text{vech}(\Sigma_\eta)')$, with $\boldsymbol{\phi} = \text{vec}(\Phi)$, where $\Phi = (\phi_1, \dots, \phi_m)$ has in the columns the vectors $\phi_j = (b_{00,j}, b_{01,j}, (B_{10,j1}, \dots, B_{10,jm}), (B_{11,j1}, \dots, B_{11,jm}))'$, $j = 1, \dots, m$; $\boldsymbol{\theta}_3 = (\nu, d, \lambda_1, \text{vech}(\bar{D}_0), \text{vech}(\bar{D}_1))'$; $\boldsymbol{\theta}_4 = (p_{1,00}, p_{1,11}, p_{2,00}, p_{2,11})'$. We specify the following prior distributions

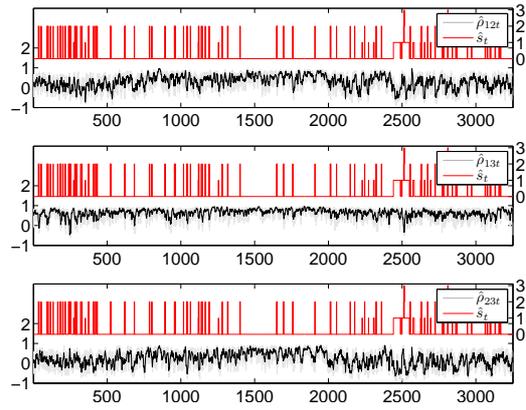
$$\begin{aligned} \boldsymbol{\theta}_1 & \sim \mathcal{N}_{24}(\mathbf{0}, 10I_{24}), \boldsymbol{\phi} | \Sigma_\eta \sim \mathcal{N}_{24}(\mathbf{0}, \Sigma_\eta \otimes 10I_8), \Sigma_\eta^{-1} \sim \mathcal{W}_3(10, 4I_3) \\ d & \sim \mathcal{U}_{(-1,1)}, \lambda_1 \sim \mathcal{U}_{(0,1)}, \nu \sim \frac{1}{\Gamma(10)} (\nu-3)^{10-1} \exp\{-\nu-3\} \mathbb{I}_{\{3,+\infty\}}(\nu) \\ \bar{D}_i^{-1} & \sim \mathcal{W}_3(10, 0.1I_3), p_{k,ii} \sim \mathcal{U}_{(0,1)}, k = 1, 2, i = 0, 1 \end{aligned}$$

We apply the Gibbs sampler given in [3] for the posterior approximation.

4 Exchange Rates Correlation Dynamics

We consider daily closing values for three exchange rates against the US\$, namely Euro, Yen and Pound. We compute the percentage log-returns of the exchange rates and denote them as $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$ for Euro, Yen and Pound, respectively. We fit the proposed Bayesian MSSC model on the exchanger rate dataset. The estimation results given in Fig. 1 show the presence of significant shifts in both volatilities and correlations. Moreover, the stepwise line in Fig. 1 indicates the presence of correlation-specific shifts (i.e., $\hat{s}_t \in \{2, 3\}$), thus suggesting the coexistence of different configurations of risk (volatility) and

Fig. 1 Posterior means (solid lines, left axes) and 95% credibility regions (gray areas, left axes) of the correlation Ω_t . Each figure includes $\hat{s}_t = \hat{s}_{1,t} + 2\hat{s}_{2,t}$ (stepwise, right axes).



contagion level (correlation) in the exchange rate markets analysed in this paper.

5 Conclusion

We apply a stochastic correlation model to detect changes in the correlations between exchange rates sampled at a daily frequency. We follow a Bayesian inference approach and find coexistence of different volatility and correlation-specific regimes.

Acknowledgements This research is supported by the Italian Ministry of Education, University and Research (MIUR) PRIN 2010-11 grant, and by the European Commission Collaborative Project SYRTO.

References

1. Asai, M., McAleer, M.: The Structure of Dynamic Correlations in Multivariate Stochastic Volatility Models, *Journal of Econometrics*, **150**, 182–192 (2009)
2. Bollerslev, T.: Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Approach, *Review of Economics and Statistics*, **72**, 498–505 (1990)
3. Casarin R., Tronzano, M., Sartore, D.: Bayesian Markov Switching Stochastic Correlation Models, Working paper, University Ca' Foscari of Venice (2013)
4. Engle, R.: Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models, *Journal of Business and Economic Statistics*, **20**, 339–350 (2002)
5. Jacquier, E., Polson, N. and Rossi, P. (1994) Bayesian Analysis of Stochastic Volatility Models, *Journal of Business and Economic Statistics*, **12**, 281–300 (1994).

Evaluating the selection effect in labour markets with a low female participation

Rosalia Castellano, Gennaro Punzo and Antonella Rocca

Abstract *The aim of this paper is to explore the main determinants of women's job search propensity as well as the mechanism underlying the selection effect across the four European countries (Italy, Greece, Hungary and Poland) with the lowest female labour force participation. The potential bias due to the overlap in some unobserved characteristics is addressed via a bivariate probit model. Significant selection effects of opposite signs are found for the Greek and Polish labour markets.*

Keywords: *female labour propensity, heckman correction, cross-country analysis*

1 Introduction

Over the last few decades the female labour force participation has been increasing throughout Europe. Nevertheless, the female activity rates are still consistently lower than their male counterpart everywhere; as they say, women are less likely than men to be employed or looking actively for a job. In 2007, the female participation rates in labour market largely varied across European countries (epp.eurostat.ec.europa.eu), from the lowest values of Southern – i.e., Italy (50.7) and Greece (54.9) – and some Eastern countries – i.e., Hungary (55.1) and Poland (56.5) – to the highest incidence for the well-developed economies of North Europe – i.e., Iceland (82.7), Sweden (76.8) and Denmark (76.4) against a EU-27 average of 63.2 per cent.

As documented [5], the usually-substantial extent of female non-participation in labour market may cause problems of sample selection because working women could be unrepresentative of the entire female population. Indeed, beyond differences in male and female behaviours in labour force participation, women who do not work may differ in some important *unmeasured* ways (i.e., individual status, family-specific

¹ R. Castellano, G. Punzo and A. Rocca, Department of Quantitative and Business Studies, University of Naples "Parthenope", e-mail: (castellano, punzo, rocca@uniparthenope.it).

or socio-cultural background) from women who choose to belong to labour market and this may even lead to biased estimates of structural parameters relevant to the behaviour of working women. For example, in the classical wage equations, it is likely that women's earnings are biased because women who are working form a self-selected (and *not* a random) sub-sample. The two-stage Heckman procedure [3] is the most used method to correct for this selectivity; in the first step, female labour propensities are estimated on a set of women's characteristics through a probit model which provides the correction term (λ), equal to the inverse of Mill's ratio, to include as additional predictor in the original regression equation.

The idea of this work was inspired by our empirical results of the women's wage equations tested over 26 EU-countries through the Heckman procedure (*not reported for brevity*). Lambda coefficients, consistently significant and negatively signed for each country (except for Czech Republic and Norway), suggest an inverse correlation between the error terms of selection probit and primary wage models. It means that unobserved factors, which make labour participation more likely, tend to be associated with lower potential returns. In this field, the paper aims at exploring the main determinants of women's job search propensity as well as the mechanism underlying the selection effect across the four EU-countries with the lowest female participation in labour market. More precisely, in the light of country-specific peculiarities of labour markets, gender equality and conciliation policies, the analysis compares countries with a different stage of economic development, say the Southern countries of Italy and Greece with the Eastern transition economies of Hungary and Poland.

2 Background and methodology

The patterns of female labour force and their changes over time arise from the interaction of institutional, cultural and socio-economic dynamics [4]. However, although Italy, Greece, Hungary and Poland strongly differ in terms of labour market flexibility, level of tertiarization of the economy, women's participation in higher education programs and policies for connecting work and family, these countries share low rates of female labour participation. In particular, in Italy and Greece, where the decline of marriages and the increase of births outside marriage undermined the male breadwinner model, the transition from care force to workforce has still weak social supports for childcare. Moreover, the higher levels of income inequalities and public debts may distract Governments from adequate gender equality policies which are officially in force but not very actively pursued. In Poland and Hungary, the female labour participation and gender pay gaps – which appeared on the surface like Nordic countries during the socialist-type regime whose policies strongly encouraged women to work – worsened for the period of transition.

Exploring 2007 EU-SILC data, the women's propensity to work (number of women actively looking for a job on the amount of unemployed women) was higher than 70% for Greek, Italian and Hungarian women, while just in Poland females' propensity was lower than their male counterpart. However, current levels of female participation in labour market strongly affect who is actively looking for a job; thus, in order to control for the potential overlap in unobserved characteristics influencing both the women's propensity to work and their propensity to look actively for a job, a

bivariate probit model is estimated [2]. The first probit model estimates the probability that a woman is not occupied:

$$y_i^* = X_i^F \gamma + v_i^F \quad \text{with} \quad v_i^F \sim N(0, \sigma_v^2) \quad (1)$$

where the latent variable y_i^* drives the observed outcome of not working y_i through the following measurement equation:

$$y_{if} = 1 \quad \text{if} \quad y_{if}^* > 0 \quad \text{and} \quad y_{if} = 0 \quad \text{if} \quad y_{if}^* \leq 0 \quad (2)$$

Focusing on the subset of women who does not work, the probability of being actively searching a job is given by:

$$S_1^* = X_i^F \gamma + W_i^F \delta + \varepsilon_i^F \quad \text{with} \quad \varepsilon_i^F \sim N(0, \sigma_\varepsilon^2) \quad (3)$$

including additional covariates (W) concerning the equalized household income and size, the individual health status and geographical area.

In this way, the potential for unobserved heterogeneity that could produce a correlation between error terms of the two probit models is considered. Therefore, not only the true effects of searching a job, but also the effect on professional condition of having these unobservable characteristics are captured [1]. If the error terms v_i and ε_i , jointly distributed as bivariate normal with zero means and unit variances, are significantly positive correlated ($\rho > 0$), unobserved factors increase both the probability of being a not-occupied female and looking for a job; for significantly negative ρ , the reverse is true, while not significant ρ shows the absence of selection effect and the equivalence of using the bivariate or two separate probit models.

3 Main results

Significant selection effects of opposite signs are found for Greece and Poland (tab. 1) while in Italian and Hungarian labour markets the non-sample selection could derive from a lack of link between the mechanisms of job search and the status of unemployed. In this light, the significance of lambda coefficients for the Heckman correction in the women's wage equations could denote a sample selection which exclusively involves women that do not participate at all to labour force. The harsh Greek scenario and the difficulties to find a job drive both the propensity of being not occupied and negatively the propensity of actively seeking employment. In Poland, the unmeasured factors associated to a lower propensity to search a job act in the opposite direction. While in all countries a higher female propensity to work concerns families where more members are already occupied, just in Poland, the women's job search propensity appears to be not linked to financial household problems or marital status; anyway, Polish high-educated women are less likely to be looking for a job. Finally, although sub-national differences occur (women living in the North-West or Centre of Italy and in the North of Hungary are more likely to be actively searching), the presence of children discourage women to be active in the labour market everywhere.

Table 1: Bivariate probit estimates of not working and actively searching for a job for females

Variables	Italy	Greece	Hungary	Poland
<i>Actively searching for a job</i>				
Intercept	- 1.2127***	- 1.7942***	- 1.4347***	- 1.3072***
Equivalised household income	- 2.2E-5**	- 2.5E-5***	- 0.0002***	- 4.19E-5
Marital status (1 if <i>married</i>)	- 0.3349***	-0.5128***	-0.2277*	0.0395
Education attainment (ref: <i>low</i>)				
<i>Medium</i> (ISCED97: 3;4)	0.2387***	0.4134***	0.2798***	0.5373***
<i>High</i> (ISCED97: 5)	0.6332***	1.1202***	0.6385***	- 0.2661***
Children (1 if <i>with children</i>)	- 0.3355***	- 0.6085***	- 0.2834**	- 0.4220***
Age class (ref.: 16-24 years)				
<i>Younger</i> [25-40 years]	0.3672***	0.4762***	- 0.3924*	0.4272***
<i>Older</i> [41-65 years]	- 0.4022***	- 0.0109	0.9193***	- 0.2661***
Health (1 if <i>chronic</i>)	- 0.0598	- 0.0731	0.7179***	- 0.3056***
Ratio ^(*)	0.2999	1.8186***	1.0821**	0.4861*
Equivalised household size	0.1588***	0.2206***	0.1319**	0.0952**
Urbanisation degree (1 if <i>densely</i>)	- 0.0576	- 0.0773	0.0446	0.0786
Geographical area (NUTS1) ^(**)				
Area 1	0.2062*	0.2018	- 0.2331**	- 0.0181
Area 2	0.1085	0.0279	- 0.1726**	- 0.2322***
Area 3	0.1532*	0.0658	–	- 0.0186
Area 4	- 0.0891	–	–	- 0.0837
Area 5	–	–	–	0.0569
<i>Not working</i>				
Intercept	3.5105***	3.5638***	3.8445***	3.2892***
Age (years)	- 0.0092***	- 0.0115***	- 0.0140***	- 0.0016
Marital status (1 if <i>married</i>)	0.0962**	0.0677	- 0.2143***	- 0.3350***
Children (1 if <i>with children</i>)	0.5082***	0.4875***	0.4473***	0.2833***
Urbanisation degree (1 if <i>densely</i>)	0.1256***	0.1002**	0.0352	0.0871***
Educational attainment (ref: <i>low</i>)				
<i>Medium</i> (ISCED97: 3;4)	- 0.4432***	- 0.4274***	- 0.5901***	- 0.8561***
<i>High</i> (ISCED97: 5)	- 0.9438***	- 0.9553***	- 0.9686***	- 1.4804***
Ratio ^(*)	- 4.4691***	- 4.3516***	- 4.3045***	- 3.5013***
Wald chi ²	429.33	317.55	188.26	313.94
Correlation (ρ)	- 0.2390	0.6858**	0.1859	- 0.3482**

^(*) (n° wage earners– 1)/(n° household members); * significant at 10%; ** 5%;*** at 1%

^(**) NUTS1 codes: *Italy*: 1 North-West, 2 North-East, 3 Centre, 4 South (ref.: Isles); *Greece*: 1 Voreia, 2 Kentriki, 3 Attiki (ref.: Nisia Aigaiou, Kriti); *Hungary*: 1 Central, 2 Transdanubia (ref.: Greath Plain and North); *Poland*: 1 Centralny, 2 Poludniowy, 3 Wschodni, 4 Polnocno-Zachodni, 5 Poludniowo-Zachodni (ref.: Polnocny)

References

1. Fleming C.M., Kler P. (2011), I'm too clever for this job: a bivariate probit analysis on overeducation and job satisfaction in Australia, *Applied Economics*, 40 (9), 1123-1138.
2. Green W.H. (1997), *Econometric Analysis*, Prentice Hall.
3. Heckman J. (1979), Sample Selection Bias as a Specification Error, *Econometrica*, 47(1).
4. Jaumotte F. (2003), Female Labour Force Participation: Past Trends and Main Determinants in OECD Countries, *OECD Economics Department Working Papers*, 376, OECD Publishing
5. Maddala G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

A statistical based H index for the evaluation of e-markets

Paola Cerchiello and Paolo Giudici

Abstract The measurement of the quality of academic research is a rather controversial issue. Recently Hirsch has proposed a measure that has the advantage of summarizing in a single summary statistics all the information that is contained in the citation counts of each scientist. From that seminal paper, a huge amount of research has been lavished, focusing on one hand on the development of correction factors to the h index and on the other hand, on the pros and cons of such measure proposing several possible alternatives. In the present work we propose an exact statistical approach to derive the distribution of the h index. To achieve this objective we work directly on the two basic components of the h index: the number of produced papers and the related citation counts vector, by introducing convolution models. What proposed has been applied to a database of homogeneous scientists in a recent paper. Here we extend what proposed in an application that concerns the evaluation of e-market performances. The results show that the approach is able to rank well e-market places on the basis of an H index based on the number of companies present in the web and their number of orders.

Key words: h index, discrete extreme value models, convolution models, e-markets

1 Foreword

The measurement of research achievements of scientists has received a great deal of interest, since the paper of Hirsch (2005) that has proposed a "transparent, unbiased

Paola Cerchiello

Dep. Economics and Management, University of Pavia, via S. Felice 5, Pavia, e-mail: paola.cerchiello@unipv.it

Paolo Giudici

Dep. Economics and Management, University of Pavia, via S. Felice 5, Pavia e-mail: giudici@unipv.it

and very hard to rig measure” (Ball, 2005): the h index. The Hirsch definition is that “a scientist has index h if h of his or her N_p papers have at least h citations each and the other (N_p-h) papers have $\leq h$ citations each”.

Following the seminal work of Hirsch many papers have dwelled on that issue, especially in the bibliometric community. Surprisingly, few papers have focused on the statistical aspects behind the H-index, apart from Glanzel (2006) that hinted at the relevance of a “statistical background” for the h-index. Recently Pratelli et al. (2012) and Beirlant and Einmahl (2010) have proposed an asymptotic distribution for the h index that can be used for inferential purposes and not only for descriptive summaries as in the typical bibliometric contributions. Our contribution follows such recent papers, with the aim of providing a statistical framework for the h index that, in addition, holds also for small sample size and respects the discrete nature of the bibliometric data at hand.

Let X_1, \dots, X_n be random variables representing the number of citations of the N_p articles (henceforth for simplicity n) of a given scientist. We assume that X_1, \dots, X_n are independent with a common citation distribution function F . Pratelli et al. 2012 and Beirlant and Einmahl 2010, among other contributions, assume that F is continuous, at least asymptotically, even if citation counts have support on the integer set. From our point of view, the definition should be as much as possible coherent with the nature of the data, thus in the present paper we assume that F is discrete and, in order to define the h index, we move to order statistics.

Given a set of n papers of a scientist to which a citations count vector \underline{X}^c is associated, we consider the ordered sample of citations $\{X_{(i)}^c\}$, that is $X_{(1)}^c \geq X_{(2)}^c \geq \dots \geq X_{(n)}^c$, from which obviously $X_{(1)}^c$ ($X_{(n)}^c$) denotes the most (the least) cited paper. Consequently the h index is defined as follows:

$$h = \max\{t : X_{(t)}^c \geq t\} \quad (1)$$

The main latent assumption behind all the above mentioned definitions is the statistical adequacy of the h index as a summary measure. Hirsch claimed that the h index is better than other measures such as the total number of citations because the latter “may be inflated by a small number of -big hits-” (Hirsch, 2005). However, from a proper statistical viewpoint, the h index is not a sufficient statistics. We proved that the h index is not a sufficient statistics. In order to prove this we need to derive the exact distribution of the h index itself. Order statistics can be profitably employed for this purpose, following the procedure outlined in Cerchiello and Giudici (2012), as the following shows.

The exact distribution of the h index is:

$$p(h_i) = [F(C_i) - F(C_i - 1)]^{(n+1-h_i)} \quad (2)$$

Using the above distribution the h-index can be shown not to be sufficient, see for more details Cerchiello and Giudici 2013. A sufficient statistics for the citation vector is obviously the total number of citations and its bijective functionals. The total number of citations has not been considered a valid summary by Hirsch because of his high sensitivity to outlying observations. Although this may be a questionable

remark, it can be naturally taken into account in an appropriate statistical framework as in the following section.

2 Proposal

In this section we present the proposed methodology using the well known academic context.

Let $X_i = (X_{i1}, X_{i2}, \dots, X_{in_i})$ be a random vector containing the citations of the n_i papers published by the i -th scientist. Note that, not only X_i but also n_i is a random quantity that can be denoted with the term 'frequency'. Consequently, the total impact of a scientist i can be defined as the sum of a random number n_i of random citations: $C_i = X_{i1} + X_{i2} + \dots + X_{in_i}$.

Our aim is to derive the distribution of the sufficient statistics C_i and of functionals of interest from it that can be interpreted as quality measures. In order to reach this objective an additional assumption has to be introduced. We assume that, for each scientist $i = 1, \dots, I$ in a homogeneous community, conditionally on the number of papers n_i , the paper impacts X_{ij} , for $j = 1, \dots, n_i$ are independent and identically distributed random variables.

For each scientist i , the distribution function of C_i , that is $F_i(x) = P(C_i \leq x)$, can be found by means of a convolution between the distributions n_i and m_i as follows: $F_i(x) = \sum_{n_i=1} p(n_i) K^{n_i*}(x_{in_i})$, where K^{n_i*} indicates the n_i -fold convolution operator of the distribution $K^*(x_{in_i})$ with itself. We remark that a more trivial assumption could be to model directly the total number of citations C_i , for example as a Poisson distribution: this simplistic assumption evidently discards the fact that the total number of citations is function of individual paper citations each of which may have a different distribution. We incorporate the latter in our convolution model. In order to specify our model we need to fit two distributions, one for the production data and one for the citation patterns. For example, a common assumption may be to take: $p(n_i) \sim \text{Poisson}(\lambda_i)$, $K(m_i) \sim \text{Poisson}(\theta_i)$, where λ_i and θ_i are unknown and strictly positive parameters to be estimated, representing, respectively, the mean number of published papers and the mean number of citations of each scientist. Once parameters are estimated the distribution functions of C_i and H_i can be calculated. From the distribution of C_i one can calculate appropriate statistical summaries that can be used for appropriate inferential purposes on science achievements. For example, the top 5% percentile of the distribution represents a high quality threshold of impact. Finally, for each scientist the point estimate corresponding to the observed C_i can be supplemented with a confidence interval.

However the above summaries and, more generally, functional of interest from $F_i(x)$ may not be obtained analytically. In this rather frequent case one can resort to Monte Carlo simulations to approximate numerically $F_i(x)$. A particular functional of interest is H_i , the h index of a scientist. Our approach can provide a natural inferential framework for the estimation of the h index which is not, differently from Pratelli et al. 2012, based on large sample assumptions.

In practice, for the distribution of the number of papers, we have observed that, in communities characterized by a high level of heterogeneity in the production process, a discrete uniform distribution may be more appropriate. Conversely, as far as citations are concerned, what observed by Hirsch (an author having very few papers with a large number of citations) can be embedded into a discrete extreme value distribution, such as the Zipf distribution or into a continuous one that is the Pareto distribution. In fact, the same distributions have been employed by the bibliometric community, in the context of power law distribution to produce community based correction factors for the h-index.

Specifically we assume that: $f(X_{i(j)}) = \left(\frac{A}{r+\beta_j}\right)^\alpha$ for $r = 1, \dots$, where for a given scientist i , r indicates the rank positions of his/her $j = 1, \dots, n_i$ publications, α is a parameter that describes the decay rate of the distribution, β is a smoothness parameter and finally A is a normalizing constant.

3 Results

We have analyzed daily e-commerce data on 200 players, grouped in 7 market places of activity: Clothing, Grocery, Electronics, Media, Furniture, Food, Other. For each market place we have considered the number of players as the analogue of the number of papers of a scientists and the daily number of orders to each company belonging to the same market place as the citations count for that company. For lack of space we report only the estimated *h-indexes* that are: $h=30$ ($n=230$) for Clothing, $h=15$ ($n=15$) for Grocery, $h=35$ ($n=178$) for Electronics, $h=33$ ($n=40$) for Media, $h=4$ ($n=24$) Furniture, $h=16$ ($n=74$) for Food, $h=20$ ($n=242$) for Other. This shows for example that Electronics is less attractive than Media because although Electronics is much more crowded ($n=130$) presents similar H index (35 vs 33).

References

1. Ball, P.: Index aims for fair ranking of scientists, *Nature* 436 (7053): 900 (2005).
2. Beirlant, J., Einmahl, J. H. J.: Asymptotics for the Hirsch index, *Scand. J. Stat.* 37, 355-364, (2010)
3. Cerchiello, P., Giudici, P.: On the distribution of functionals of discrete ordinal variables, *Statistics and Probability Letters*, 82, 2044-2049, (2012).
4. Cerchiello, P., Giudici, P.: H index: a statistical proposal, *Technical Report*, (2013).
5. Glanzel, W.: On the h-index a mathematical approach to a new measure of publication activity and citation impact, *Scientometrics* 67, 315-321 (2006)
6. Hirsch, J. E.: An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572 (2005)
7. Pratelli, L., Baccini, A., Barabesi, L., Marcheselli, M.: Statistical Analysis of the Hirsch Index. *Scandinavian Journal of Statistics*, 39: 681694. (2012)

Bayesian nonparametric estimation of global disclosure risk

Annalisa Cerquetti

Abstract In disseminating microdata a classical measure of global disclosure risk is the number of sample uniques that are also population uniques. A Bayesian estimation of global risk of disclosure, based on frequencies of frequencies under the superpopulation approach, has been first introduced in Samuels (1998). Here we propose a Bayesian nonparametric solution under two-parameter Poisson-Dirichlet priors on the relative abundances of cross-classifications in the total population. We rely on recent results for posterior prediction estimation of rare species richness under Gibbs priors, a large class of models generalizing the partition structure of the Dirichlet process prior.

Key words: Bayesian nonparametrics, Disclosure risk, Gibbs priors, Sample uniques, Sampling formula, Two-parameter Poisson-Dirichlet priors.

1 Introduction

Statistical agencies that provide microdata for public use need to keep the risk of disclosure of confidential information negligible, ensuring that the file to be released is safe and the risk of reidentification of sampled individuals is low. In samples from social surveys, records consist mainly of categorical attributes. Since attackers might reidentify an individual by matching records with respect to a set attributes called key variables, individuals with sets of attributes that are unique or rare both in the sample and in the population may be identifiable. Consequently a classical measure of global disclosure risk is the proportion of sample uniques which are also population uniques, i.e. the combinations of characteristics represented just one time in the sample, which do not have further representative in the whole population.

Annalisa Cerquetti
MEMOTEF, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome
e-mail: annalisa.cerquetti@uniroma1.it

The idea of working with frequencies of frequencies in disclosure risk estimation was first introduced in Samuels (1998) under the superpopulation approach, then revisited by Fienberg and Makov (2001) and Hoshino (2001). A Bayesian hierarchical model for a record-level measure of re-identification is due to Polettini and Stander (2004). Here we propose a Bayesian nonparametric approach applying some recent results for posterior predictive estimation of rare species richness as introduced in Favaro et al. (2013) and further investigated in Cerquetti (2013). Let (n_1, \dots, n_k) be the multiplicities of the first k combinations of categorical attributes observed in the n -sample to be released, and (c_1, \dots, c_n) the corresponding vector of the frequencies of multiplicities of size $1, 2, \dots, n$. Samuels (1998) proposes a Bayesian nonparametric solution under the assumption that the relative abundances $(P_i)_{i \geq 1}$ of the different cross-categories present in the population in decreasing order are distributed according to a Poisson-Dirichlet (θ) prior (Kingman, 1975), which corresponds to assume that for each n the random vector (C_1, \dots, C_n) has distribution, for $\theta > 0$

$$\mathbb{P}_\theta(C_1 = c_1, \dots, C_n = c_n) = \frac{n! \theta^k}{(\theta)_n} \prod_{i=1}^n \frac{1}{(i)^{c_i} c_i!}. \quad (1)$$

For N the total size of the population, and n the size of the sample to be released, Samuels (1998) shows that the posterior expected proportion of sample uniques c_1 which are population uniques, (DR), corresponds to

$$\mathbb{E}_\theta(DR | n_1, \dots, n_k) = \frac{n + \theta + 1}{N + \theta - 1}.$$

We extend this result to the two-parameter (α, θ) Poisson-Dirichlet class (Pitman and Yor, 1997), and devise a further generalization to the entire Gibbs priors class (Gnedin and Pitman, 2006) which will be the subject of a future paper (Cerquetti and Polettini, 2013).

2 Main results

While the standard approach in disclosure risk analysis is the superpopulation approach in which one assumes that the total population, whose size is known, is a sample from an unknown superpopulation, here we adopt a full Bayesian posterior predictive perspective. We assume that both the total number of possible cross classifications with non zero multiplicities in the total population and the size of the total population are extremely large and unknown. Given the observed sample of size n to be released, we obtain a Bayesian nonparametric estimator of the proportion of cross classifications of size one in the sample, that remain of size one after an additional sample of size m .

Let (X_1, \dots, X_n) be a n sample from the population, and let X_1^*, \dots, X_j^* be the first j labels identifying different combinations of categories observed. The relevant information in the sample is contained in the vector (n_1, \dots, n_j) of the multiplicities

of the first j combinations observed, or in (c_1, \dots, c_n) , for $c_l = \sum_{i=1}^j 1\{n_i = l\}$ the counting vector of the combinations of different sizes. Here c_1 is the number of sample uniques. Given the general form of the exchangeable partition probability function corresponding to the two-parameter (α, θ) model of Pitman and Yor (1997) which is well-known to correspond to

$$p(n_1, \dots, n_j) = \frac{(\theta + \alpha)_{j-1} \alpha}{(\theta + 1)_{n-1}} \prod_{i=1}^j (1 - \alpha)_{n_i - 1}, \quad (2)$$

for $\alpha \in (0, 1)$ and $\theta > -\alpha$, we can state the following:

Proposition 1. *Under two-parameter (α, θ) Poisson Dirichlet priors on the relative abundances of the different cross-classifications arising in the population, a Bayesian posterior predictive estimator of the global disclosure risk (DR), given (n_1, \dots, n_j) the multiplicities of the j combinations observed in the sample to be released, is given by*

$$\mathbb{E}_{\alpha, \theta}(DR | n_1, \dots, n_j) = \frac{(\theta + n + \alpha - 1)_m}{(\theta + n)_m}. \quad (3)$$

For $c_1 \geq 2$, the posterior variance of DR is given by,

$$\begin{aligned} \text{Var}_{\alpha, \theta}(DR | n_1, \dots, n_j) &= \frac{c_1 - 1}{c_1} \frac{(\theta + n + 2\alpha - 2)_m}{(\theta + n)_m} + \\ &+ \frac{1}{c_1} \frac{(\theta + n - 1 + \alpha)_m}{(\theta + n)_m} \left[1 - c_1 \frac{(\theta + n - 1 + \alpha)_m}{(\theta + n)_m} \right]. \end{aligned} \quad (4)$$

Proof. By a result in Favaro et al. (2013, p. 32) the specific form of the falling factorial moments for $O_{l,m}^{(n)} = \sum_{i=1}^j 1\{(n_i + M_i) = l\}$, the number of cross-classifications of size l after an additional sample of size m , with random allocation in the sample categories M_1, \dots, M_j for $\sum_k M_k \leq m$, under (α, θ) priors, for $\{\xi_i : n_{\xi_i} \leq l\}$ corresponds to

$$\begin{aligned} \mathbb{E}_{\alpha, \theta}[(O_{l,m}^{(n)})_{[r]}] &= r! \sum_{(\xi_1, \dots, \xi_r)} \frac{m! \prod_{i=1}^r (n_{\xi_i} - \alpha)_{l - n_{\xi_i}}}{\prod_{i=1}^r (l - n_{\xi_i})! (m - rl + \sum_{i=1}^r n_{\xi_i})!} \times \\ &\times \frac{(\theta + n + r\alpha - \sum_{i=1}^r n_{\xi_i})_{m - lr + \sum n_{\xi_i}}}{(\theta + n)_m}. \end{aligned} \quad (5)$$

For $l = 1$, and $c_{1,n} \geq r$ it reduces to

$$\mathbb{E}_{\alpha, \theta}[(O_{1,m}^{(n)})_{[r]}] = r! \binom{c_{1,n}}{r} \frac{(\theta + n + r\alpha - r)_m}{(\theta + n)_m},$$

and since $DR = (O_{1,m}^{(n)})/c_{1,n}$ it yields (3). Exploiting the known relationship $E[(X)^2] = E[(X)_{[2]}] + E(X)$ the result in (4) easily follows.

Remark 1. To generalize the result in Proposition 1 under Gibbs priors (Gnedin and Pitman, 2006) it is enough to resort to the general form of equation (5) as obtained in Favaro et al. (2013, Th. 1) and further studied in Cerquetti (2013), namely

$$\begin{aligned} \mathbb{E}_{\alpha, V} \left[(O_{l,m}^{(n)})_{[r]} \right] &= r! \sum_{(\xi_1, \dots, \xi_r)} \frac{m! \prod_{i=1}^r (n_{\xi_i} - \alpha)_{l-n_{\xi_i}}}{\prod_{i=1}^r (l - n_{\xi_i})! (m - rl + \sum_{i=1}^r n_{\xi_i})!} \times \\ &\times \sum_{k=0}^{m-rl+\sum_{i=1}^r n_{\xi_i}} \frac{V_{n+m, j+k}}{V_{n, j}} S_{m-rl+\sum_{i=1}^r n_{\xi_i}, k}^{-1, -\alpha, -(n-(j-r)\alpha - \sum_{i=1}^r n_{\xi_i})} \end{aligned} \quad (6)$$

for $\xi_i : n_{\xi_i} \leq l$. Here $S_{n,k}^{-1, -\alpha, \gamma}$ are generalized non central Stirling numbers, and $V_{n,k}$ stands for the coefficient identifying the specific Gibbs model as from the representation of the corresponding Gibbs exchangeable partition probability function $p_{\alpha, V}(n_1, \dots, n_k) = V_{n,k} \prod_{j=1}^k (1 - \alpha)_{n_j-1}$ for $\alpha \in (-\infty, 1)$ and $(x)_s = (x)(x+1) \cdots (x+s-1)$, which generalizes (2).

Acknowledgements The author wishes to thank Silvia Poletini for suggesting disclosure risk estimation as a possible field of application of the Bayesian nonparametric approach to species sampling problems under Gibbs partition models and for pointing out the relevant references.

References

1. Cerquetti, A.: Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling. *Elect. J. Statist.* **7**, 697–716 (2013)
2. Cerquetti, A. and Poletini, S.: A Bayesian nonparametric solution to global disclosure risk estimation under Gibbs priors. *Manuscript in preparation* (2013)
3. Favaro, S., Lijoi, A. and Prünster, I.: Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* (to appear) (2013)
4. Fienberg, S.E. and Makov, U.E.: Uniqueness and disclosure risk: Urn models and simulation. *Research in Official Statistics*, **4**, 23–40 (2001).
5. Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and the Stirling triangles. *Journal of Math. Science* **138**, 3, 5674–5685 (2006)
6. Hoshino, N.: Applying Pitman’s Sampling Formula to Microdata Disclosure Risk Assessment. *J. of Official Statistics* **17**, 499–520 (2001)
7. Kingman, J. F. C.: Random discrete distributions. *J. Roy. Statist. Soc. B.*, **37**, 1–22 (1975).
8. Pitman, J. and Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900 (1997)
9. Poletini, S. and Stander, J. : A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation, in: *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer, V. Torra, LNCS vol. 3050, Springer-Verlag, Berlin, pp. 247–261 (2004)
10. Samuels, S. M.: A Bayesian, Species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. of Official Statistics* **14**, 373–383 (1998)

The Forecasting side of Sovereign Risk: a Generalized Cross Entropy Approach

Enrico Ciavolino, Roberto Savona

Abstract Focusing on the Italian case over the period 01/2007 to 09/2012 we inspect how sovereign risk, proxied by sovereign credit default swap monthly quotes, is connected to previsions and actual data for 5 macroeconomic indicators, namely *inflation*, *consumer confidence*, *industrial orders*, *retail sales*, and *industrial production*. To handle our question we estimate a Generalized Cross Entropy (GCE) regression model using previsions as priors and release data as true realizations. Empirical results prove that priors add informational gain when jointly considered with true realizations and they matter most than simple economic surprises (differences between actual and prevision data).

Key words: Generalized Cross Entropy (GCE), sovereign risk, macroeconomic announcements, credit default swap.

1 Introduction

How sovereign risk dynamics move with information on macroeconomic fundamentals? This is the inspiring question of this paper which may seem, at first sight, common to other many works on the connection between sovereign debt crisis and country fundamentals. However, we take a different route in our search path, by looking into the role played by forecasts on macroeconomic data, made by a pool of experts, *together* with their actual values, announced by institutional authorities on pre-scheduled dates. Suppose you know which are the expectations on some key economic indicators about Italy, which are however mere previsions since the re-

Enrico Ciavolino
University of Salento, Palazzo Parlangeli, Via V.M. Stampacchia, 45, 73100, Lecce, Italy, e-mail: enrico.ciavolino@unisalento.it

Roberto Savona
University of Brescia, C.da S. Chiara, 50, 25122 Brescia, Italy, e-mail: savona@eco.unibs.it

lease data will be available, say, the next week. How previsions reflect on sovereign risk price when true data will be available? This is exactly what we inspect in this paper. We focus on the Italian case over the period 01/2007 to 09/2012 using the sovereign credit default swap monthly quotes, to proxy the market price of sovereign risk, and 5 macroeconomic indicators, namely *inflation*, *consumer confidence*, *industrial orders*, *retail sales*, and *industrial production*. To handle our question we define a regression model based on the Generalized Cross Entropy (GCE) estimator using previsions as priors and release data as true realizations. The way we inspect how previsions and actual macroeconomic data translate into market risk prices is the main innovation of this paper. Indeed, existing literature [1, 2] is focused on how macroeconomic “surprises” reflect on prices, i.e. the difference between actual and forecast data. Instead, the functional form we propose is conceived to say something about the role of previsions (the priors) and true data: do the priors add gain in information content of macroeconomic indicators? This is what we inspect. Main results of empirical analysis show that previsions matter most than actual values, and that priors together with release data explain more than surprises.

2 Generalized Cross Entropy

Generalized Maximum Entropy (GME) and Generalized Cross Entropy (GCE) are approaches proposed to held with some problems could be found in the statistical analysis [3]. The most attractive properties are the possibility to deal with ill-posed data, with matrices short and fat, it is possible to impose inequality constrains on the relationships between the variables and it is possible to introduce prior information on the model.

In this paper we give an introduction to the GCE for regression models, since this method can be seen as a general case of the GME. Let us consider the following regression model for the i^{th} unit:

$$y_i = \alpha + \sum_j^m x_{ij} \beta_j + \varepsilon_i \quad (1)$$

the idea of GCE is the re-parametrization of the coefficients (α and β_j) and the error term (ε_i) as a convex combination of expected value of a discrete random variable, particularly:

$$\beta_j = \sum_k^K z_{jk}^\beta p_{jk}^\beta$$

where z_{jk}^β is the generic element of the *support* \mathbf{z}_j^β , symmetric vector around zero whose dimension is $K \times 1$ (with $3 \leq K \leq 7$), while p_{jk}^β is the generic element of the probability vector \mathbf{p}_j^β associated to \mathbf{z}_j^β . The error term ε_i can also be written as:

$$\varepsilon_i = \sum_h^H z_{ih}^\varepsilon p_{ih}^\varepsilon$$

where z_{is}^ε is the generic element of the support \mathbf{z}_i^ε , a symmetric vector around zero, while p_{ih}^ε is the generic element of the probability vector \mathbf{p}_i^ε associated to \mathbf{z}_i^ε (note that also in this case $3 \leq H \leq 7$). Finally, the GCE regression model for the i^{th} unit can be written as follows:

$$y_i = \left(\sum_k^K z_k^\alpha p_k^\alpha \right) + \sum_j^m x_{ij} \left(\sum_k^K z_{jk}^\beta p_{jk}^\beta \right) + \left(\sum_h^H z_{ih}^\varepsilon p_{ih}^\varepsilon \right) \quad (2)$$

The vectors \mathbf{z}_j^β and \mathbf{z}_i^ε have an important role in the probability estimation procedure together with the choice of their supports. These vectors could be shaped starting from particular prior information or rather could be chosen ad-hoc (for example, by considering \mathbf{z}_i^ε the choice of its supports can be made by the adoption of the *three-sigma-rule* [4]). The parameters of the regression model are estimated by recovering the probabilities distribution of the corresponding parameters and error terms, $\mathbf{p}_j^\beta \quad \forall j = 1, \dots, m$ and $\mathbf{p}_i^\varepsilon \quad \forall i = 1, \dots, n$. The idea of GCE is the minimization of the following entropy function:

$$I(P) = \sum_j \sum_k p_{jk}^\beta \log p_{jk}^\beta / q_{jk}^\beta + \sum_i \sum_h p_{ih}^\varepsilon \log p_{ih}^\varepsilon / q_{ih}^\varepsilon \quad (3)$$

Where q_{jk}^β and q_{ih}^ε represent the prior probabilities distribution for the coefficients and the error term.

3 Estimation Results

The data used in the empirical analysis on market expectations from Bloomberg Financial Services. This data vendor reports on a monthly basis the release data and the corresponding market expectations computed as median response of the respective polls collected approximately one week before the macroeconomic announcements. To proxy the sovereign risk we used the Sovereign Credit Default Swap (SCDS) 5 yrs. The SCDS is a financial swap agreement for which the seller will compensate the buyer in the event of a default or other credit event pertaining to government bonds (the so-called “reference entity”). The price is expressed as a percentage of the notional amount (the “spread”), indicating how expensive is an “insurance” to cover potential losses from defaults. Thus, an increase in the SCDS is because the market perceives as more risky that sovereign, and viceversa.

Computationally, we applied the GCE regression model using data for Italy, which represents an interesting case study especially in lights of the 2011 events which led the spreads of government bonds and SCDS to very high levels and near to default zone. The time period covers the months from January 2007 to September 2012 over which we compared the SCDS quotations for the 5 yrs contract (USD de-

nominated) with 5 covariates: (1) Final CPI monthly variation; (2) ISAE Consumer Confidence Indicator (Trend, 1980=100); (3) Industrial Orders monthly variations; (4) Retail Sales monthly variations; (5) Industrial Production monthly variations.

The method we propose consider two steps of estimation: in the *first* step, we estimate the beta coefficients by using the *prevision* data given from experts, without imposing any prior to the model, that means the q_{jk}^β and q_{ih}^ε probabilities are defined as uniform distribution; in the *second* step, the probability distribution of the beta coefficients and the error term are used as prior for the definition of the GCE regression model. The results in the following table show the estimations for the *prevision* data, for the *released* data, for the model with priors and by the the last column reports the model using the standardized difference between realized data and corresponding prevision, what is known as economic surprise. Parameter estimations reveal that increases in inflation ($\hat{\beta}_1$), and lowerings of confidence ($\hat{\beta}_2$) industrial orders ($\hat{\beta}_3$) and production ($\hat{\beta}_5$) reflect on increases in sovereign risk, which is consistent with economic theory. We also computed the Root Mean Squared Errors (RMSE) obtaining the following results: (1) prevision: 0.802; (2) actual: 0.912 (3) prior: 0.918; (4) diff: 0.975. Interestingly, forecasts made one 1 week before than corresponding true realizations are more informative about the sovereign risk dynamics than data announced by authorities. Priors really add informational gain as proved by our proposed functional form, which shows better results (RMSE) than pure surprises. These findings will change the way with which economists deal with macroeconomic announcements relative to market price dynamics.

Table 1 Parameters Estimations with GCE

Betas	Prevision	Real	Prior	Diff
$\hat{\beta}_1$	0.125	0.019	-0.280	-0.069
$\hat{\beta}_2$	-0.233	-0.371	-0.510	0.058
$\hat{\beta}_3$	0.143	0.032	-0.276	-0.065
$\hat{\beta}_4$	-0.374	-0.012	0.364	0.118
$\hat{\beta}_5$	-0.131	0.001	0.398	0.008

References

1. Ederington, L., and J. Lee: How markets process information: news releases and volatility. *Journal of Finance* **48(4)**, 1161–1191 (1993).
2. Ehrmann, M., M. Fratzscher, R. Gurkaynak, and E. Swanson: Convergence and anchoring of yield curves in the Euro area. *The Review of Economics and Statistics*, **93(1)**, 350–364 (2011).
3. Golan A.: *Information and entropy econometrics-a review and synthesis*. Now Pub (2008)
4. Pukelsheim, F.: The three sigma rule. *The American Statistician* **48(2)**, 88–91 (1994)

What data tell you that models can't say

Nicoletta Cibella, Tiziana Tuoto, Luca Valentino

Abstract Latent class model is a well-known solution for record linkage problem when probabilistic approach is followed. However, the identification of the two unknown distributions (the matches and the non-matches) is not always straightforward. In real data application, even when matching variables are chosen at the best of their availability, latent class models may converge to estimate something different from the expected matching distributions. This paper shows what happens in terms of record linkage quality when the external values are plugged-in parameters of the latent class models aiming at identifying the matching populations. An application to the Post enumeration survey of the 2010 Census of Agriculture is presented.

1 Introduction

Record linkage procedure aims at matching records referring to the same entities, both within a dataset and from two or more different data sources. In the field of the official statistics, the probabilistic record linkage is becoming extremely important [1,2] in order to correctly identify matches and also to evaluate automatically the accuracy of the integration procedure. Following the seminal Fellegi and Sunter [3] theory, for all candidate pairs the probabilistic record linkage model assumes that the unknown linkage status is a latent dichotomous variable. The estimation is based on the results of the comparisons between the selected matching variables. In other words, the probabilistic model adopted for the estimation [4] assumes that the frequency distribution of the observed comparison patterns is a mixture of the matches and non-matches distributions and can be obtained by means of the EM algorithm as proposed in Jaro [5]. This paper shows how the effectiveness of the record linkage process is

¹

Nicoletta Cibella, Istat, Italian National Statistical Institute; email: cibella@istat.it
Tiziana Tuoto, Istat, Italian National Statistical Institute; email: tuoto@istat.it
Luca Valentino, Istat, Italian National Statistical Institute; email: luvalent@istat.it

affected by alternatives values assigned to the parameters of the probability model; in other words, the aim is on the effect of the choices of the parameters on the effectiveness in the identification of the true linkage.

2 Record linkage estimation model

When linking two datasets, say A and B, of size N_A and N_B respectively, the pairs (a,b) in the cross product $A \times B$, of size $N=N_A \times N_B$, need to be classified in two subsets M and U, independent and mutually exclusive, such that M is the set of matches ($a=b$) and U is the set of non-matches ($a \neq b$).

Following Fellegi and Sunter [3] pairs are classified on the basis of comparisons on K common identifiers (matching variables). For each pair (a,b) , a comparison function is applied to each matching variable in order to obtain a comparison vector $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_K\}$. The ratio r , between the probabilities of γ given the pair (a,b) belongs either to the subset M or U,

$$r = \frac{P(\gamma|(a,b) \in M)}{P(\gamma|(a,b) \in U)} = \frac{m(\gamma)}{u(\gamma)} \quad (1)$$

is used to classify pairs on the basis of two thresholds T_m and T_u ($T_m > T_u$): those pairs for which r is greater than the T_m value can be considered as linked; those pairs for which r is smaller than T_u value can be considered as not-linked, if r falls in the range (T_m, T_u) no-decision is made and the pair is held out for the clerical review. The thresholds are chosen so to minimize the number of false matches and false non-matches. A very crucial point in linkage process is the estimation of the $m(\gamma)$ and the $u(\gamma)$ distributions: a common approach is to consider the unknown linkage status as a latent dichotomous variable, C, related to the manifest variables, the k selected identifiers. If the comparisons among variables produce dichotomous results, a contingency table can be calculated. The $m(\gamma)$ and the $u(\gamma)$ probability distributions become the parameters of a log-linear model, together with p the probability of the M set, and they can be estimated, for instance by means of the EM algorithm. The joint distribution of the observations γ and the latent variable $C=c$ ($c=(0,1)$) is given by:

$$P(C = c, \gamma) = [pm(\gamma)]^c [(1-p)u(\gamma)]^{1-c} \quad (2)$$

In order to make the estimation of the parameters easier, the conditional independency assumption between the elements of the vector γ is generally introduced. A heavy complication in the model parameters estimation is due to the very low number of matches compared to the number of non-matches, when considering the whole set of pairs given by the cross-product of the input files. In the real cases, when huge amount of data are managed, it is usual to apply blocking/sorting/clustering methods to reduce the size of the pairs to compare, i.e. the search space dimension [6]. For an analysis of the effect of blocking/sorting methods see Cibella and Tuoto [7].

3 Forcing data: an example on the coverage survey of the 2010 Census of Agriculture

Model definition and parameter estimation are the most crucial aspects of the linkage process. In this paper we are interested in investigating the robustness of the linkage results when different models are chosen and also when the parameter estimation phase is overcome and probability distributions values, evaluated from external sources, are plugged-in the decision rule. This study exploits data from the coverage survey of the 2010 Italian Census of Agriculture. As well-known, in spite of all efforts in enumerating units in the Census, some units are not caught and it is a common practice to reach them by means of an accurate coverage survey. Then capture-recapture models [8] are used in order to estimate the unknown true total amount of the interest population and the rate of census under-coverage requiring the error-free linkage procedures between units collected in Census and Post Enumeration Survey (PES). So, a very accurate and complicated linkage procedure was carried out, including also a huge clerical review activities. As a matter of fact, we can assume that the true linkage status is known for all records involved in Census and PES. The linkage activities for Census and PES are very complex also for record linkage experts, first of all because perfect results are needed. It is usual to approach the problem in steps and perform several linkage passes to achieve the final outcome. From now on we refer to a specific steps in the whole linkage procedure that involved about 5 000 records remaining in the PES that need to be matched with almost 770 000 records of Census. This step was particularly difficult from the estimation point of view, for several reason related to the strong difference in file sizes, the errors and inaccuracies in the available matching variables. The variables considered for linkage are the name of the farms operator (*cdt_nome*), the first six letters of the farms operator fiscal code (*sub_cf*) and its date of birth (*datanas*), the municipality (*cdt_com*) and the province (*cdt_pro*) of the residence where the farms operator lives, the address of the headquarter (*cea_ind*) with its municipality (*cea_com*) and its province (*cea_pro*). Table 1 shows the characteristics of the models adopted.

Table 1: *Characteristics of the models*

	<i>Reduction method</i>	<i>Matching variables (comparison function adopted with the threshold used for the agreement)</i>
Model1	Simhash on <i>cdt_nome</i>	<i>cdt_nome</i> (jaro 0.9) <i>sub_cf</i> (jaro 0.9) <i>cea_ind</i> (jaro 0.8)
Model2	Sorted neighbourhood on <i>cdt_nome</i> (sliding window=100)	<i>cea_pro</i> (equal) <i>cea_com</i> (equal)
Model3	Sorted neighbourhood on <i>sub_cf</i> (sliding window w=180)	<i>cdt_pro</i> (equal) <i>cdt_com</i> (equal)
Model 4		<i>cea_ind</i> (3 grams 0.3) <i>datanas</i> (equal) *

* *datanas* is included only in Model3

These models differ for several aspects including the methods adopted for the search space reduction. In table 1 only the models that allow to achieve the highest number of true links are reported. Some of the selected variables (*cea_com* and *cea_pro* for

models 1 and 2, cdt_pro and cdt_com for models 3 and 4) have a high distinguish power in order to detect the matches (or the un-matches) and for them the models fits well data, the parameters estimation converge to the real matching status. For the other matching variables the parameters estimation of the models is not effective, the estimated probabilities for the m distribution is below 0.5 (the thresholds for reliable estimates) in case of an agreement. In those case some external and suitable probabilities are considered in order to detect the matches.

4 Preliminary results

Figures 1 and 2 show some preliminary results in terms of the effectiveness of the compared models, with $T_m=0.95$ and $T_u=0.5$. The precision indicator is the ratio of the true matches on the total number of matches identified by the procedure while the recall denotes the ratio between the true matches identified on the total number of the true matches. In models 1 and 2 the use of external probabilities (in the figures called *Marginals*) for the variables giving inconsistent estimates produce a relevant effect on the quality of the linkage, see the level of the precision. Instead in the model 3 and 4 the effects are less relevant due to the small number of variables with high identification power. In general, in presence of not reliable estimates the total number of matches raise without producing an increase in the recall indicator.

Figure 1: Precision of models using threshold 0.95 or 0.5

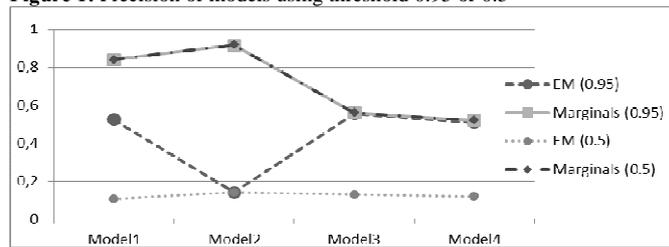
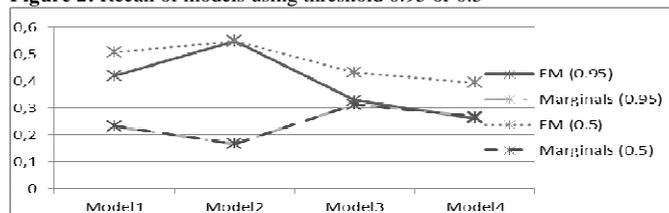


Figure 2: Recall of models using threshold 0.95 or 0.5



References

1. Essnet Project Data Integration, DI, Eurostat, (2009-2011), <http://www.crosportal.eu/content/data-integration-1>.

2. Essnet Project Integration of Survey and Administrative Data, ISAD, Eurostat, (2006-2008), <http://www.cros-portal.eu/content/isad-finished>.
3. Fellegi I.P., Sunter A.B. (1969), "A Theory for record linkage", *Journal of the American Statistical Association*, 64, 1183-1210.
4. Armstrong J., Mayda, J.E. (1993): "Model-based estimation of record linkage error rates", *Survey Methodology*, 137-147.
5. Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", *Journal of the American Statistical Society*, 84 (406), 414-20.
6. Baxter R, Christen P, Churches T (2003) "A Comparison of fast blocking methods for record linkage", <http://www.act.cmis.csiro.au/rohanb/PAPERS/kdd03clean.pdf>.
7. Cibella N., Tuoto T., "Statistical perspectives on blocking methods when linking large data-sets" (2011), in A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, *Studies in Theoretical and Applied Statistics*, DOI 10.1007/978-3-642-21037-2_8, Springer-Verlag Berlin Heidelberg.
8. Wolter K. (1986), "Some coverage error models for Census data", *Journal of the American Statistical Association*, 81, 338-346.

Multiple Hidden Markov Models for Categorical Time Series

Roberto Colombi and Sabrina Giordano

Abstract We introduce multiple hidden Markov models (MHMMs) where an observed multivariate categorical time series depends on an unobservable multivariate Markov chain. MHMMs provide an elegant framework for specifying various independence relationships between multiple discrete time processes. These independencies are interpreted as Markov properties of a mixed and a chain graph associated to the latent and observable components of MHMMs, respectively.

Key words: Granger noncausality, Conditional independence, Graphical models

1 Introduction

In this work, we focus on discrete hidden Markov models with a multivariate categorical observable process depending on a multivariate latent chain, so we observe more variables at each time and assume that their distributions can be affected by one or more latent variables. This is a new extension of the traditional hidden Markov models and we will refer to as multiple hidden Markov models (MHMMs).

With the class of MHMMs, we generalize the hidden Markov models, admitting more than one latent process, that have been proposed in the literature, see [5] for a review. According to our approach, the latent process can satisfy general hypotheses of Granger noncausality and contemporaneous independence, as described by [2] for Markov chains, and different sets of observable categorical time series are allowed to depend on different sets of unobservable processes. Moreover, in the framework of MHMMs, the observable variables are not required to be independent

Roberto Colombi
Department of Engineering, University of Bergamo, Italy, e-mail: colombi@unibg.it

Sabrina Giordano
Department of Economics, Statistics and Finance, University of Calabria, Italy, e-mail: sabrina.giordano@unical.it

given the latent states (local independence assumption), but the association between them is also modeled.

To enhance the advantages of parsimony and interpretability of MHMMs, we will present graphical Markov models for MHMMs where the transition probabilities of the latent process and the distributions of the observable variables conditioned on the latent states are required to obey Markov properties encoded by mixed and chain graphs.

2 Graphical modeling of multiple hidden Markov models

Let $\mathbf{E}_{\mathcal{U}}$ be a r -variate process of categorical variables, $\mathbf{E}_{\mathcal{U}} = \{E_{\mathcal{U}}(t) : t \in \mathbb{N}\} = \{E_i(t) : t \in \mathbb{N}, i \in \mathcal{U}\}$, $\mathcal{U} = \{1, \dots, r\}$, $\mathbb{N} = \{0, 1, 2, \dots\}$ and let $\mathbf{F}_{\mathcal{V}}$ be a s -dimensional process of categorical variables $\mathbf{F}_{\mathcal{V}} = \{F_{\mathcal{V}}(t) : t \in \mathbb{N}\} = \{F_j(t) : t \in \mathbb{N}, j \in \mathcal{V}\}$, $\mathcal{V} = \{1, \dots, s\}$. The following Definition states when $(\mathbf{E}_{\mathcal{U}}, \mathbf{F}_{\mathcal{V}})$ is an MHMM.

Definition 1. The joint process $(\mathbf{E}_{\mathcal{U}}, \mathbf{F}_{\mathcal{V}})$ is an MHMM if: a) $\mathbf{E}_{\mathcal{U}}$ is not observable; b) $(\mathbf{E}_{\mathcal{U}}, \mathbf{F}_{\mathcal{V}})$ is a first order multivariate Markov chain; c) $E_{\mathcal{U}}(t) \perp\!\!\!\perp F_{\mathcal{V}}(t-1) | E_{\mathcal{U}}(t-1)$; d) $F_{\mathcal{V}}(t) \perp\!\!\!\perp E_{\mathcal{U}}(t-1), F_{\mathcal{V}}(t-1) | E_{\mathcal{U}}(t)$.

Note that the markovianity of the process $(\mathbf{E}_{\mathcal{U}}, \mathbf{F}_{\mathcal{V}})$ is an assumption whereas in the usual definition of hidden Markov models it is a consequence and that condition c) implies that $\mathbf{E}_{\mathcal{U}}$ is a first order Markov chain as proved in [1]. Hereafter, for every subset $\mathcal{T} \subset \mathcal{U}$ and $\mathcal{R} \subset \mathcal{V}$, marginal processes of the latent chain and the observed variables are represented by $\mathbf{E}_{\mathcal{T}} = \{E_i(t) : i \in \mathcal{T}, t \in \mathbb{N}\}$ and $\mathbf{F}_{\mathcal{R}} = \{F_j(t) : j \in \mathcal{R}, t \in \mathbb{N}\}$.

We now address the use of graphical Markov models for MHMMs. Such models associate missing edges of a graph with some conditional independence restrictions imposed on the probabilities of the observable variables given the latent states and the transition probabilities of the latent process.

We consider two types of graphs: a mixed graph for the latent component of the MHMM and a chain graph for the observable component given the latent states. Specifically, the transition probabilities of the latent chain $\mathbf{E}_{\mathcal{U}}$ in the MHMM are required to obey a set of Markov properties encoded by a mixed graph G whose basic features are discussed in [2] for Markov chains, while the Markov properties satisfied by the distribution of the observed variables conditioned to the latent states are read off a chain graph G^* . See [3] for a review of the Markov properties of chain graphs.

In the mixed graph G , a node i corresponds to the marginal process \mathbf{E}_i , for every $i \in \mathcal{U}$, and Granger noncausality and contemporaneous independence restrictions are associated with missing directed and bi-directed edges, respectively. In particular, missing bi-directed edges lead to independencies of marginal processes at the same point in time; missing directed edges, instead, refer to independencies which involve marginal processes at two consecutive instants. The chain graph G^* with two chain components τ_0 and τ_1 serves the need to encode the independence rela-

tions among observable and latent variables of the MHMM at a given time point. The nodes of the chain graph G^* , belonging to τ_0 correspond to the variables $E_i(t^*)$, $i \in \mathcal{U}$, and the nodes in τ_1 correspond to $F_j(t^*)$, $j \in \mathcal{V}$, for any arbitrary $t^* \in \mathbb{N}$. All the edges in the subgraph induced by a chain component are bi-directed and the graph induced by the chain component τ_0 is bi-complete. Furthermore, the directed edges in graph G^* point in the same direction from τ_0 towards τ_1 .

Definition 2. An MHMM $(\mathbf{E}_{\mathcal{U}}, \mathbf{F}_{\mathcal{V}})$ is Markov with respect to (wrt) a mixed and a chain graph when the latent component is Markov wrt a mixed graph, and the observation component given the latent states is Markov wrt a chain graph. In particular, the latent process $\mathbf{E}_{\mathcal{U}}$ is Markov wrt a mixed graph G if and only if its transition probabilities satisfy the following conditional independencies for all $t \in \mathbb{N} \setminus \{0\}$ associated to missing directed and bi-directed edges of G , respectively:

$$E_{\mathcal{T}}(t) \perp\!\!\!\perp E_{\mathcal{U} \setminus pa_G(\mathcal{T})}(t-1) | E_{pa_G(\mathcal{T})}(t-1) \quad \forall \mathcal{T} \in \mathcal{U} \quad (1)$$

$$E_{\mathcal{T}}(t) \perp\!\!\!\perp E_{\mathcal{U} \setminus sp_G(\mathcal{T})}(t) | E_{\mathcal{U}}(t-1) \quad \forall \mathcal{T} \in \mathcal{U} \quad (2)$$

The observable process $\mathbf{F}_{\mathcal{V}}$ is Markov wrt a chain graph G^* if and only if the distribution of the observable variables given the latent states satisfies the following conditional independencies for all $t \in \mathbb{N} \setminus \{0\}$ associated to missing bi-directed and directed edges of G^* , respectively:

$$F_{\mathcal{R}}(t) \perp\!\!\!\perp F_{\mathcal{V} \setminus sp_{G^*}(\mathcal{R})}(t) | E_{\mathcal{U}}(t) \quad \forall \mathcal{R} \in \mathcal{V} \quad (3)$$

$$F_{\mathcal{R}}(t) \perp\!\!\!\perp E_{\mathcal{U} \setminus pa_{G^*}(\mathcal{R})}(t) | E_{pa_{G^*}(\mathcal{R})}(t) \quad \forall \mathcal{R} \in \mathcal{V}. \quad (4)$$

For a set \mathcal{S} of nodes of a graph \mathcal{G} , the sets of parents $pa_{\mathcal{G}}(\mathcal{S})$ and spouses $sp_{\mathcal{G}}(\mathcal{S})$ are defined, for example, in [2]. In the context of first order multivariate Markov chains, condition (1) corresponds to the classical notion of Granger noncausality and ensures that the marginal process $\mathbf{E}_{\mathcal{T}}$ is a Markov chain (see [2], [4]). Henceforth, we will refer to (1) with the term *Granger noncausality* condition saying that the latent process $\mathbf{E}_{\mathcal{T}}$ is not G-caused by $\mathbf{E}_{\mathcal{U} \setminus pa(\mathcal{T})}$ with respect to $\mathbf{E}_{\mathcal{U}}$. Condition (2) states that the transition probabilities satisfy the bi-directed Markov property [6] with respect to the graph obtained by removing the directed edges from the mixed graph. Here, we will refer to (2) with the term *contemporaneous independence* con-

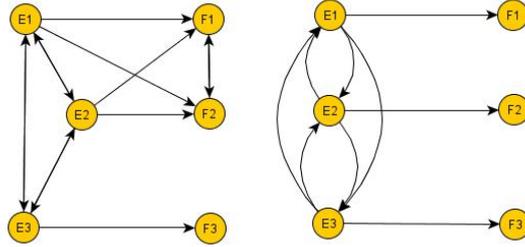


Fig. 1 Mixed-chain graphs associated to MHMMs

dition and say that the latent processes $\mathbf{E}_{\mathcal{T}}$ and $\mathbf{E}_{\mathcal{Z} \setminus sp(\mathcal{T})}$ are contemporaneously independent. On the other hand, conditions (3) and (4) encoded by the chain graph G^* refer to the observable component of the MHMM given the latent states and are equivalent to the type IV Markov properties *C2b* and *C3b* discussed by [3]. Note that, the independencies among observable variables given the latent ones, at time t , are bi-directed Markov properties and the independencies of a set of observable variables from a set of latent variables, at time t , are conditioned with respect to the remaining latent variables, but not to the remaining observable variables.

As a matter of convenience, the mentioned mixed graph and the two component chain graph can be superimposed as shown in Figure 1 to form a unique graph that we will refer to as mixed-chain graph. When reading the independencies (1, 2) off the sub-graph induced by the first component of a mixed-chain graph associated to an MHMM, it is relevant to remind that the nodes are intended to represent marginal processes. The independencies (3, 4) are instead encoded in the chain graph where nodes represent random variables.

Example 1. The mixed-chain graph on the right of Figure 1 encodes that the latent variables of the three dimensional MHMM are Granger caused reciprocally but are contemporaneously independent; moreover, every observed variable depends only on its own latent variable and, at every time point, the observable variables given the latent states are independent. The mixed-chain graph on the left of Figure 1 is associated to a 3-variate MHMM where the latent variables are not contemporaneously independent, and each of them depends only on its own past; at every time, given the first two latent variables, the first two observable variables are independent of the third latent variable; at every time, the third observable variable does not depend on the first two latent variables, conditionally upon the third latent variable and, finally, at every time, the third observable variable is independent of the other two given all the latent variables.

Finally, we mention that the above independencies are equivalent to zero constraints on the parameters of marginal models for the transition and observation probabilities, property which greatly simplifies the fit of MHMMs.

References

1. Colombi, R., Giordano, S.: Testing lumpability for marginal discrete hidden Markov models. *Adv. Stat. Anal.*, **95**, 293–311 (2011)
2. Colombi, R., Giordano, S.: Graphical models for multivariate Markov chains. *J. Multivariate Anal.*, **107**, 90–103 (2012)
3. Drton, M.: Discrete chain graph models, *Bernoulli* **15(3)**, 753–763 (2009)
4. Florens, J.P., Mouchart, M., Rolin, J.M.: Noncausality and marginalization of Markov processes. *Econometric Theory* **9**, 241–262 (1993)
5. Koski, T.: *Hidden Markov Models for Bioinformatics*. London: Kluwer Academic Publishers (2001)
6. Richardson, T.: Markov properties for acyclic directed mixed graphs, *Scand. J. Stat.* **30**, 145–157 (2003)

Asymptotics in survey sampling for high entropy sampling designs

Pier Luigi Conti and Daniela Marella

Abstract The aim of the paper is to establish asymptotics in sampling finite populations. Asymptotic results are first established for an analogous of the empirical process based on the Hájek estimator of the population distribution function, and then extended to Hadamard-differentiable functions. As an application, asymptotic normality of estimated quantiles is provided.

Key words: Hájek estimator, sampling entropy, quantiles.

1 Introductory aspects

Asymptotic results in sampling finite populations are widely used in different contexts. All results are concerned with the asymptotic normality of (usually linear) statistics, under different conditions and sampling plans. Among several papers devoted to this subject, we mention [4], where asymptotic properties of L -statistics are obtained under stratified two-stage sampling, [3], where asymptotic properties for the Horvitz-Thompson estimator under rejective sampling are obtained, as well as [7], [1], where Hájek's results are extended to different sampling designs. In [2] the estimation of the population distribution function and quantiles is studied in case of a stratified cluster sampling. Clusters are selected from each stratum without replacement and with equal inclusion probabilities. The results obtained are similar to those in [4].

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy
e-mail: pierluigi.conti@uniroma1.it

Daniela Marella
Università Roma Tre, Via del Castro Pretorio 20, 00185 Roma, Italy
e-mail: daniela.marella@uniroma3.it

Here we attempt to construct an asymptotic theory for sampling finite populations that parallels, as far as possible, the classical theory in nonparametric statistics. For the important class of “high entropy” sampling designs, similarities and differences between finite populations results and classical nonparametrics will be discussed.

In the sequel, the symbols used are introduced. Let \mathcal{U}_N be a finite population of N units, labeled by integers $1, \dots, N$. Let Y be the variable of interest, and for each unit i , denote by y_i the value of Y ($i = 1, \dots, N$). Let further $y_N = (y_1, \dots, y_N)$. For each real y , the *population distribution function* (p.d.f., for short) is defined as

$$F_N(y) = \frac{1}{N} \sum_{i=1}^N I_{(y_i \leq y)}, \quad y \in \mathbb{R} \quad (1)$$

where $I_{(y_i \leq y)}$ is 1 if $y_i \leq y$, and 0 otherwise.

For each $0 < p < 1$, the p th *population quantile* $Q_N(p)$, say, is the left-continuous inverse of F_N computed at point p . In symbols:

$$Q_N(p) = F_N^{-1}(p) = \inf\{y : F_N(y) \geq p\}. \quad (2)$$

Now, for each unit i in \mathcal{U}_N , define a Bernoulli random variable (r.v.) D_i , such that the unit i is included in the sample if and only if (iff) $D_i = 1$, and let D_N be the N -dimensional vector of components D_1, \dots, D_N . A (unordered, without replacement) sampling design P is the probability distribution of D_N . In particular, $\pi_i = E_P[D_i]$ is the inclusion probability of unit i . The suffix P denotes the sampling design used to select population units.

The class of sampling designs we consider asymptotically behave as the rejective sampling. Let p_1, \dots, p_N be N real numbers, with $p_1 + \dots + p_N = n$. The sampling design is a *Poisson design* with parameters p_1, \dots, p_N if the r.v.s D_i s are independent with $Pr_{Po}(D_i = 1) = p_i$ for each unit i , the suffix Po denoting the Poisson design.

The *rejective sampling*, or *normalized conditional Poisson sampling* ([3], [5]) corresponds to the probability distribution of the random vector D_N , under Poisson design, conditionally on $n_s = n$.

The *Hellinger distance* between a sampling design P and the rejective design is defined as

$$d_H(P, P_R) = \sum_{D_1, \dots, D_N} \left(\sqrt{Pr_P(D_N)} - \sqrt{Pr_R(D_N)} \right)^2. \quad (3)$$

Our basic assumptions are listed below.

- A1. $(\mathcal{U}_N; N \geq 1)$ is a sequence of finite populations of increasing size N .
- A2. For each N , y_1, \dots, y_N are realizations of a superpopulation (Y_1, \dots, Y_N) composed by *i.i.d.* r.v.s Y_i with common d.f. F . In the sequel, we will denote by \mathbb{P} the probability distribution of r.v.s Y_i s, and by \mathbb{E}, \mathbb{V} the corresponding operators of mean and variance, respectively.
- A3. For each population \mathcal{U}_N , sample units are selected according to a fixed size sample design with inclusion probabilities π_1, \dots, π_N , and sample size n . Fur-

thermore, for $d > 0$, $f > 0$,

$$d_N = \sum_{i=1}^N \pi_i(1 - \pi_i) \rightarrow \infty, \quad \frac{1}{N}d_N \rightarrow d, \quad \lim_{N \rightarrow \infty} \frac{n}{N} = f \text{ as } N \rightarrow \infty.$$

A4. For each population $(\mathcal{U}_N; N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities π_1, \dots, π_N , and let P be the actual sampling design (having the same inclusion probabilities). Then

$$d_H(P, P_R) \rightarrow 0 \text{ as } N \rightarrow \infty.$$

A5. There exist two positive real numbers A, B such that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{\pi_i} = A < \infty, \quad \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{1}{(i\pi_i)^2} = B < \infty.$$

2 Main Basic results

In order to estimate the p.d.f. F_N , consider the Hájek estimator

$$\widehat{F}_H(y) = \frac{\sum_{i=1}^N \frac{1}{\pi_i} D_i I_{(y_i \leq y)}}{\sum_{i=1}^N \frac{1}{\pi_i} D_i}. \quad (4)$$

Our main goal is to study the asymptotic “global” behaviour of the random function $\widehat{F}_H(\cdot)$. In order to accomplish this, we define the (sequence of) random function(s)

$$W_N^H(y) = \sqrt{n}(\widehat{F}_H(y) - F_N(y)), \quad y \in \mathbb{R}; \quad N \geq 1. \quad (5)$$

The main results of the present section is Proposition 1.

Proposition 1. *If the sampling design P satisfies assumptions A1-A5, with \mathbb{P} -probability 1, conditionally on y_N the sequence $(W_N^H(\cdot); N \geq 1)$, converges weakly, in $D[-\infty, +\infty]$ equipped with the Skorokhod topology, to a Gaussian process $W^H(\cdot) = (W^H(y); y \in \mathbb{R})$ that can be represented as*

$$W^H(y) = \sqrt{f(A-1)}B(F(y)), \quad y \in \mathbb{R} \quad (6)$$

where $(B(t); 0 \leq t \leq 1)$ is a Brownian bridge.

In classical nonparametric statistics, the empirical process $\sqrt{n}(\widehat{F}_n(y) - F(y))$ converges weakly to a Gaussian process of the form $B(F(y))$. This result is apparently similar to Proposition 1, with two differences: (i) the centering factor F instead of F_N ; (ii) the absence of the finite population correction term $f(A-1)$,

since in classical nonparametric statistics the observations are (realizations of) *i.i.d.* r.v.s, and essentially there is non sampling design.

The asymptotic result of Proposition 1 are obtained conditionally on the population values y_N ; hence, the involved probability is *only* the sample design probability. Proposition 1 refers to design-based inference, where the values y_i s are considered as *fixed*. In other words, the role played by the superpopulation model of assumption A2 is of secondary importance.

Consider now a map $\phi(\cdot) : D[-\infty, +\infty] \rightarrow E$, E being an appropriate normed space. Assume further that $\phi(\cdot)$ is Hadamard-differentiable at F (cfr. [6]). Then, the following result holds.

Proposition 2. *If $\phi(\cdot)$ is (continuously) Hadamard-differentiable at F , with Hadamard derivative $\phi'_F(\cdot)$, then the asymptotic law of $\sqrt{n}(\phi(\tilde{F}_H(y)) - \phi(F_N))$ coincides with the asymptotic law of $\phi'_F(W^H)$, as N increases.*

As an application of Proposition 2, we may prove the asymptotic normality of estimated quantiles.

Proposition 3. *Suppose that F is continuously differentiable with derivative $f = dF/dy$, and let $\tilde{Q}_H(p) = \tilde{F}_H^{-1}(p) = \inf\{y : \tilde{F}_H(y) \geq p\}$, for $0 < p < 1$. Then, the sequence of random processes $\tilde{T}_N^H(\cdot) = (\sqrt{n}(\tilde{Q}_H(p) - \tilde{Q}_N(p)))$; $\varepsilon \leq p \leq 1 - \varepsilon$ converges weakly, in $D[\varepsilon, 1 - \varepsilon]$ equipped with the Skorokhod topology, to a Gaussian process $T^H(\cdot) = (T^H(p))$; $\varepsilon \leq p \leq 1 - \varepsilon$ that can be represented as*

$$T^H(p) = \sqrt{f(A-1)} \frac{B(p)}{f(Q(p))}, \quad \varepsilon \leq p \leq 1 - \varepsilon \quad (7)$$

where $B(p)$ is a Brownian bridge.

References

1. Berger, Y G.: Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 209–226 (1998)
2. Francisco, C A., Fuller, W A.: Quantile Estimation with a Complex Survey Design. *The Annals of Statistics*, **19**, 454–469 (1991)
3. Hájek, J.: Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, **35**, 1491–1523 (1964)
4. Shao, J.: L-Statistics in Complex Survey Problems. *The Annals of Statistics*, **22**, 946–967 (1994)
5. Tillé, Y.: *Sampling Algorithms*. Springer, New York (2006)
6. Van Der Vaart, A.: *Asymptotic Statistics*. Cambridge University Press, Cambridge (1998)
7. Vížek, J A.: Asymptotic Distribution of Simple Estimate for Rejective, Sampford and Successive Sampling. In: *Contributions to Statistics* (Ed. J. Jurečková), 263–275, Reidel Publishing Company, Dordrecht, Holland (1979)

On the Use of Recursive Partitioning in Casual Inference: A Proposal

Claudio Conversano, Massimo Cannas and Francesco Mola

Abstract A tree-based method for identification of a balanced group of observations in casual inference studies is presented. The method derives from an algorithm which uses a multidimensional balance measure criterion to recursively split the dataset based on the values of the covariates. Observations are finally partitioned in subsets characterized by different degrees of homogeneity. An ad-hoc resampling scheme is used to select the units for which causal inference can be carried out.

Key words: Regression trees, Resampling, Average Treatment Effect, Balancing Recursive Partitioning.

1 Introduction

In experimental studies about the estimation of the effect of a treatment on a set of individuals the randomization of treatment assignment implies that the treated and control groups are balanced with respect to observed and unobserved covariates. Thus, it is possible to obtain unbiased estimator of causal effects via direct comparison of treated and control units. In observational studies the treated and control groups may differ in the distribution of covariates; in these situations the difference in the distribution of the outcome variable between the two groups cannot be attributed solely to the treatment and a direct comparison may give a biased estimate. However, under suitable conditions [2, 4] unbiased estimates of treatment effects can be obtained after balancing the distribution of covariates across treated and control units. Matching methods are useful to identify similar observations and for this reason they have been suggested for achieving balance. Matching requires that the covariates distributions of treated and control units share a common support of val-

Claudio Conversano, Massimo Cannas and Francesco Mola
Dipartimento di Scienze Economiche ed Aziendali, Università di Cagliari, Viale S. Ignazio 17,
09123 Cagliari, e-mail: conversa@unica.it, massimo.cannas@unica.it, mola@unica.it

ues otherwise (i.e., in case of lack of overlap) the comparison is not possible. Usually, the identification of the common support is done either indirectly, by imposing a minimum distance for two units to be matched together, or directly, by restricting the analysis to those units that are considered to belong to the common support. In the latter case the common support must be a-priori identified.

An hybrid approach to the definition of the common support called Random Recursive Partitioning (RRP) has been introduced in [3]. RRP defines, at each iteration, a random partition of the covariates space by growing regression trees with fictitious outcome $Z \sim U(0, 1)$. Each random partition is used to derive a proximity matrix whose elements are used to weight the difference in mean across treated and control units and/or to select a subset of observations belonging to the common support. Analyzing the data used in [1], RRP shows that both weighted estimators and estimators based on a selected subset of treated and control units provide reasonable results in comparison with traditional methods. In addition, it seems that the subsets selected through RRP have better covariate balance than the unselected ones.

In this paper we propose an approach, based on recursive partitioning, which exploits the balancing property of tree-based methods. It uses a resampling scheme to derive a sequence of trees working on resampled versions of the original data and whose main aim is to balance the set of covariates by considering as splitting criterion a multidimensional balancing measure. The method uses a certain number of samples and, with the tree obtained from each of these, assigns more sampling weight to the units belonging to the most homogeneous subsets (or terminal nodes). As in RRP, the final proximity matrix can be used to weight causal estimates or to select a subset of observations. Since the proposed approach is mainly aimed at balancing covariates by using recursive partitioning, we name it *Balancing Recursive Partitioning Algorithm* (BaRPA).

The main features of BaRPA are described in Section 2, whereas Section 3 presents the results of the performance of BaRPA on simulated data.

2 Balancing Recursive Partitioning

Given an outcome variable Y , a set of covariates X_j ($j = 1, \dots, p$) and a treatment variable T observed on N cases, BaRPA can be applied in all the situations in which the presence of the treatment variable T influences the distribution of Y and X_j in a way that the distribution $Y|T = 0$ can differ from $Y|T = 1$, and the same can happen for each $X_j|T = 0$ with respect to the corresponding $X_j|T = 1$. In this framework, the basic idea supporting the implementation of BaRPA is that the balance of each $X|T$ on the basis of the set of covariates X_j and the treatment variable T can be obtained in a recursive manner. Specifically, it is possible to exploit the capability of one of the $X_j|T$ in improving the balance of the whole set of covariates. BaRPA tries to obtain this balancing in a recursive way: data are partitioned by selecting a splitting covariate and its associated split point that minimize the global imbalance of the set of covariates in at least one of the two resulting child nodes. As in [3], BaRPA

allows us to estimate a proximity matrix which measures how close is a treated unit with all the untreated ones: this matrix is obtained by growing a tree on resampled subsets of the original data.

BaRPA is also tailored to the identification of a subset of matched observations for which balance in covariates distribution holds. To this aim, BaRPA is orientated towards the search of a (possibly small) subset of observations whose covariate distribution is, on average, more balanced than the original distribution observed on the whole dataset. The balanced subset is detected in the first iteration (i.e., after growing the first tree). Resampling subsets of observations helps to assess if this detection was obtained by chance or if it really identifies a balanced subset.

2.1 Main steps of BaRPA

1. Tree growing. A binary recursive partitioning of the data is performed in order to identify subregions of the covariates space in which the distributions of each $X_j|T = 0$ and its corresponding $X_j|T = 1$ are more balanced. Node splitting is based on the idea that a node is split if either the right or the left child node are more balanced than the parent node. BaRPA uses the Average Standardized Absolute Mean difference (ASAM) to split a node: It searches the splitting covariate X_L^* and its associated split point x_L^* that minimize the average ASAM of all the covariates in the left child node, as well as the splitting covariate X_R^* and its associated split point x_R^* that minimize the the average ASAM of all the covariates in the right child node. The split point (x_L^* or x_R^*) providing the maximum decrease in imbalance compared to the same measure concerning the parent node is used for splitting.

The tree growing process proceeds until all current terminal nodes cannot be split since they do not provide any improvement in the imbalance of $Y|T$. In addition the user can specify, as stopping criterion, a minimum number of treated and control units (n_{min}) that are required to split a node.

2. Subset selection and weights updating. Once that a tree has been grown, BaRPA looks at the terminal nodes presenting the best balancing of each $X_j|T = 0$ with respect to $X_j|T = 1$, the so called *best-balanced nodes*. These nodes are also characterized by a ratio between the number of treated units belonging to a given node and that of the corresponding control units which is between 0.5 and 2. Denoting with N_T and N_C the number of treated and control units in the original data and with n_t and n_c the treated and control units belonging to the best-balanced nodes, the procedure selects the subset $n_t + n_c$ and uses this subset to estimate the proximity matrix.

3. Estimation of the proximity matrix Π and of the average treatment effect. The proximity matrix Π is made up of N_T rows and N_C columns and is estimated after growing R binary trees on resampled versions of the original data. Once that, in iteration r ($r = 1, \dots, R$), a binary tree has been grown and a subset has been selected, a proximity measure $\pi_{ij}^{(r)}$ ($i = 1, \dots, N_T; j = 1, \dots, N_C$) is derived as follows:

$\pi_{ij}^{(r)}$ is set to 1 for treated and control units belonging to the selected subset $n_t + n_c$, whereas it is set to zero for the remaining $N - (n_t + n_c)$ units. The final proximity matrix is the average of the $\pi_{ij}^{(r)}$ s over the R samples:

$$\Pi = \left[\pi_{ij} = \frac{1}{R} \sum_{r=1}^R \pi_{ij}^{(r)} \right] \quad (1)$$

To estimate the average treatment effect, BaRPA uses the same estimators proposed in [3]: A weighted ATT estimator, based on weights $f_{ij} = \pi_{ij} / \sum_{j=1}^{N_C} \pi_{ij}$, is:

$$ATT_W = \frac{1}{N_T} \sum_{i=1}^{N_T} \left[(Y|T=1)_i - \sum_{j=1}^{N_C} f_{ij} (Y|T=0)_j \right] \quad (2)$$

Moreover, an ATT estimator based on normalized weights is considered. Here, $\pi_i^{max} = \max_{j \in (1, \dots, N_C)} \pi_{ij}$ indicates the maximum number of times a treated unit i has been matched to a control unit. These weights can be normalized by defining $q_i = \pi_i^{max} / \sum_{i=1}^{N_T} \pi_i^{max}$ in such a way that $\sum_{i=1}^{N_T} q_i = 1$. The normalized ATT estimator is

$$ATT_N = \frac{1}{N_T} \sum_{i=1}^{N_T} \left[(Y|T=1)_i - \sum_{j=1}^{N_C} f_{ij} (Y|T=0)_j \right] q_i \quad (3)$$

BaRPA also implements the *selected* ATT estimators, which are built solely on the treated units that have been matched at least λ^* % times with some other control units. The value of $\lambda^* \in (0, 1)$ corresponds to the maximum value of λ for which either more than n_{min} treated units or more than n_{min} control units can be selected. Alternatively, the selection threshold can be specified as $n_{min} + k \cdot \sigma_\tau$ (τ is the number of units selected by the R trees and σ_τ is its standard deviation; k is a constant) in order to select larger subsets. Thus, selected estimators evaluate the average treatment effect restricted to the portion of treated units that can be reliably matched. These estimators are denoted by $S.ATT_W$ and $S.ATT_N$. Of course, if the goal is estimating the average treatment effect on the control (ATC), then ATC_W , ATC_N , $S.ATC_W$ and $S.ATC_N$ can be defined in the same way.

4. Subsets Sampling and Stopping rule. As previously stated, after growing the first tree BaRPA selects a balanced subset $n_t + n_c$ of cases. In order to define weights for the estimation of the proximity matrix Π as well as to overcome the well-known instability problem characterizing recursive partitioning algorithms, BaRPA investigates, by growing additional trees on resampled data, if the selection of the first balanced subset is sensitive to small perturbation in the original data and if the same subset can be further refined. The sampling scheme depends on which average treatment effect is being estimated (ATT or ATC) and is motivated by the idea of finding the best (set of) counterpart(s) for each (treated or control) unit. In this respect, when estimating ATT resampling works by retaining, in each run, N_T units and by drawing $N_C = N_T$ cases from the original data with weights \mathbf{w}_{N_C} . Whereas, when

estimating ATC , N_C cases are retained in each run and $N_T = N_C$ are drawn from the original data with weights w_{N_T} . Weights w_{N_C} and w_{N_T} are updated, in each run, on the basis of the selected subset $n_t + n_c$. BaRPA stops the process of identification of the sequence of trees as soon as the relative change in value of one of the previously-defined estimators is lower than a previously specified threshold.

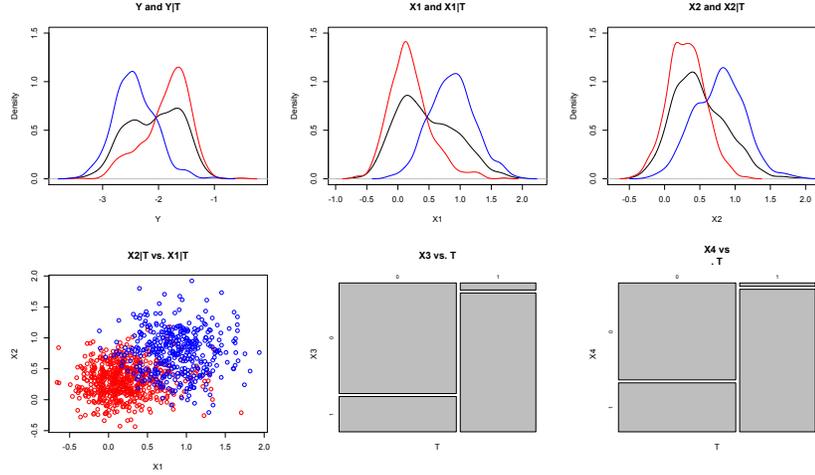


Fig. 1 *Top panel.* Imbalance in the distribution of the outcome Y (left) and of the continuous variables X_1 (center) and X_2 (right): red lines identifies treated units, blue lines identifies control ones. *Bottom panel.* Identification of the common support of X_1 and X_2 (left): red points identifies treated units, blue points identifies control ones; Imbalance in the distribution of the categorical variables X_3 (center) and X_4 (right).

3 Simulation Study

The performance of BaRPA has been evaluated on simulated data. The design factors consider a dataset composed of $N = 1000$ cases on which a treatment effect is generated in order to obtain $P(T = 1) \approx 0.4$ through the logistic function:

$$P(T = 1) = (1 + \exp(-(-.22 + \sum_{i=1}^4 B_i \cdot X_i + B_5 \varepsilon)))^{-1} \quad (4)$$

In (4), X_1 , X_2 and ε are mixtures of two normal random variables while X_3 and X_4 are two dummy variables deriving from a mixture of two uniform random variables. The B coefficients equals 0.10 for B_1 , B_2 , B_3 and B_5 , and zero for B_4 .

The outcome Y is generated as a linear combination of the 4 covariates and of the treatment T , and is such that the *true treatment effect* G is 0.60:

$$Y = -1.5 + 0.60X_1 - 0.72X_2 - 0.73X_3 - 0.20X_4 + 0.60T + \varepsilon \quad (5)$$

Simulated data are highly imbalanced: this imbalance can be observed from the empirical distribution of Y , X_1 and X_2 , as well as from their conditional distributions, as represented in Figure 1.

Table 1 shows the results about the performance of BaRPA on the simulated data.

Table 1 Results provided by BaRPA on simulated data (relative bias and average ASAM).

ATT	\hat{G}	$ r_b(\hat{G}) $	R	λ^* rule	λ^*	$\hat{n}_t(\hat{n}_c)$	μ_{ASAM}	$\Delta(\mu_{ASAM})$	$\mu_{ASAM999}$
ATT_W	-0.746	2.24	311						
ATT_N	0.459	0.23	199						
$S.ATT_W$	0.612	0.02	19	$n_{min} = \sqrt{N_T}$	0.48	20(20)	0.08	-0.96	1.96
$S.ATT_W$	0.501	0.16	79	$\sqrt{N_T} + \sigma_\tau$	0.27	54(54)	0.41	-0.79	1.97
$S.ATT_W$	0.427	0.29	48	$\sqrt{N_T} + 2\sigma_\tau$	0.11	86(86)	0.34	-0.83	1.97
$S.ATT_W$	0.418	0.30	103	$\sqrt{N_T} + 3\sigma_\tau$	0.03	137(122)	0.66	-0.67	1.97
$S.ATT_N$	0.463	0.23	816	$n_{min} = \sqrt{N_T}$	0.53	20(21)	0.46	-0.77	1.96
$S.ATT_N$	0.479	0.20	202	$\sqrt{N_T} + \sigma_\tau$	0.28	54(54)	0.23	-0.88	1.96
$S.ATT_N$	0.503	0.16	300	$\sqrt{N_T} + 2\sigma_\tau$	0.10	93(86)	0.38	-0.81	1.96
$S.ATT_N$	0.482	0.20	182	$\sqrt{N_T} + 3\sigma_\tau$	0.04	123(119)	0.62	-0.69	1.96

^a The table reports, in each row, the results of the estimators defined in Section 2.1. \hat{G} is the estimated value for the true treatment effect G and $|r_b(\hat{G})|$ is its relative bias; R denotes the number of samples used by BaRPA; λ^* rule is the criterion used to define the empirical threshold λ^* for $S.ATT_W$ and $S.ATT_N$; $\hat{n}_t(\hat{n}_c)$ is the number of selected treated (control) units; μ_{ASAM} is the average ASAM obtained for the 4 covariates on the selected subset $\hat{n}_t(\hat{n}_c)$ and $\Delta(\mu_{ASAM})$ denotes the relative change of μ_{ASAM} with respect to the same measure computed on the original data; $\mu_{ASAM999}$ is the average μ_{ASAM} obtained on 999 independent samples composed of \hat{n}_t treated and \hat{n}_c control units.

All the estimators, except ATT_W , present a reasonable, and in some cases extremely low, relative bias. As for balancing, all the selected estimators consistently reduce the average ASAM of the original data (column $\Delta(\mu_{ASAM})$): in particular, $S.ATT_W$ with the n_{min} rule for the selection of λ^* provides the maximum reduction in imbalance by detecting a small subset composed of 20 treated and 20 control units.

References

- Dehejia, R., Wahba, S.: Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**(448), (1999)
- Heckman, J.H., Ichimura, H., Smith, J., Todd, P.: Characterizing selection bias using experimental data. *Econometrica*, **66**(5), 1017–1098 (1998)
- Porro, G., Iacus, S.M.: Random Recursive Partitioning: A matching method for the estimation of Average Treatment Effects. *J. of Appl. Econom.* **24**, 363–385 (2009)
- Rubin, D.B.: The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics* **29**, 184–203 (1973)

Keeping the pace with higher education. A fuzzy states gender study

Franca Crippa, Marcella Mazzoleni e Mariangela Zenga

Abstract Gender indicators on education show, in so called developed countries, a situation where women have reached high standards. At any rate, these indicators are apt to global descriptions and classifications, since they are conceived to compare highly different contexts. Therefore the specific situation of developed countries require further and deeper insight in a complex system [Mecatti, and Farina, 2012]. In particular, the educational situation in Italy, as in other western countries, is increasingly in favour of females, under many respects: grades, ability to keep in pace with the requirements of the institution, both in terms of faster attainment of titles and of lower dropout rates. With reference to the Italian university system, namely concerning a cohort of students, we apply the theory of non homogeneous Markov systems with fuzzy states, in order to describe the differential students' educational progress in terms of gender. The levels of students' career progression will be related to the academic years using a stochastic model that assumes the progress levels as fuzzy states [Symeonaki, Kalamatianou, 2011] with a membership function related both to the exam's final grade and to the time needed to pass the exam in comparison with the schedule provided in the syllabus. The membership function will be derived separately for women and men, in order to relate each fuzzy state to the corresponding administrative situation and to derive further insight in differential students' progression.

Key words: higher education, gender, fuzzy state

Franca Crippa
Department of Psychology, Universit di Milano-Bicocca, piazza dell'Ateneo Nuovo, 1, Milan e-mail: franca.crippa@unimib.it

Marcella Mazzoleni an Mariangela Zenga,
Department of Statistics and Quantitative Methods, Universit di Milano-Bicocca, via Bicocca degli Arcimboldi,8 e-mail: m.mazzoleni8@campus.unimib.it, mariangela.zenga@unimib.it

1 Introduction

Gender indicators [1] on education show, in so called developed countries, higher standards in favour of the female student population, particularly remarkable when compared with the following working outcome, to an extent that has given birth to the saying 'Learn more, earn less' referred to women. In particular, European dropout rates prove to be remarkably lower for women and, conversely, higher for men. An analogous trend seem to concern the ability to keep in pace with the formative path states in the formative offer, where female students show more regular trajectories. In order to understand in depth this educational gap, with reference to the Italian university system, we consider a cohort of students, applying the theory of non homogeneous Markov systems with fuzzy states. The aim is the description of the students' educational progress respectively for females and males. The levels of students' career progression will be related to the academic years using a stochastic model that assumes the progress levels as fuzzy states [2] with a membership function related both to the exam's final grade and to the time needed to pass the exam in comparison with the course syllabus. The membership function will be derived separately for women and men, in order to relate each fuzzy state to the corresponding administrative situation and to derive further insight in differential students' progression. It should be highlighted that issue of respecting the higher education (HE) official timing in Italian university asks for a neat specification of the undergraduates' degree of freedom in answering to the curricular requirements. As a matter of fact, in the Italian HE system the administrative registration to the subsequent academic years does not correspond to a unique situation in terms of accumulated formative credits. In fact, unlike several other countries in the European Union, in Italy registering to the subsequent year is admissible to all students at the end of a certain academic year, regardless of the number of exams passed in the current year and in the previous ones. Therefore, late graduations and dropouts can originate from several behavioural patterns in terms of curricular developments. In addition, it should be underlined how HE students' credit balance might be due, on the one side, to the differential difficulty in complying with specific curricular requirements, on the other side to the development of optimal strategies in coping with the requirements themselves [3].

2 Towards new indicators for gender differential behaviour. Markov chains with fuzzy states

Markov chains methods have been widely applied to career paths in several context, from education to the workplace. Their fundamental form, where states are mutually exclusive and can be clearly identified and each state is influenced only by the previous one, though, does not apply to situation analogous to the Italian HE system, where the administrative situation does of being registered to a specific academic

year corresponds to multiple conditions in terms of credits. The clarity of identification of states can fail in some HE systems therefore fails, giving place to fuzzy states, when two sorts of situations essentially occur. In the university system, the actual states can be exactly measured and observed, but their number is so large that decisions are not associated with the exact states of the system. In these two sorts of situations, decisions are associated with non mutually exclusive states which can be defined as fuzzy sets on the original non-fuzzy state space of the system, where the system has a large number of states. The optimal condition of having accomplished all due requirements can be identified in terms of successful strategies [4,5], that can differ in terms of gender. Furthermore, the relation of every combination of credits and exams to the fuzzy state can be quantified and associated to dropouts. The fuzzy space state is, the, $F = \{F_1, F_2, ..F_N\}$ where the number of elements N is lower than the one in the state space S , $k, k < N$. The relation between the observable k space state elements and the fuzzy state space N elements is specified in terms of the function $\mu_{F_r}(j)$ of the element of the fuzzy set $F_r, r=1,2,..,N$:

$$F_r() : S[0; 1] \tag{1}$$

where the expression above is said the membership function $\mu_{F_r}(j)$ and it quantifies the relative compositional quota of each observable state with respect to the fuzzy state under scrutiny. Besides, $F = \{F_1, F_2, ..F_N\}$ is assumed to define a fuzzy partition on S so that:

$$\sum_{r=1} \mu_{F_r}() = 1 \tag{2}$$

Each actual condition in terms of the combination of the amount of gained credits and the exams passes needs to be referred to the administrative year of registration by means of an explicit quantification of intensity, in order to specify analytically the membership function in equation (2). With respect to the to sth student, his record of $l+1$ passed examinations are considered. With respect to each i passed exam, $i = 1, 2, .., l + 1$, the index md_i [6] is computed, in order to evaluate its relative difficulty compared with the whole student's accomplishment, apart from the exam under consideration:

$$md_i = \frac{\sum_{i=1}^n [y_{si} - \frac{\sum_{k=1}^l y_{sk}}{l}]}{n} \tag{3}$$

being y_{si} the grade the student obtained at that examination, l is the total number of exams the student has passed in the same administrative year of registration, excluding the i^th one under scrutiny, n the number of students that successfully passed the exam l^th in the year considered. Therefore the index in equation 4 relies on the individual difference between the specified mark and the average mark in all other passed exams in the same academic year and it estimates the relative difficulty of examinations in different subjects. Then, from this index and adapted for the Italian setting, the membership function $\mu_{\{F_r(.)\}}$ for the fuzzy sets F_r is estimated :

$$\mu_{F_r} = \frac{1 + \frac{\sum md_i * f c_i}{33-18+1}}{m_1} \tag{4}$$

where $\sum md_i$ refers to the exams neither taken nor passed in the year of enrollment and m_1 is the total number of examinations included in the syllabus for that year and it quantifies the relative compositional quota of each observable state with respect to the fuzzy state under scrutiny. Lastly, for each curricular year, the probability for a student to belong to the fuzzy state has been derived by means of the following expression:

$$P[Y^f(t-1) = F_r] = \sum_{j=1}^k P[Y(t-1) = j] \cdot \mu_{F_r}(j). \quad (5)$$

where $Y(t)$ e $Y^f(t)$ represent respectively the state and fuzzy state of a student in the system S at time t. In case the time-event of interest be the dropout, it needs to be highlighted that it is an absorbing state, in the sense that, once experimented, it is impossible to leave for any other state.

3 Preliminary results

In our preliminary data analysis, membership functions and transition probabilities have been estimated in the R environment, on the base of a originally developed *ad hoc* code [7]. Henceforth we refer to a undergraduate course in a different former faculty, Psychological Sciences and Techniques, Faculty of Psychology, with respect to the cohort enrolling in autumn 2005. Only enrollments that start their credit attainment anew have been taken into account, their cohorts accounting to 485 students at the Faculty of Psychology, 360 females and 185 males respectively. Undergraduates' career was followed up to five year after matriculation, so as to observe most of them up until graduation. On the whole, the membership function shows an extremely wide range of credits attainments where, in any case, the female student population proves more adherent to the institutional requirements. As shown in Table 1 females show a higher value of the membership function for the correspondence between the administrative and the actual academic situation, when compared with their male colleagues in Table 2. For instance, the membership function for female students registered at the second year (second row of the tables) is equal to 0,3528, versus 0.3232 for male. This implies that females, at the end of their second administrative year, have accomplished approximately 3% more credits than males related to that very year, and have around 7% less credits still to gain from the previous administrative year. Differences in dropout are more remarkable: females show a rate of 0.17 at the end of first year, versus a male value of 0.24. The following two academic years show a male rate almost double than the female one.

The Italian

In line with applications to other national contexts, our preliminary results confirm higher educational standard for women. In particular, they underline a wider behavioural divergence in more extreme decisions, namely in not continuing university, whereas the gap in adhering to institutional requirements is far more reduced, in terms of credit progressions, for students who stay in HE.

Level of progression	1 f.s.	2 f.s.	3f.s.	graduation	dropout
1st	0.3354	0.4924	0.0000	0.0000	0.1722
2nd	0.2287	0.3528	0.3514	0.0000	0.0671
3rd	0.1437	0.1955	0.1089	0.4979	0.0540
1st out of programme	0.1818	0.2105	0.0953	0.4272	0.0853
2nd out of programme	0.1471	0.1988	0.0974	0.2425	0.3143

Table 1 Probabilities of meeting syllabus requirements in relation to the academic year of registration, undergraduate course in Psychological and Technical Sciences, female population

Level of progression	1 f.a.	2 f.s.	3f.s.	graduation	dropout
1st	0.3912	0.3688	0.0000	0.0000	0.2400
2nd	0.2937	0.3232	0.2568	0.0000	0.1263
3rd	0.2169	0.2211	0.0824	0.3712	0.1084
1st out of programme	0.2909	0.3139	0.1320	0.1925	0.0682
2nd out of programme	0.1106	0.1816	0.0988	0.2454	0.3636

Table 2 Probabilities of meeting syllabus requirements in relation to the academic year of registration, undergraduate course in Psychological and Technical Sciences, male population

References

1. Mecatti, F., Crippa, F. and Farina, P. A Special Gen(d)er of Statistics. Development and Methodological Prospects of Gender Statistics *International Statistical Review*, vol. 80, 3, 452-467 (2012)
2. Symeonaki, M. and Stamoub, G.B. . Theory of Markov systems with fuzzy states, *Fuzzy Sets and Systems*, Volume 143, 3, 427-445, (2004)
3. Betti, G., Chelli B. and Cambini. R.. A statistical model for the dynamics between two fuzzy states: theory and an application to poverty analysis, *Metron - International Journal of Statistics*, vol. LXII, 3, 391-411 (2004)
4. Sah., M. and Degtiarev, Y. *Forecasting enrollment model based on first order fuzzy time series*, Proceedings of world academy of science, engineering and technology, vol.1, 2005.
5. Symeonaki, M. and Kalamatianou, A. *Markov system with fuzzy states for describing students' educational progress in Greek universities*, World Statistics Congress (ISI), Dublin, (2011)
6. Kelly, A. The relative standards of subject examinations, *Research Intelligence* 2, .34-38(1975)
7. Mazzoleni, M. *Le catene di markov con stati fuzzy: un' applicazione al caso delle carriere universitarie*, Final Dissertation, Master Degree In Economics, University of Milano-Bicocca, Milan, Italy, (2011)

CUB model to validate FACIT TS-PS measurement instrument

F. Cugnata, C. Guglielmetti and S. Salini

1 Introduction

The evaluation of healthcare quality is certainly a complex subject, also due to the fact that it is a multidimensional concept. Complexity is even higher when considering chronic illness patients. For them, besides the quality of the medical therapy, good quality of care is important as they have more articulated needs and more frequently interact with their physicians comparing to acute disease patients. Moreover, besides good medical therapy, good quality of care determines patient satisfaction. Satisfaction with medical care was first identified as an integral component of health care quality assurance programs by the World Health organization in 1989 [4]. It has been defined as the assessment of the fulfillment of individual needs and expectations of those receiving care by means of indirect or direct questions about the quality of care provided [2].

The Functional Assessment of Chronic Illness Therapy version Treatment Satisfaction and Patient Satisfaction (FACIT-PS-TS) is part of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System which is a comprehensive and extensive set of self-report instruments for the assessment of health-related quality of life (QOL) in patients with cancer or other chronic illnesses. This specific version has been developed in order to assess the patients perception of the quality of care and the related satisfaction in chronic illness health care services.

The FACIT TS-PS is a 25-item instrument, subdivided into seven core quality of care domains:

Federica Cugnata
DEMM, Università degli Studi di Milano, e-mail: federica.cugnata@unimi.it

Chiara Guglielmetti
DEMM, Università degli Studi di Milano, e-mail: chiara.guglielmetti@unimi.it

Silvia Salini
DEMM, Università degli Studi di Milano, e-mail: silvia.salini@unimi.it

- A. *Explanations* (4 items) received about their illness
- B. *Interpersonal* (3 items) relations with healthcare personnel (physicians and nurses)
- C. *Comprehensive* (3 items) care in term of ability of multiprofessional team as a whole to be responsible for all aspects of the disease including the impact on personal, relationship and work
- D. *Technical quality* (3 items) competence of physicians
- E. *Decision Making* (5 items)
- F. *Nurse* (3 items)
- G. *Trust* (4 items) in physicians

Patients are required to evaluate the 6 different factors on a 4 points scale (0=No, not at all; 1=Yes, but not as much as I wanted; 2=Yes, almost as much as I wanted; 3= Yes, and as much as I wanted). Additionally the FACIT TSPS includes a 3 items overall measure. The first (referred to as the recommendation item) asked patients if they would recommend the hospital to others, the second (referred to as the repeat choice item) asked patients if they would choose the same clinic or office again. Both are on a 3 point scale (with possible response categories of yes, may be and no). The third (referred to as the satisfaction item) asked patients to rate their overall evaluation of care on a 4 point scale (with response categories of poor, fair, good, very good and excellent).

2 Results

The instrument FACIT-PS has not yet validation neither in English nor in Italian. The Italian version validation presents some problems. An attempt to validate it through a confirmatory factor analysis (CFA) was done, see Figure 1. The results, when analyzed by dimension, confirmed the internal consistency of the items (in fact, the Cronbach's Alpha were all close to 0.80), but seems that the whole model is not confirmed ($Chi-Squares = 1072,055$, $df = 295$, $p-value = 0.000$) although almost acceptable ($CFI = 0.904$ and $RMSEA = 0.078$).

We tried to apply CUB approach, in order to select only the significant items and to obtain a more parsimonious version of the instrument.

Mainly motivated by psychological arguments, CUB models has been proposed by Piccolo [3]. In these models, the answers to ordinal response items in a questionnaire are interpreted as the result of a cognitive process, where the judgement is intrinsically continuous but is expressed in a discrete way within a prefixed scale of m categories. The rationale of this approach stems from the interpretation of the final choices of respondents as a result of a complex mechanism whose main components are the *feeling* of the subject towards the item and an intrinsic *uncertainty* in choosing the ordinal value of the response [1].

In order to show the typical result of the CUB model in Figure 2 we present observed relative frequencies and fitted probability distributions obtained by fitting CUB(0,0) models for the two overall variables, *recommendation* and *satisfaction*.

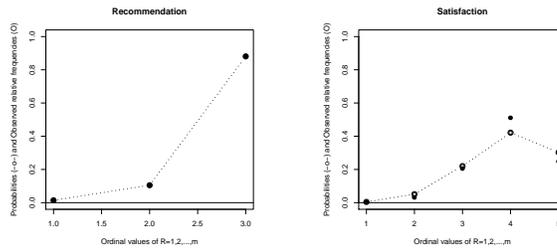


Fig. 1 CUB(0,0)

The models are substantially different, in particular in *feeling* (as measured $1 - \xi$). These aspects are neatly displayed in Figure 3 where estimated models are located in the parametric space of CUB models. Figure 3 show the classical CUB map Feeling vs Uncertainty obtained applying CUB(0,0) model to all items. The thing you notice immediately is that the overall is not a summary of the items, as might be expected.

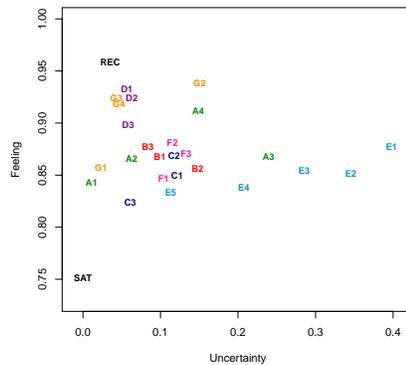


Fig. 2 CUB map Feeling vs Uncertainty

We estimate a series of CUB(0,q) models with a overall measure as a dependent variable and items satisfaction as covariates to explain feeling and we use the step-wise algorithm to select the best CUB(0,q) model with covariates to explain feeling. For each dimension we select the the significant items to explain *feeling* of *recommendation* and *satisfaction*. We compare the significant items of the *satisfaction* and of the *recommendation* and we consider all the significant items of the *satisfaction*, of the *recommendation*, or of both.

We apply a CUB(0,25) containing all the initial items and CUB(0,15) using the significant items with *recommendation* and *satisfaction* as a dependent variable.

The maximized log-likelihood for the CUB (0, 25) for *recommendation* is 109, which is higher than the value for the CUB (0, 15) model that is -115. Moreover, for the first model $AIC = 272$ and $BIC = 381$, for the second model $AIC = 263$ and $BIC = 332$, so according to these criteria the second model is preferable than the first, even if the likelihood ratio test is not significant: $2(\ell_{25} - \ell_{15}) = 12$ and $\chi_{g=10}^2 = 18.31$ at the significance level of $\alpha = 0.05$.

The maximized log-likelihood for the CUB (0, 25) for *satisfaction* is -431, which is higher than the value for the CUB (0, 15) model that is -434. Moreover, for the first model $AIC = 916$ and $BIC = 1026$, for the second model $AIC = 903$ and $BIC = 972$, so according to these criteria the second model is preferable than the first, even if the likelihood ratio test is not significant: $2(\ell_{25} - \ell_{15}) = 6$ and $\chi_{g=10}^2 = 18.31$.

The model with 125 items is, according to the likelihood ratio test, equal to the one with 25 is an important result that encourages even more the need for validation of the instrument.

The confirmatory factor analysis (CFA) was run using only the 15 significant items. The results, reported on Table 1, confirmed that the reduced model could be accepted.

Table 1 CFA results

	Complete Model	Reduced Model
CFI	0.904	0.970
RMSEA	0.078	0.059

A future effort in the direction of validation could be to submit a shortened version of the questionnaire to patients and to analyze the results.

References

- [1] Iannario M, Piccolo D (2012) Cub models: Statistical methods and empirical evidence. In: Kenett R, Salini S (eds) *Modern Analysis of Customer Surveys: with Applications using R, Statistics in Practice*, Wiley
- [2] Jenkinson C, Coulter A, Bruster S, Richards N, Chandola T (2002) Patients experiences and satisfaction with health care: results of a questionnaire study of specific aspects of care. *Quality and Safety in Health Care* 11(4):335–339
- [3] Piccolo D (2003) On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* 5:86–104
- [4] Vrijhoef HJ, Berbee R, Wagner EH, Steuten LM (2009) Quality of integrated chronic care measured by patient survey: identification, selection and application of most appropriate instruments. *Health Expectations* 12(4):417–429

Multilevel Approach in Meta-Analysis of Pre-Election Poll Accuracy

Rosario D'Agata, Venera Tomaselli

Abstract. Following a meta-analysis approach as a *special* case of multilevel modelling, we are interested to identify the potential sources of dissimilarities in accuracy measures of pre-election polls, carried out in Parliamentary elections in Italy from 2001 to 2008. Multilevel model approach decomposes variance components as well as meta-analysis random models but differently from traditional meta-analysis, multilevel approach makes estimate procedure easier and more flexible.

1 Meta-analysis: a case of hierarchical modelling

This paper addresses the topic of meta-analysis as a secondary method of research to connect the results of several empirical studies about the relationship between explanatory variables and dependent variables in a common metric called the effect size (Higgins *et al.*, 2002; 2003).

In order to calculate an average effect size, we can use fixed or random effects models (Borenstein *et al.*, 2009). Estimating fixed effects model, we assume that the true effect size across the studies is always the same. The only error source is the sampling error. Formally:

$$d_j = \delta_j + e_j \quad [1]$$

where d_j is the effect size of study j ($j = 1, \dots, J$), δ_j is the value observed in the population j and e_j is within study error, that is, the sampling error. So the effect size is function of a common effect plus an error term assumed normally distributed with variance σ_e^2 .

Usually, however, we can not assume that the effect size is identical across the study. So we estimate different effect sizes by the means of a random effects model. Then, the true effect sizes δ_j , assumed to vary across the studies, is defined as mean of population effect sizes. The observed effect size d_j in [1], obtained from each primary study, is sampled from a distribution of effects with true effect sizes δ_j , and variance σ_u^2

Rosario D'Agata, University of Catania, Vitt. Emanuele II, 8 Catania, e-mail: rodagata@unict.it

Venera Tomaselli, University of Catania, Vitt. Emanuele II, 8 Catania, e-mail: tomavene@unict.it

In turn, δ_j is function of the mean of all true effects γ_0 plus *between*-studies error u_j . Formally:

$$\delta_j = \gamma_0 + u_j \quad [2]$$

So we can rewrite the [1] as:

$$d_j = \gamma_0 + u_j + e_j \quad [3]$$

where d_j is the effect observed in the j -study, γ_0 is the estimate for the mean outcome across all studies, u_j is *between*-studies residual error term and e_j is *within*-studies error term (Hox and de Leeuw, 2003). As a consequence, the effect size estimates are affected by two error sources: sampling variance and variance of the population effect sizes, due to the dissimilarity across the studies outcomes.

Employing a multilevel approach, the [2] can be written as following:

$$\delta_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_k Z_{kj} + u_j \quad [4]$$

where δ_j is the effect size assumed varying across the studies; γ_0 is the mean of all true effects; Z_{kj} are study features; γ_k are the coefficients and u_j is the error term representing the differences across the studies. On assume that u_j is normally distributed with variance σ_j^2 .

Substituting the [4] in the [1], the model can be written as:

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_k Z_{kj} + u_j + e_j \quad [5]$$

Multilevel model approach as well as traditional meta-analysis random models is oriented to decompose variance of study outcomes into two components: the *within* study variance, as sampling variance and the *between* study variance, due to the differences across the study results, computed as predictive accuracy measures. Estimating the effect size by the means of a multilevel models is simpler than by traditional meta-analysis methods, because a multilevel approach is more flexible (Hox and de Leeuw, 2003). It allows us to avoid to cluster the studies, due to heterogenous effect sizes across them. So, we do not need to identify any variable defining the membership of studies to a cluster. Into the multilevel model, furthermore, we can include the study characteristics as predictors to explain the variance across the j -studies (Hedges and Olkin, 1985).

2 The accuracy of pre-electoral polls in Italy

In order to measure how much is accurate a poll outcome, we choose to use the poll accuracy measure A_i (Martin *et al.*, 2005) as a predictor of an election result:

$$A_i = \ln \left\{ \frac{s_{ij}/(1 - s_{ij})}{[S_j / (1 - S_j)]} \right\} \quad [6]$$

where:

- s_{ij} is the proportion of respondents favoring the s -competitor (party, coalition or candidate) in the i -th poll referring to the j -th population
- $1-s_{ij}$ is the proportion of respondents favoring other else competitors in the same i -th poll, for the same j -th population

- S_j is the real proportion of votes polled by the same S -competitor in the same j -th population
- $1-S_j$ the actual proportion of votes polled by other else competitors in the same j -th population.

As an important assumption is that the distribution of results is normal (Bryk *et al.* 1992), the logarithmic transformation of the odds ratio is used both to create a symmetric measure around 0 and to simplify the computation of the variance (Martin *et al.*, 2005, 351), taking into account the sampling error of the poll measure, assumed normally distributed and with a known variance.

Now, we specify a multilevel model aiming to identify statistically significant relations between forecast accuracy and polls characteristics. We collected 42 polls carried out in the two weeks before ‘embargo’ week on the occasion of the elections for Chamber of Deputies in Italy from 2001 to 2008 and published on website: www.sondaggielettorali.it.

For each poll we gathered: poll outcome, the customer, survey method, sample size, poll length, the day distance between survey and election and predicted gap between the two most important coalitions. The dependent variable is computed as shown in [6], on the basis of poll outcomes and official electoral data. We consider A_i as the accuracy measure of electoral forecast for Centre-Left alliance.

In the first step, an ‘empty’ model is estimated¹ in order to control for the homogeneity of outcomes across the polls (Table 1). The value of intercept (0.078) informs that the electoral outcome of Centre-left alliance is on the average overestimated.

Table 1: Empty model: $y = \text{Accuracy Centre-Left alliance}$.

	β_{0j}	<i>E.S.</i>	<i>Z</i>	<i>p-value</i>
<i>Fixed effects</i>				
Intercept	0.078	0.035	2.229	< 0.02
<i>Random effects</i>				
σ_{uj}^2	0.047	0.011	4.273	< 0.001
<i>Deviance = - 4.839</i>				
<i>Deviance difference test: $\chi^2 = 1089,7$; <i>gdl</i> = 41; <i>p-value</i> = < 0.001</i>				

The model, furthermore, shows the random component (σ_{uj}^2) which is significant. Since the Wald test, implemented in MIWin, is inaccurate to test between polls variance (Hox and de Leeuw, 2003), we use the *deviance difference* test for a null-hypothesis that σ_{uj}^2 is equal to zero. The test produces a χ^2 of 1089,7 (p -value < 0.001) and indicates the presence of heterogeneity in accuracy measures across the polls.

In the second step, we estimate the complete model in order to identify the characteristics of polls correlated with the accuracy measure. All predictors appear significant (Table 2).

Notably, when customer is *political organ* or *agency* (compared with *Mass Media*), the accuracy measure decreases. All of the survey methods introduced in the model, respect to *CATI* method, appear positively correlated with the accuracy. The greater is sample size and the longer is survey period, the more accurate is the poll. The less is the distance between poll and election, the greater is the predicted gap between the two

¹ The models are estimated by employing MIWin software using RML algorithm.

alliances, more accurate is the forecast. Finally, in those elections where *Centre-Left* wins accuracy measure decreases.

Table 2: Complete model: $y = \text{Accuracy Centre-Left alliance}$

	β_{0i}	<i>E.S.</i>	<i>Z</i>	<i>p-value</i>
<i>Fixed effects</i>				
Intercept	0.675	0.224	3.013	< 0.002
Customer: <i>agency</i> (Ref. <i>Mass Media</i>)	-0.179	0.056	-3.196	< 0.001
Customer: <i>political organ</i> (Ref. <i>Mass Media</i>)	-0.280	0.110	-2.545	< 0.006
Survey method: <i>CATI e CAWI</i> (Ref. <i>CATI</i>)	0.200	0.068	2.941	< 0.002
Survey method: <i>CAWI</i> (Ref. <i>CATI</i>)	0.280	0.071	3.944	< 0.001
Survey method: <i>CASI</i> (Ref. <i>CATI</i>)	0.369	0.14	2.636	< 0.005
Sample size: $\ln n_i/N_i$	0.033	0.013	2.538	< 0.006
Poll length (days)	0.055	0.021	2.619	< 0.005
Days between poll and election	-0.019	0.006	-3.167	< 0.001
Predicted gap	0.024	0.006	4.000	< 0.001
Electoral winner: <i>Centre-Left</i> (Ref. <i>Centre-Right</i>)	-0.217	0.055	-3.945	< 0.001
<i>Random effects</i>				
σ_{ui}^2	0.012	0.003	4.000	< 0.001
<i>Deviance</i> = -51.596				
<i>Deviance difference test:</i> $\chi^2 = 238.3$; <i>gdl</i> = 31; <i>p-value</i> = < 0.001				

The complete model reduces the variance across the polls from 0.047, observed in the empty model, to 0.012. It means that the heterogeneity in the accuracy depends on polls characteristics considered to estimate the model.

References

1. Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R., Introduction to Meta-Analysis. John Wiley & Sons, Ltd (2009).
2. Bryk A., Raudenbush S. W., Hierarchical linear models: applications and data analysis methods. Sage publications, Newbury Park (2000).
3. Hedges, L.V. and Olkin, I., Statistical Methods for Meta-Analysis. Academic Press, INC., Orlando, Florida (1985).
4. Higgins, J. and Thompson, S. G., Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*. (21): 1539-1558 (2002).
5. Higgins, J., Thompson, S. G., Deeks, J. J., and Altman, D. G., Measuring inconsistency in meta-analysis. *British Medical Journal*. (327): 557-560 (2003).
6. Hox, J. and de Leeuw, E., Multilevel models for meta-analysis, in: Reise, S. P. and Duan, N. (eds). *Multilevel Modeling: Methodological Advances, Issues, and Applications*. Lawrence Erlbaum Associates. Mahwah (NJ): 90-111 (2003).
7. Martin, E. A., Traugott, M. W. and Kennedy, C., A review and proposal for a new measure of poll accuracy. *Public opinion Quarterly*. (69): 342-369 (2005).

Low-dimensional tracking of association structures in categorical data

Alfonso Iodice D'Enza and Angelos Markos

Abstract Multiple correspondence analysis (MCA) is a well established dimension reduction method to explore the associations within a set of categorical variables and it consists of a singular value decomposition (SVD) of a suitably transformed matrix. The high computational and memory requirements of ordinary SVD makes its application impractical on massive or sequential data sets that characterize several modern applications. The aim of the present contribution is to allow for incremental updates of existing MCA solutions, which lead to an approximate yet highly accurate solution; this makes it possible to track, via MCA, the association structures in data flows. To this end, an incremental SVD approach with desirable properties is embedded in the context of MCA.

Key words: Singular value decomposition, Correspondence analysis, Incremental methods

1 Introduction

Multiple correspondence analysis (MCA) is a suitable dimension reduction method for the visual exploration of the association structure characterizing a set of categorical attributes [3]. Classic applications of MCA range from marketing to psychology, to social and environmental sciences. In the last decade,

Alfonso Iodice D'Enza
Department of Economics and Law, Università di Cassino e del Lazio Meridionale,
Italy e-mail: iodicede@unicas.it

Angelos Markos
Department of Primary Education, Democritus University of Thrace, Greecee-mail:
amarkos@eled.duth.gr

new frameworks of application emerged, that usually involve large/massive amounts of categorical data.

MCA can be accomplished via a Singular Value Decomposition (SVD) of a suitably transformed data matrix. The MCA applicability suffers from the same limitations of SVD, typically: unfeasible computational requirements for large and high-dimensional data; unsuitable for application to data flows, since all the data being analyzed must be available from the start, or eventually kept in memory.

In the literature there are several proposals aiming to overcome the SVD-related limitations via efficient eigensolvers (for example, [1]) or via incremental updates of existing SVD solutions according to new data (see [2] for an overview). The aim of the present contribution is to extend the use of MCA as a visual tracking tool for evolving association structures. To this end, we propose a modification of MCA to deal with incremental updates of existing solutions, that we refer to as ‘Live’ MCA and leads to an approximate, albeit accurate, solution.

The paper is organized as follows: In Section 2 we briefly recall MCA as a dimension reduction method for categorical data. An incremental modification of MCA for tracking association structures is proposed in Section 3. In Section 4, we provide experimental results on synthetic datasets to illustrate that the incremental approach delivers very similar results with batch MCA. The paper concludes in Section 5.

2 Dimension reduction for categorical data

Let \mathbf{Z} be a $n \times Q$ binary matrix, where n is the number of observations and Q the total number of categories that characterize q categorical variables. The general element is $z_{ij} = 1$ if the i^{th} statistical unit is characterized by the j^{th} category, $z_{ij} = 0$ otherwise; let $\mathbf{P} = \frac{1}{n \times q} \mathbf{Z}$ be the correspondence matrix, where $n \times q$ is the grand total of \mathbf{Z} . The core step of MCA is the matrix decomposition of the standardized residual matrix \mathbf{S} , defined as follows

$$\mathbf{S} = \mathbf{D}_{\mathbf{r}}^{-1/2} (\mathbf{P} - \mathbf{r}\mathbf{c}^{\top}) \mathbf{D}_{\mathbf{c}}^{-1/2}, \quad (1)$$

where \mathbf{r} and \mathbf{c} are the row and column margins of \mathbf{P} , respectively; $\mathbf{D}_{\mathbf{r}}$ and $\mathbf{D}_{\mathbf{c}}$ are diagonal matrices with values in \mathbf{r} and \mathbf{c} . The MCA solution can be obtained via the singular value decomposition (SVD) of $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$, where \mathbf{U} is a $n \times Q$ orthonormal matrix with left singular vectors on columns, $\mathbf{\Sigma}$ is a diagonal matrix containing the Q singular values, \mathbf{V} is a $Q \times Q$ matrix of right singular vectors. The j^{th} singular value corresponds to the standard deviation of data along the direction of j^{th} singular vector, $j = 1, \dots, Q$. The principal coordinates of the statistical units are $\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1/2} \mathbf{U}\mathbf{\Sigma}$, whereas $\mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1/2} \mathbf{V}\mathbf{\Sigma}$ are the attribute coordinates.

3 Association tracking

In order to describe an incremental algorithm for tracking association structures, i.e., when data arrive sequentially, we rewrite the standardized residual matrix of Eq. 1 as:

$$\mathbf{S} = \underbrace{\frac{\mathbf{Z}}{Q\sqrt{n}}\mathbf{D}_c^{-1/2}}_{\mathbf{X}_1} - \mathbf{1}_n \underbrace{\frac{1}{\sqrt{n}}\mathbf{1}_c^\top\mathbf{D}_c^{1/2}}_{\mu_1^\top}, \quad (2)$$

where $\mathbf{X}_1 = \frac{1}{Q\sqrt{n}}\mathbf{Z}\mathbf{D}_c^{-1/2}$ is the $n_1 \times Q$ row-wise centered matrix of the first data block and $\mu_1 = \frac{1}{\sqrt{n}}\mathbf{D}_c^{1/2}\mathbf{1}$ is the data mean. For an incoming data block, we obtain the corresponding $n_2 \times Q$ matrix \mathbf{X}_2 and calculate the eigenspaces of \mathbf{X}_1 and \mathbf{X}_2 using ordinary SVD, as $\Omega_1 = (n_1, \mu_1, \mathbf{U}_1, \Sigma_1, \mathbf{V}_1)$ and $\Omega_2 = (n_2, \mu_2, \mathbf{U}_2, \Sigma_2, \mathbf{V}_2)$.

In order to obtain the eigenspace Ω_3 of the super matrix $[\mathbf{X}_1^\top \ \mathbf{X}_2^\top]$, using uniquely the information in Ω_1 and Ω_2 , we adopt and briefly describe the incremental approach proposed by [4]. The total number of statistical units and the global data mean can be easily updated: $n_3 = n_1 + n_2$ and $\mu_3 = \frac{n_1\mu_1 + n_2\mu_2}{n_3}$.

In order to take into account the varying mean, the vector $\sqrt{\frac{nm}{n+m}}(\mu_2 - \mu_1)$ is added to the \mathbf{X}_2 matrix. Given the eigenspace of \mathbf{X}_1 , the projection \mathbf{L} of $\tilde{\mathbf{Y}}$ onto the orthogonal basis \mathbf{U}_1 is described by $\mathbf{L} = (\mathbf{U}_1)^\top \mathbf{X}_2$.

Let $\mathbf{H} = \mathbf{X}_2 - \mathbf{U}_1\mathbf{L}$ be the component of $\tilde{\mathbf{Y}}$ orthogonal to the subspace spanned by \mathbf{U}_1 . \mathbf{H} is decomposed such that an orthogonal matrix \mathbf{U}_h is obtained via a Gramm-Schmidt orthogonalization, as $\mathbf{U}_h = \text{orth}(\mathbf{X}_2 - \mathbf{U}_1\mathbf{L})$.

Then, $[\mathbf{X}_1^\top \ \mathbf{X}_2^\top] = ([\mathbf{U}_1 \ \mathbf{U}_h] \ \mathbf{U}_k) \Sigma_k \begin{pmatrix} \mathbf{V}_k^\top [\mathbf{V}_1 \ \mathbf{0}]^\top \\ \mathbf{0} \ \mathbf{I} \end{pmatrix}^\top,$

where $\mathbf{U}_k \Sigma_k \mathbf{V}_k^\top$ the SVD of $\mathbf{K} = \begin{bmatrix} \Sigma_1 & \mathbf{L} \\ \mathbf{0} & \mathbf{U}_h^\top \mathbf{H} \end{bmatrix}$.

Finally, $\mathbf{V}_3 = ([\mathbf{U}_1 \ \mathbf{U}_h] \ \mathbf{U}_k) \Sigma_k, \ \Lambda_3 = \Sigma_k, \ \mathbf{U}_3 = \begin{bmatrix} \mathbf{V}_1 \ \mathbf{0} \\ \mathbf{0} \ \mathbf{I} \end{bmatrix} \mathbf{V}_k.$

4 Experimental results

The proposed approach was applied to synthetic datasets in order to investigate its effectiveness to approximate the Batch MCA solution. In our experimental setup, each dataset could have 2 to 5 categories (attributes) per variable, randomly generated according to the uniform distribution model. Each experiment was defined by the following parameters:

- rows $\in \{10,000, 20,000, \dots, 100,000\}$, units of the original data matrix
- columns $\in \{50, 100, \dots, 500\}$, variables of the original data matrix

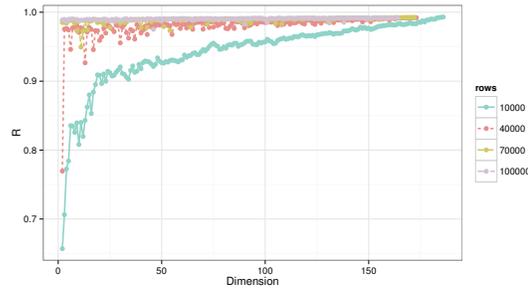


Fig. 1 Protocol setup:
rows = 10,000 to 100,000,
columns = 70, nb = 1000
rows, sb = 100

- $nb \in \{100, 200, \dots, 1000\}$, number of total data blocks
- $sb \in \{100, 200, \dots, 1000\}$, size of the starting data block.

The degree of similarity between Batch and Live MCA configurations was assessed via a Procrustes based similarity index, R , ranging from 0 to 1. Due to lack of space, we only discuss the effect of the number of rows. Figure 1 shows the evolution of similarity, as rows increase from 10,000 to 100,000, for increasing number of dimensions. It is evident that Batch and Live configurations become hardly different when the number of dimensions approaches its maximum value ($R > 0.99$).

5 Conclusions

A method has been proposed that makes possible the low-dimensional tracking of association structures characterizing categorical data flows. The method in question extends the applicability of MCA when data is not entirely available from the start. Such implementations become then feasible, for instance, for continuous monitoring of word associations that are present in data pulled on-the-fly from social networking sites, or for revealing and visualizing web-page visiting patterns via web-log analysis.

References

1. Baglama J. & Reichel L.: Augmented implicitly restarted Lanczos bidiagonalization methods. *Siam J Sci Comput* 27, 19–42 (2007)
2. Baker C., Gallivan K. & Van Dooren P.: Low-rank incremental methods for computing dominant singular subspaces. *Linear Algebra Appl* 436(8), 2866–2888 (2012)
3. Greenacre M.J.: *Correspondence Analysis in Practice*, second edition. Chapman and Hall/CRC (2007)
4. Ross D., Lim J., Lin R.S. & Yang M.H.: Incremental Learning for Robust Visual Tracking. *Int J Comput Vis* 77, 125–141 (2008)

Self-censored Categorical Responses

A device for recovering latent behaviors

Giulio D'Epifanio¹

Abstract This work deals with the problem of recovering the distribution of choices (eg in electoral polls) on a finite set of alternatives (eg political parties), using a randomly drawn sample from a certain population of deciders (eg electors) when the responses are censored (eg due to the intentional choice of non responding) on some of the sampled deciders. The identification problem, which arises in separating the latent behavior of interest (eg voting intention) from the self-selecting mechanism which influences it, is tackled through the “recursive identification” which is operatively implemented by a pseudo-Bayesian updating process in the Prior Feedback Setup (Casella & Robert, 2002).

Key words: Non-ignorable Categorical Non-response, Prior Feedback Setup, Self Selection, Survey Data

1 Introduction

For instance in electoral polls, the decision of responding may depend on the latent intention (“*what I actually say depends on what I would like to say*”). Thus, the reference data in table (1) (see [3]) may be interpreted on the sketched conceptual scheme. Here, on the population \mathcal{P} of the British electors, for any voter $i \in \mathcal{P}$, its latent choice is potentially reported by Y^* which takes values on the set of $L := 4$ competing parties: $\{“cons.”(1), “lab.”(2), “lib.”(3), “other.”(4)\}$. Therefore, Y^* denotes the hidden response provided by the “*latent phenomenon of interest*”. The decision of “self-exclusion in responding” is reported by $\chi \in \{0, 1\}$ so that, applied to any voter $i \in \mathcal{P}$, $\chi(i) = 0$ whenever $i \in \mathcal{P}$ refuses to respond, $\chi(i) = 1$ otherwise. Properly re-coded, the actually observed response is $Y := \chi \cdot Y^*$ so that $Y \in \{0(“nonresp.”), 1(“cons.”), 2(“lab.”), 3(“lib.”), 4(“other.”)\}$.

It is supposed that both the self-selection mechanism and the latent behavior of interest may be influenced by (professional and gender class based) conditions on a finite set $x \in X := \{x_1, x_2, \dots, x_{R:=10}\}$ of situations.

X: Sex	Social class	Y (intention of vote)				Intention unknown
		Conserv.	Labour	Liberal	Other	
		(cons.)	(lab.)	(lib.)	(other)	
Male	Professional	26	8	7	0	11
	Managerial and technical	87	37	30	6	64
	Skilled	66	77	23	8	77
	Semiskilled and unskilled	14	25	15	1	12
	Never worked	6	6	2	0	7
Female	Professional	1	1	0	1	2
	Managerial and technical	63	34	32	2	68
	Skilled	102	52	22	4	77
	Semiskilled and unskilled	10	32	10	2	38
	Never worked	20	25	8	2	19

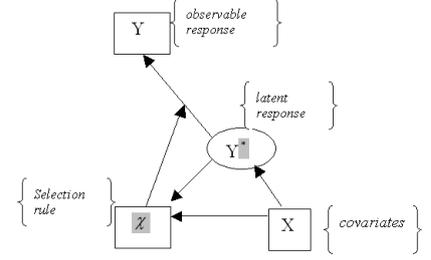


Table 1 (1) British general election panel survey. (2) Conceptual scheme

Based on scheme above, let us consider the data generating model¹ below.

$$Y_i := T_i * Y_i^* \quad (1)$$

$$T_i = \chi_i$$

$$\chi_i | y_i^*, x_i \sim_{ind} Bin(\chi_i; p_i(y_i^*, x_i)) \quad (2)$$

$$\begin{aligned} \logit p_i(y_i^*, x_i) := & \mu_0 + \sum_{l=2}^{L:=4} \delta_l \cdot I_{(l=y_i^*)} + \sum_{l=2}^{L:=4} \beta_l^{Sex} \cdot I_{(Sex(i)=\"{Fem.}\")} \cdot I_{(l=y_i^*)} + \\ & + \sum_{l=2}^{L:=4} \sum_{r=2}^{K:=5} \beta_{rl}^{SocClass} \cdot I_{(r=SocClass(i))} \cdot I_{(l=y_i^*)} \\ Y_i^* | \theta_i \sim_{ind} & Mult(y^*; \theta_{i1}, \dots, \theta_{il}, \dots, \theta_{iL}) \quad (3) \end{aligned}$$

$$\theta_i := (\theta_{i1}, \dots, \theta_{il}, \dots, \theta_{i4}) = (\theta_1, \dots, \theta_l, \dots, \theta_4)(x_i), \quad x_i \in State(X) := \{1, 2, \dots, r, \dots, R := 10\}$$

$$i := 1, \dots, n$$

¹ For any interviewed person $i := 1, \dots, n$, response Y depends on both latent response Y^* , which is generated by the “hidden phenomenon of our interest”, and the latent self-censoring decision $\chi \in \{0, 1\}$. Although redundant, we assume that χ is revealed by a sensor $T := \chi$ while χ is generated by the hidden self-selection process which is regulated by parameter p . But, p depends on both condition $X = r$, on $R := 10$ gender-social strata, and latent response Y^* itself. Latent intention Y^* is driven by parameter profile of interest θ which may depend on condition X . In eq (2), the reference base-line μ_0 is on the type of individual: “male professional voter of conserv. party”. Here, notation as $I_{(l=y_i^*)}$ denotes the (0/1)indicator-function such that $I_{(l=y_i^*)} = 1$ whenever $l = y_i^*$ (0 otherwise).

Unfortunately, inference on table $[\theta_{rl}]$, $\theta_{rl} := Pr\{Y^* = l | X = r\}$ is marked troublesome² in that censored response Y depends on the self-censoring mechanism which depends on table $[\theta_{rl}]$ itself. From the literature, pioneered by Little and Rubin, it emerges that the inferential problem above is inherently problematic (see [4] which used "data-dependent priors" in categorical setting, and see [3] for a full Bayesian approach).

2 Methodological outlines

Attempting to separate hidden table $[\theta_{rl}]$ from the unknown self-censoring mechanism, we will use the concept of "recursive identifiability" in the "prior feedback setup" of Robert (see [1]). In a pseudo-Bayesian setting, it is based on a proper "belief modeling" (see model (4)-(5)-(6) below) which re-interprets data-generating mechanism (1)-(2)-(3), at macro-level, through a set of constraints specifications about internal hidden mechanisms (self-censoring and voting intentions) on a space of hyper-parameters $\gamma \in \Gamma$. A formal device is proposed which, given a certain "*default assumption*" on the self-censoring mechanism (eg MCAR for reference), takes into input actual data (y, x) to provide in output the "belief representation", on $[\theta_{rl}]$ and self-censoring mechanism, which is not furtherly susceptible to be updated by data using a Bayesian updater (eg using posterior expectations). According to a "minimum information principle", recursively, a status γ^* of equilibrium in updating is searched from which a table could be recovered for $[\theta_{rl}]$ to be compared against the default one for sensitivity checks. Technically, based on the model below, the "knowledge space state" is represented by an hyper-parameter profile $\gamma := (\gamma_{sel}; \gamma_{beh}) \in \Gamma$ which is splitted in two parts such that the former identifies the self-selection ("sel") mechanism and the latter the latent behavior ("beh") of interest. Then, a recursive process is implemented which uses a double feedback in updating. It starts from a "*default assumption*" on self-censoring, specified by setting γ_{sel} . Then, it updates belief on $[\theta_{rl}]$ (by regulating γ_{beh}) given the current temporary assumption on self-censoring (having fixed current status of γ_{sel}); afterwards, it updates belief on the self-censoring mechanism (by regulating γ_{sel}) given the current temporary recovery of $[\theta_{rl}]$ (having fixed current status of γ_{beh}), and so on, until further updating is null. Thus, provided the process above converged, at the equilibrium configuration $\gamma^* := (\gamma_{sel}^*; \gamma_{beh}^*) \in \Gamma$ the "belief status" has been "recursively identified" for which actual data could not carry "innovative information" (on both the latent process of interest and the hidden self-censoring mechanism) at light of the set of constraints specifications on the assumed belief model. In practice, a powerful (constrained fixed point) recursive algorithm can be implemented which processes data for identifying

² unless the type of selection rule is "missing at random" (MAR), ie $\delta_l := 0, \beta_l^{Sex} := \beta^{Sex}$, $\beta_{rl}^{Sex} = \beta_r^{Sex}$, $l := 2, ..L$

that configuration $\gamma^* := (\gamma_{sel}^*; \gamma_{beh}^*) \in \Gamma$ for which the (perhaps, weighted) vectorial residuals from updating $\Delta(\gamma^*; x) := E(\psi|y, x; \gamma^*) - E(\psi|x; \gamma^*)$ (ie the variation between the vector of “conditional expectations given data at condition x ” against its unconditional counterpart) at γ^* is orthogonal to the constraints³ which were been assumed admissible for specifying the hidden mechanisms (self-censoring and parties choosing). The “belief model” below, interpreting data generating mechanism (1)-(2)-(3), is structured on two main layers⁴: the “observation outer-layer” (4) and the “internal-layer” (5).

$$Y_r := (Y_{r0}, Y_{r1}, \dots, Y_{r4}) \mid \psi_r := (\psi_{r0}, \psi_{r1}, \dots, \psi_{r4}) \stackrel{ind.}{\sim}_{r:=1, \dots, R} Mult(y_r; \psi_r, n_r) \quad (4)$$

$$\Psi_r \mid m_r, a_r \stackrel{ind.}{\sim}_{r:=1, \dots, R} Dirich(\psi_r; m_r, a_r) \quad (5)$$

$$m_r := (m_{r0}, m_{r1}, m_{r2}, m_{r3}, m_{r4}), \quad 0 < m_{rl} := E[\psi_{rl}] < 1, \quad a_r := w_r, \quad w_r > 0, \\ m_{r0} := (1 - p_r) \\ m_{rl} := q_{rl} * \theta_{rl}, \quad \text{party } l := 1, 2, 3, 4 \quad (6)$$

$$p_r := Pr\{\text{"non resp"} \mid x = r\} = \sum_{l=1}^4 q_{rl} \theta_{rl}, \quad q_{rl} > 0$$

$$q_{rl} := Pr\{\text{"non resp"} \mid Y^* = "l", x = r\} = Pr\{Y_r = "0" \mid Y_r^* = "l", x = r\}$$

$$\text{logit } q_{rl} = \mu_0 + \sum_{s=2}^{L:=4} \delta_l \cdot I_{(s=l)} + \sum_{s=2}^{L:=4} \beta_l^{Sex} \cdot I_{(Sex(r)="Fem.")} \cdot I_{(s=l)} + \\ + \sum_{s=2}^{L:=4} \sum_{SocClass w:=2}^{K:=5} \beta_{lw}^{SocClass} \cdot I_{(SocClass(r)=w)} \cdot I_{(s=l)}$$

$$\theta_{rl} := Pr\{Y^* = "l" \mid x = r\} = \varphi_{r(l-1)}, \quad 0 < \varphi_{r(l-1)} < 1, \quad l := 2, 3, 4 \\ \theta_{r1} := 1 - (\varphi_{r1} + \varphi_{r2} + \varphi_{r3})$$

$$r := x \in State(X) := \{1, 2, \dots, R := 10\}$$

References

- [1] Casella G, Robert C P (2002) Monte Carlo Statistical Methods (third printing). Springer, New York
- [2] D'Epifanio G (2005) Data Dependent Prior Modeling and Estimation in Contingency Tables. New York <http://www.springerlink.com/content/j73233521955n624/?p=6505e914841943a6bbeaf8cbcd144238b&pi=3>
- [3] Forster J.J., Smith P. W. F. (1998), Model-based Inference for Categorical Survey-data subject to Non-ignorable non-response, J. R. Statist. Soc. B, 60, pp. 57-70
- [4] Park T., Brown M.B., 1994, Models for Categorical data with Nonignorable Non Response, Journal of the American Statistical Association, Vol. 89, 44-52.

³ we interpreted (see [2]) Δ as a “gradient of updating”, in a geometric setting

⁴ The former relates actual categorical responses Y to parameters profile $\psi_{rs} := Pr\{Y_r = s\}$ which governs multinomials crossing social strata, n_r denotes the number of interviewees at stratum r . The latter, uses (as a convenient and practical belief-carrier model) a set of mean-parametrized Dirichlets on profiles $(\psi_1, \dots, \psi_r, \dots, \psi_R)$, whose mean-profile m is sub-structured to represent the self-censoring mechanism, which governs $q_{rl} := Pr\{\text{"non resp"} \mid Y^* = "l", x = r\}$ regulated by $\gamma_{sel} := \{(\mu_0, \delta, \beta^{Sex}, \beta^{SocClass})\}$, and the latent process of interest $[\theta_{rl}]$ regulated by $\gamma_{beh} := \{(\varphi_{r1}, \varphi_{r2}, \varphi_{r3}), r := 1, \dots, R\}$. Thus, $\gamma := (\gamma_{sel}; \gamma_{beh})$. The identification problem arises from eq (6).

Tourism Market Segmentation with Imprecise Information

Pierpaolo D'Urso, Marta Disegna, Riccardo Massari

Abstract Aim of the paper is to find homogeneous groups of visitors according to their satisfaction with different aspects of the visited destination. The case study used for this analysis is represented by 1,000 foreign tourists visiting the South-Tyrol region, in the Northern Italy, in 2010–2011. Since the segmentation variables are measured by a 10-points Likert-type scale, we formalize these data as fuzzy numbers. This formalization allows us to take into consideration the vagueness and uncertainty in evaluation tourists' judgments, influenced by the subjective meaning that one attributes to each value of a rating scale, i.e. by the heterogeneity in individual evaluation. Moreover, to deal with the uncertainty in assigning visitors to each segment, we adopt the Fuzzy C-Means Clustering Algorithm (FCM) for fuzzy data.

Key words: Fuzzy clustering method, Fuzzy number, Tourism.

1 Introduction

A review of the literature shows that customer satisfaction exerts a positive effect on both economic returns and brand loyalty ([1]). Overall satisfaction of tourists with a particular destination is a function of satisfaction with individual attributes, which characterize the destination and create the experience, influencing future intentions ([2], [3]) and expenditure behavior ([1]). Profiling visitors with respect to their satisfaction is of crucial importance for local policymakers, managers and mar-

Pierpaolo D'Urso
La Sapienza, Roma, Italy, e-mail: pierpaolo.durso@uniroma1.it

Marta Disegna
School of Economics and Management, Free University of Bolzano, Italy e-mail:
marta.disegna@unibz.it

Riccardo Massari
La Sapienza, Roma, Italy e-mail: riccardo.massari@uniroma1.it

keting analysts since developing and sustaining competitive advantage in competitive tourism markets largely depends upon the understanding of visitors needs ([4]; [5]). Therefore, market segmentation can provide important information to destination planners in order “to allocate resources more effectively in attracting distinct and unique groups of travellers” ([6]).

Since the introduction of market segmentation in the late 1950s, the number and type of approaches for segmentation has grown enormously [7, 8], dealing with different types of segmentation variables and different framework. When variables are ordinal (e.g., Likert scale), the items are often represented by linguistic expressions that are subjectively evaluated by respondents. [9] observed that subjective evaluations are best represented in a fuzzy framework, reflecting the uncertainty and the heterogeneity in individual evaluation.

The aim of this paper is to perform a segmentation analysis of the foreign visitors of South–Tyrol, Northern Italy, according to their satisfaction with the visit using a fuzzy segmentation approach to fuzzy data.

2 The tourism case study

The dataset for our study is drawn from with the annual inbound survey “International Tourism in Italy”, conducted by the Bank of Italy since 1996.

We focus our case study on 997 foreign visitors who visited the province of Bolzano, in the South–Tyrol region (Northern Italy), during 2010 and 2011 and whose main purpose was “tourism, holiday, leisure”. The segmentation variables used in this paper regard the evaluation of tourists’ satisfaction linked to different aspects of the visited destination. Satisfaction is analyzed through 9 different aspects. The overall satisfaction with the destination is also collected. All items are measured by asking respondents to report their level of satisfaction with each aspect considered on a 10–point Likert–type scale ranged from “1” (*Very unsatisfied*) to “10” (*Very satisfied*). Likert–type scales are a linguistic (qualitative) common way to investigate on subjective evaluation of tourists, and in general individuals.

3 Methodology

Likert–type variables entail a certain degree of imprecision, due to the subjective meaning that one attributes to each value of a rating scale. To cope with this inaccuracy in the measurement of the phenomena of interest, we formalize linguistic data in terms of fuzzy numbers [9]. A general class of fuzzy data, called LR fuzzy data, can be defined in a metric form as follows [10]:

$$\tilde{\mathbf{X}} \equiv \{\tilde{x}_{ik} = (c_{ik}, l_{ik}, r_{ik})_{LR} : i = 1, \dots, N; k = 1, \dots, K\}, \quad (1)$$

where $\tilde{x}_{ik} = (c_{ik}, l_{ik}, r_{ik})_{LR}$ denotes the *LR* fuzzy variable k observed on the i -th unit; c_{ik} indicates center, i.e. the “core” of the fuzzy number; l_{ik} and r_{ik} represent the left and right spread. In particular, we consider the “triangular” *LR* fuzzy variable.

In our framework, there is another source of uncertainty, due to the assignment of units to clusters when dealing with complex structure of data, such as that at hand. This source of uncertainty can be tackled with the adoption of fuzzy clustering, in which units are assigned to each cluster with a membership degree that represents a measure of the level of uncertainty (vagueness) in the assignment process. In particular, we consider the Fuzzy *C*–Means Clustering Algorithm (FCM) [11].

Since we deal with fuzzy variables, we make use of the distance measure proposed by [12] which takes into account the vagueness in the observed data:

$$d_F^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = [w_C^2 (\|\mathbf{c}_i - \mathbf{c}_{i'}\|^2) + w_S^2 (\|\mathbf{l}_i - \mathbf{l}_{i'}\|^2 + \|\mathbf{r}_i - \mathbf{r}_{i'}\|^2)], \quad (2)$$

where $\tilde{\mathbf{x}}_i \equiv \{\tilde{x}_{ik} = (c_{ik}, l_{ik}, r_{ik})_{LR} : k = 1, \dots, K\}$ denote the fuzzy data vector for the i -th unit; \mathbf{c}_i , \mathbf{l}_i and \mathbf{r}_i are the vectors of the centers and of the left and right spreads, respectively; $\|\mathbf{c}_i - \mathbf{c}_{i'}\|^2$ is the squared Euclidean distances between the centers; $\|\mathbf{l}_i - \mathbf{l}_{i'}\|^2$ and $\|\mathbf{r}_i - \mathbf{r}_{i'}\|^2$ are the squared Euclidean distances between the left and right spread, respectively; $w_C, w_S \geq 0$ are suitable weights for the center component and the spread component of (2), constrained by the following conditions: $w_C + w_S = 1$ (*normalization condition*) and $w_C \geq w_S \geq 0$ (*coherence condition*) ([12]). Using the distance (2) into the FCM algorithm, we obtain the FCM algorithm for fuzzy data ([12]).

Note that the prototypes obtained with the FCM algorithm for fuzzy data are of *LR* fuzzy type, inheriting their typology by the observed data ([12]).

4 Results and conclusions

The aim of this paper is to find homogeneous groups of foreign visitors of South–Tyrol according to their satisfaction linked to different aspects of the destination taking into consideration the uncertainty of the subjective evaluation collected through linguistic (qualitative) variables. The FCM algorithm for fuzzy data was adopted in order to capture this uncertainty and to allow a more flexible allocation of units to each cluster. Indeed, some units can be fuzzy allocated to more than one cluster, if their characteristics are compatible with the profile of different clusters, a situation that cannot be detected with crisp clustering method. As a result we have obtained three clusters of foreign visitors of about the same size: the least satisfied, those in the middle, and the enthusiasts (see Figure 1: dotted lines represent the uncertainty in subjective evaluations). In order to profile these clusters we can use the remaining variables collected by the survey and regarding: socio–demographic characteristics of the interviewees; information on the trip; information on the travel expenditure.

References

1. Zhang, L., Qu, H., Ma, J.(E.): Examining the Relationship of Exhibition Attendees' Satisfaction and Expenditure: The Case of Two Major Exhibitions in China. *Journal of Convention & Event Tourism* **11:2**, 100–118 (2010)
2. Kozak, M., Rimmington, M.: Tourist satisfaction with Mallorca, Spain, as an off-season holiday destination. *Journal of Travel Research* **38:3**, 260–269 (2000)
3. Kim, Y.G., Suh, B.W., Eves, A.: The relationships between food-related personality traits, satisfaction, and loyalty among visitors attending food events and festivals. *International Journal of Hospitality Management* **29**, 216–226 (2010)
4. Lee, J., Beeler, C.: An Investigation of Predictors of Satisfaction and Future Intention: Links to Motivation, Involvement, and Service Quality in a Local Festival. *Event Management* **13:1**, 17–29 (2009)
5. Koc, E., Altinay, G.: An analysis of seasonality in monthly per person tourist spending in Turkish inbound tourism from a market segmentation perspective. *Tourism Management* **28:1**, 227–237 (2007)
6. Kau, A.K., Lim, P.S.: Clustering of Chinese Tourists to Singapore: An Analysis of Their Motivations, Values and Satisfaction. *International Journal of Tourism Research* **7**, 231–248 (2005)
7. Liao, S.H., Chu, P.H., Hsiao, P.Y.: Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications* **36**, 11772–11781 (2012)
8. Dolnicar, S., Leisch, F.: Segmenting markets by bagged clustering. *Australasian Marketing Journal* **12:1**, 51–65 (2004)
9. Coppi, R., D'Urso, P.: Fuzzy k-mean clustering models for triangular fuzzy time trajectories. *Statistical Methods and Applications* **11**, 21–24 (2002)
10. Dubois, D., Prade, H.: Possibility theory, Plenum press, New York (1988).
11. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA (1981)
12. Coppi, R., D'Urso, P., Giordani, P.: Fuzzy and possibilistic clustering for fuzzy data. *Computational Statistics & Data Analysis* **56:4**, 915–927 (2012)

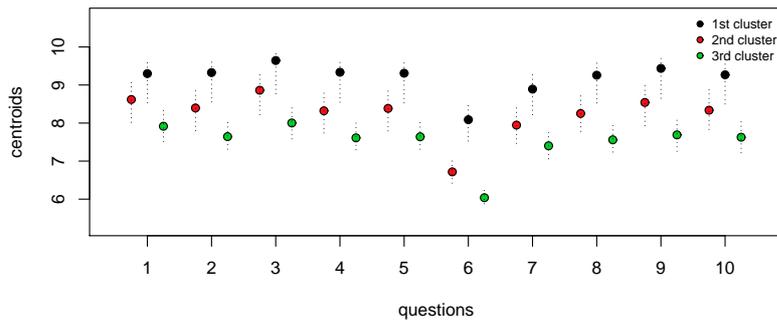


Fig. 1 The three clusters solution.

Cluster-weighted models for multivariate response and extensions

Utkarsh J. Dang, Salvatore Ingrassia, Paul D. McNicholas and Ryan Browne

Abstract A multivariate extension to cluster-weighted modeling that can deal with multivariate response is proposed. An expectation-maximization algorithm for maximum likelihood estimation of the parameters in the model is presented. A parsimonious family of models is also proposed using an eigen-decomposed covariance structure.

Key words: Cluster-weighted models, multivariate data, EM-algorithm, mixture models

1 Introduction

Mixture models have seen increasing use over the last decade or so with important applications in clustering and classification. Various mixture-based methods have emerged for clustering multivariate data. Arguably, the most famous model-based clustering methodology is the Gaussian parsimonious clustering models (GPCM) family (GPCM; Celeux and Govaert, 1995), which is supported by the `mclust` (Fraley and Raftery, 1999), `mixture` (Browne and McNicholas, 2013), and `Rmixmod` (Lebet et al., 2012) packages for R.

Incorporating a regression structure can yield important insight when there is a clear regression relationship between some variables. Here, traditional model-based

Utkarsh J. Dang
University of Guelph, Canada. e-mail: udang@uoguelph.ca

Salvatore Ingrassia
Università di Catania, Italy. e-mail: s.ingrassia@unict.it

Paul D. McNicholas
University of Guelph, Canada. e-mail: paul.mcnicholas@uoguelph.ca

Ryan Browne
University of Guelph, Canada. e-mail: rbrowne@uoguelph.ca

clustering methods that fail to take into account such a relationship may not perform as well. Some methodologies that deal with such regression data include finite mixtures of regressions (FMR; DeSarbo and Cron, 1988; Leisch, 2004) and finite mixtures of regression with concomitant variables (FMRC; Wedel, 2002). However, these methodologies do not explicitly model the distribution of the covariates.

Ingrassia et al. (2012) introduced cluster-weighted modeling (CWM) in a general statistical mixture framework. As opposed to FMR and FMRC, both the distribution of the response given the covariates and the covariates is modelled. The joint probability of the univariate response and multivariate covariates is decomposed in an expectation-maximization (EM) framework (Dempster et al., 1977) for parameter estimation. Ingrassia et al. (2012) presented theoretical and numerical properties and discussed the performance of the model under both Gaussian and student-t distributional assumptions.

Here, we present an extension of CWM that can deal with multivariate response vectors. Parameter estimation is done in the EM framework. A more parsimonious version of the model will also be developed and results will be demonstrated on some simulated and real data sets.

2 Methodology

Let \mathbf{X} and \mathbf{Y} be random vectors defined on Ω with joint probability distribution $p(\mathbf{x}, \mathbf{y})$. Here, the response vector \mathbf{Y} has values in $\mathcal{Y} \subseteq \mathbb{R}^d$ and the explanatory vector \mathbf{X} has values in $\mathcal{X} \subseteq \mathbb{R}^p$. Let Ω be partitioned into G disjoint groups, such that $\Omega = \Omega_1 \cup \dots \cup \Omega_G$. Then, in a CWM framework, the joint probability $p(\mathbf{x}, \mathbf{y})$ can be decomposed as

$$p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) = \sum_{g=1}^G p(\mathbf{y} | \mathbf{x}, \Omega_g) p(\mathbf{x} | \Omega_g) \pi_g,$$

where $p(\mathbf{y} | \mathbf{x}, \Omega_g)$ is the conditional density of \mathbf{y} given \mathbf{x} and Ω_g , $p(\mathbf{x} | \Omega_g)$ is the probability density of \mathbf{x} given Ω_g , and $\pi_g = p(\Omega_g)$ is the mixing weight, where $\pi_g > 0$ ($g = 1, \dots, G$) and $\sum_{g=1}^G \pi_g = 1$. $\boldsymbol{\theta}$ denotes the set of all parameters.

2.1 Parameter Estimation

Let $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$ be a sample of N independent observations. Then, the incomplete likelihood function under Gaussian distributional assumptions is

$$L_0(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^N p(\mathbf{x}_i, \mathbf{y}_i|\boldsymbol{\theta}) = \prod_{i=1}^N \left[\sum_{g=1}^G \phi_d(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\chi}_g) \phi_p(\mathbf{x}_i|\boldsymbol{\psi}_g) \pi_g \right].$$

Here, ϕ_d and ϕ_p denote the probability density functions for a d and p dimensional multivariate Gaussian distribution, respectively. $\boldsymbol{\chi}_g = (\mathbf{B}_g, \boldsymbol{\Sigma}_{yg})$ refers to the parameters of the distribution of the response given the covariates where \mathbf{B}_g is the matrix of regression coefficients. Similarly, $\boldsymbol{\psi}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{xg})$ denotes the mean and covariance for the predictor vector. Note that $(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{y}_1, \dots, \mathbf{y}_N)$ are considered incomplete data. The complete data are $(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{y}_1, \dots, \mathbf{y}_N, \mathbf{z}_1, \dots, \mathbf{z}_N)$ where the missing variable z_{ig} is the component label vector such that $z_{ig} = 1$ if $(\mathbf{x}_i, \mathbf{y}_i)$ comes from the g th population and 0 otherwise. Now, the corresponding complete-data likelihood is

$$L_c(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^N \prod_{g=1}^G [\phi_d(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\chi}_g) \phi_p(\mathbf{x}_i|\boldsymbol{\psi}_g) \pi_g]^{z_{ig}}.$$

The complete-data log-likelihood function can be decomposed as

$$\mathcal{L}_c(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N \sum_{g=1}^G [z_{ig} \log \phi_d(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\chi}_g) + z_{ig} \log \phi_p(\mathbf{x}_i|\boldsymbol{\psi}_g) + z_{ig} \log \pi_g]$$

The E-step involves calculating the expected complete data log-likelihood:

$$\tau_{ig}^{(k)} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}} \{z_{ig}|\mathbf{x}_i, \mathbf{y}_i\} = \frac{\pi_g^{(k)} \phi_d(\mathbf{y}_i|\mathbf{x}_i, \mathbf{B}_g^{(k)}, \boldsymbol{\Sigma}_{yg}^{(k)}) \phi_p(\mathbf{x}_i|\boldsymbol{\mu}_g^{(k)}, \boldsymbol{\Sigma}_{xg}^{(k)})}{\sum_{j=1}^G \pi_j^{(k)} \phi_d(\mathbf{y}_i|\mathbf{x}_i, \mathbf{B}_j^{(k)}, \boldsymbol{\Sigma}_{yj}^{(k)}) \phi_p(\mathbf{x}_i|\boldsymbol{\mu}_j^{(k)}, \boldsymbol{\Sigma}_{xj}^{(k)})}$$

provides the current value on the k -iteration. The M-step on the $(k+1)$ th iteration of the EM algorithm involves the maximization of the conditional expectation of the complete-data log-likelihood with respect to $\boldsymbol{\theta}$. The updates for $\pi_g^{(k+1)}$, $\boldsymbol{\mu}_g^{(k+1)}$ and $\boldsymbol{\Sigma}_{xg}^{(k+1)}$, $g = 1, \dots, G$, are

$$\begin{aligned} \hat{\pi}_g^{(k+1)} &= \frac{1}{N} \sum_{i=1}^N \tau_{ig}^{(k)}, \\ \hat{\boldsymbol{\mu}}_g^{(k+1)} &= \frac{\sum_{i=1}^N \tau_{ig}^{(k)} \mathbf{x}_i}{\sum_{i=1}^N \tau_{ig}^{(k)}}, \\ \hat{\boldsymbol{\Sigma}}_{xg}^{(k+1)} &= \frac{\sum_{i=1}^N \tau_{ig}^{(k)} (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{i=1}^N \tau_{ig}^{(k)}}. \end{aligned}$$

These closed form updates can also be found in McLachlan and Peel (2000). The updates for $\mathbf{B}_g^{(k+1)}$ and $\boldsymbol{\Sigma}_{yg}^{(k+1)}$, $g = 1, \dots, G$, are

$$\hat{\mathbf{B}}_g^{(k+1)} = \sum_{i=1}^N \tau_{ig}^{(k)} \mathbf{y}_i \mathbf{x}_i' \times \left(\sum_{i=1}^N \tau_{ig}^{(k)} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}.$$

$$\hat{\boldsymbol{\Sigma}}_{yg}^{(k+1)} = \frac{\sum_{i=1}^N \tau_{ig}^{(k)} (\mathbf{y}_i - \mathbf{B}'_g \mathbf{x}_i) (\mathbf{y}_i - \mathbf{B}'_g \mathbf{x}_i)'}{\sum_{i=1}^N \tau_{ig}^{(k)}}.$$

2.2 Ongoing work

Note that the number of parameters for both component covariance matrices increases quadratically. Here, the covariance matrices $\boldsymbol{\Sigma}_{xg}$ and $\boldsymbol{\Sigma}_{yg}$ can be decomposed to give families of mixture models. We will perform an eigen-decomposition of these matrices in the fashion of Gaussian parsimonious clustering models (Celeux and Govaert, 1995). The algorithm will be run on simulated and real data in R (R Development Core Team, 2011) to evaluate the performance of the model.

Acknowledgements This work is supported by a Alexander Graham Bell Canada Graduate Scholarship (CGS-D), a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and a travel grant from the Università di Catania, Italy.

References

- Browne, R. P. and P. D. McNicholas (2013). *mixture: Mixture Models for Clustering and Classification*. R package version 1.0.
- Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern recognition* 28(5), 781–793.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B* 39, 1–38.
- DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification* 5(2), 249–282.
- Ingrassia, S., S. C. Minotti, and G. Vittadini (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification* 29(3), 363–401.
- Lebet, R., S. Iovleff, and A. Longeville (2012). *Rmixmod: Mixture Modelling Package*. R package version 1.0.
- Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11(8), 1–18.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica* 56(3), 362–375.

Unsupervised Classification through Quantile Regression

Cristina Davino, Domenico Vistocco

Abstract This paper aims to propose an innovative approach to identify group effects through a quantile regression model. Quantile regression is a quite recent regression technique that allows to focus on the effects that a set of explanatory variables has on the entire conditional distribution of a dependent variable. The proposal concerns the use of multivariate techniques to detect effects attributable to different group membership and it is illustrated using an empirical analysis. In particular the impact of student features on the University outcome measured by the degree mark is evaluated taking into account that the dependence structure could be different according to the Faculty membership.

Key words: Quantile regression, unsupervised classification

1 Introduction

The relationship between a response variable and a set of explanatory variables can be different if units belong to different groups. The aim of the paper is to propose an innovative approach based on the conjoint use of multivariate methods and quantile regression to identify a typology in a dependence model. The proposed approach provides an unsupervised classification able to take into account the dependence structure. The methodological framework is represented by quantile regression, as introduced by Koenker and Basset (1978). This method is an extension of the classical estimation of the conditional mean to the estimation of a set of conditional

Cristina Davino

University of Macerata - Department of Political Sciences, Communications and Intern. Relations
62100 Macerata, Italy e-mail: cdavino@unimc.it

Domenico Vistocco

University of Cassino - Department of Economics and Law
03043 Cassino (FR), Italy e-mail: vistocco@unicas.it

quantiles. It offers a complete view of a response variable providing a method for modelling the rates of changes at multiple points (conditional quantiles) of its conditional distribution (Davino *et al.*, 2013). This paper is an extension of the supervised approach provided by the authors (Davino and Vistocco, 2007, 2008). What characterizes this proposal is the use of an unsupervised approach to classify units in the dependence model, whereas the former proposals exploit a pre-specified stratification variable.

In the following section the methodology is described together with results deriving from a real data application aiming at grouping students according to the relationship between the degree mark and their features.

2 The proposed approach: data, methodology and results

The proposed unsupervised classification is described exploiting an empirical analysis. The application aims to evaluate if and how the student features (socio-demographic and University experience attributes) affect the outcome of the University career, measured through the degree mark. As stated above, the underlying idea is that this effect can be very different if the students belong to different groups.

The evaluation of the factors influencing the degree mark is based on a random sample of 685 students graduated at University of Macerata (Davino and Vistocco, 2007) which is located in the Italian region Marche. The following features of the student profile have been observed: gender, place of residence during University education (Macerata and its province, Marche region, outside Marche), course attendance (no attendance, regular), foreign experience (yes, no), working condition (full time student, working student), number of years to get a degree, diploma mark. The density plot of the response variable (Fig. 1, left-hand side) shows the presence of a strong right skewness.

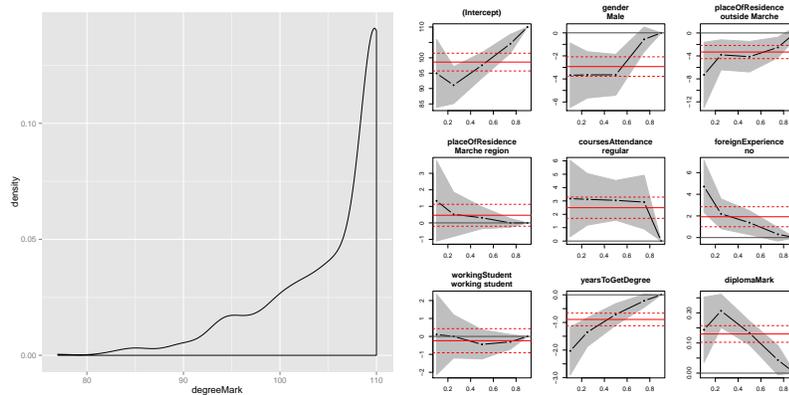


Fig. 1 Degree mark density (left-hand side) and QR coefficients (right-hand side).

The proposed unsupervised classification is based on the use of multivariate methods and quantile regression. The approach is structured in the following four

steps: *i*) global estimation, *ii*) identification of the best model for each unit, *iii*) identification of the groups and of the best model for each group, *iv*) partial estimation. Each step is detailed in the following.

i) In the first step, a QR model is estimated on the whole sample:

$$Q_\theta(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta(\theta) + \mathbf{e} \quad (1)$$

where $0 < \theta < 1$, $Q_\theta(\cdot|\cdot)$ denotes the conditional quantile function for the θ^{th} quantile, \mathbf{y} the dependent variable (degree mark) and \mathbf{X} the matrix of the explanatory variables (students features).

In Fig. 1 (right-hand side), QR coefficients are graphically represented for the different features of the student profile. The horizontal axis displays the different quantiles while the effect of each feature holding constant the others is represented on the vertical axis. The shaded region in each subplot shows the confidence band ($\alpha=0.1$). Finally, the lines parallel to the horizontal axis correspond to LS coefficients, the related confidence intervals are in dashed lines using the same level for α . The graphical representation allows to visually catch the different effect of the student characteristics on the degree mark: the coefficients related to males, place of residence outside Marche and years employed to get a degree are negative even if increasing moving from lower to upper quantiles. On the other hand, place of residence in Marche region, diploma mark and foreign experience play a positive but decreasing effect. Moreover a regular course attendance and working condition have a slighter decreasing effect on the degree mark.

ii) In the second step, the coefficients matrix $\hat{\Theta}_{[p \times k]}$ (where p is the number of explanatory variables and k the number of estimated conditional quantiles) and the regressors data matrix \mathbf{X} are used to estimate the conditional distribution matrix of the response variable: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\Theta}$. The generic element $\hat{y}_{i\theta}$ of the $\hat{\mathbf{Y}}_{[n \times k]}$ matrix represents the value of the estimated value for the i^{th} units according to the θ^{th} quantile. In the proposed empirical analysis, quantiles vary from 0.01 to 0.99 with a step of 0.01. The best model for each unit i is identified by the quantile able to better estimate the response variable, namely through the quantile which minimize the absolute difference between the observed value and the estimated density value:

$$\hat{\theta}_i^{best} : \underset{\theta=1, \dots, k}{\operatorname{argmin}} |y_i - \hat{y}_{i\theta}| \quad (2)$$

The $\hat{\theta}^{best}$ vector allows to extract from the $\hat{\mathbf{Y}}$ matrix the best estimated value for each unit: \hat{y}_θ^{best} . For some considerations on the added value provided by considering \hat{y}_θ^{best} instead of the classical LS predicted values, the interest reader is referred to Davino and Vistocco (2008).

iii) The third step of the proposed strategy aims to identify a typology on the basis of the QR results. Units are grouped according to the best quantile they have been assigned in step *ii*) because it can be considered as an indicator of a similar dependence structure. Different criteria can be followed to identify the clusters, for example dividing the $\hat{\theta}^{best}$ vector in classes (equal frequency, width, etc.) albeit each of them introduces a certain degree of subjectivity.

In the present paper a multivariate approach is proposed performing a cluster analysis on the estimated \hat{Y} matrix in order to classify units sharing similar patterns for the predicted values for all the considered quantiles. In the specific application, three clusters seems the best partition. For each group, the median of $\hat{\theta}^{best}$ has been considered a robust reference quantile: $G1=0.75$, $G2=0.59$, $G3=0.53$.

iv) In the last step, QR is again executed on the total sample but considering only the three reference quantiles previously defined. Fig. 2 shows the observed response distribution for group G1 compared with the estimated distributions of all the groups obtained using the G1 reference quantile.

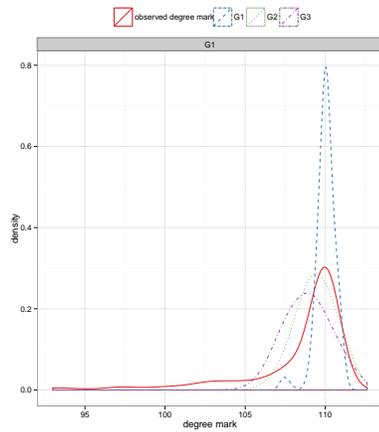


Fig. 2 Density estimations of the first group against the other groups.

The proposed approach can represent a valid tool to cluster units taking into account the dependence structure in the data. It is based on the observed similarity among units in terms of their conditional quantile patterns and its main advantage relies on the automatic and multivariate identification of the groups.

References

- Davino C., Furno M., Vistocco D. (2013) *Quantile Regression: Theory and Applications*. Wiley.
- Davino C., Vistocco D. (2007) The evaluation of University educational processes: a quantile regression approach. *STATISTICA*, n.3, pp. 267-278.
- Davino C., Vistocco D. (2008) Quantile regression for the evaluation of student satisfaction. *Statistica Applicata* **20**, 179–196.
- Koenker R. (2005) *Quantile Regression*. Econometric Society Monographs, Cambridge University Press.
- Koenker R., Basset G.W. (1978) Regression Quantiles, *Econometrica*, 46, 33-50.

A copula-based approach to discover inter-cluster dependence relationships

F. Marta L. Di Lascio and Simone Giannerini

Abstract In this work we focus on clustering dependent data according to the multivariate structure of the generating process. Di Lascio and Giannerini [1] proposed an algorithm based on copula function [3], called CoClust, that accomplishes this task. The CoClust has a good performance in many different scenarios [1]. However, it allocates all the observations and it has an high computational burden. In this work we propose a modified version of the CoClust that overcomes these problems. By means of a detailed Monte Carlo study we compare the two algorithms and prove the superiority of the new version of the CoClust over the old version.

Key words: CoClust algorithm, Copula function, Dependence structure.

1 Introduction

Clustering is a useful exploratory technique for multivariate data as it allows the identification of potentially meaningful relationships between objects. The extensive literature of clustering includes both methods based on distance/dissimilarity and methods based on probability models. In practice, all such methods are able to cope with pairwise and linear relationships and are not suitable to model multivariate complex dependence. To our knowledge, the first attempt to overcome these limits in the context of model-based clustering techniques was made in [1] by proposing a copula-based clustering algorithm, called CoClust. Copulas [3] are multivariate distribution functions that allow to describe a variety of complex multivariate dependence structures without any assumptions on the margins. The CoClust produces a clustering through a joint density function defined via copula; the dimension of the copula is the number of clusters since each cluster is represented by a (marginal) uni-

F. Marta L. Di Lascio, School of Economics and Management, Free University of Bozen-Bolzano, Italy, e-mail: marta.dilascio@unibz.it · Simone Giannerini, Department of Statistical Sciences, University of Bologna, Italy, e-mail: simone.giannerini@unibo.it

variate density function. The CoClust can model complex dependence relationships between observations and focuses on the inter-cluster dependence relationship: in fact observations in different clusters are dependent while observations in the same cluster are i.i.d. realizations from the same marginal distribution. The CoClust performs very well in many different scenarios ([1, Section 4]). However, in its present version it has some shortcomings, e.g. it is not able to discard irrelevant observations and it has an high computational burden. In this work we propose a modified version of the CoClust that overcomes these problems.

In Section 2 we briefly describe the proposal while in Section 3 we give some results from a simulation study. Conclusions are outlined in Section 4.

2 CoClust algorithm

The algorithm we propose inherits all the benefits of the original version of the CoClust, e.g. it does not require either to choose a starting classification or to set a priori the number of clusters. Moreover, our proposal allows to overcome the drawbacks of the original version: in fact the new algorithm allows to discard observations that do not belong to the clustering and it is computationally feasible.

In brief, at the first step, the algorithm selects the candidate observations through pairwise dependence measures. This solution allows to lower considerably the computational complexity if compared to the algorithm in [1]. Also, observations are classified through a criterion based on the log-likelihood of a copula fit. The copula is estimated through a semi-parametric version of the inference for margins estimation method [2]; for technical details see [1, Section 2.3]. Starting from a data matrix of size $G \times S$, the procedure treats each row of such matrix as a single element to be allocated to a cluster. At the first two steps the algorithm selects the optimal number of clusters K and from the second step onwards, it allocates the remaining observations/rows to the K clusters. The procedure can be summarized as follows:

1. for $k = 1, \dots, K_{max}$, where $K_{max} \leq G$, specified by the user, is the maximum number of clusters, select a subset of n_k k -plets on the basis of the Spearman's rank correlation coefficients;
2. the subset of k -plets that maximizes the log-likelihood copula function is selected as the starting clustering; this allow to select the number of clusters K ;
3. select the K -plet with the maximum Spearman's correlation coefficient, compute all its permutations and estimate $K!$ copulas by using the observations already clustered and the candidate K -plet;
4. if there is a permutation that increases the log-likelihood of the copula fit, then allocate the K -plet to the clustering by assigning each observation to the corresponding cluster, otherwise drop it;
5. repeat the last two steps until all the observations are either allocated or discarded.

At the end of the procedure we obtain K clusters whose dependence is modelled by a K -dimensional copula function.

3 A Simulation study

We evaluate the performance of our proposal by *i*) comparing it with the original version of the CoClust algorithm and *ii*) investigating its capability to discard observations irrelevant to the final clustering. We use the following performance measures:

1. $p.n.c.$: percentage of Monte Carlo replications for which the identified number of clusters is correct;
2. $SEN.o$: sensitivity for observations; percentage of single observations belonging to the data generating process (DGP hereafter) that are correctly allocated;
3. $SEN.k$: sensitivity for k -plets; percentage of k -plets belonging to the DGP that are correctly allocated;
4. $\hat{\theta}_2^*$: average over replications of the post-clustering estimates of the dependence parameter;
5. $r.p.$: rejection percentage of the null hypothesis $H_0 : \theta_2 = 0$ with $\alpha = 0.05$.

As for the comparison with the original version of the CoClust we consider the simulation experiments in [1, Section 4]. Note that the $SEN.k$ coincides with the measure $p.c.a$ in [1, Section 4]. The sample size n is set to 30 and $K = 3$ for all the scenarios investigated. Table 1 shows the results of our proposal, some information on the simulated DGP and the kind of copula model used in the algorithm. If we

Table 1 CoClust performance: comparing our proposal with the previous version of the CoClust.

DGP	Margins	Copula	$p.n.c.$	$SEN.k$	$SEN.o$	$\hat{\theta}$	$r.p.$
Frank copula	Gamma, Beta, Gaussian	Frank	100	99.50	99.83	9.98	100
Skew-Normal	Skew-Normal	Clayton	92.50	82.03	84.46	1.12	100
Mixed Gaussian	Gaussian	Clayton	71.25	79.00	78.30	0.90	100

compare table 1 with tables 6,7,8 in [1] we can argue that our proposal clearly overcomes its previous version with respect to all the performance measures. In all the three scenarios investigated the new version improves the allocation of both k -plets and observations as well as the identification of the correct number of clusters.

Next, we simulate from a trivariate Skew-Normal DGP with skew-normal margins and from an independent trivariate normal copula such that the 30% of the total number of observations are independent. Table 2 shows the results of this set of

simulations. The performance of our proposal is satisfactory also when independent data from a different distribution contaminate the data set.

Table 2 Trivariate Skew-Normal DGP plus 30% of independent observations.

DGP	Margins	Copula	n	$p.n.c.$	$SEN.k$	$SEN.o$	$\hat{\theta}$	$r.p.$
Skew-Normal	Skew-normal	Gumbel	60	83.75	66.31	71.00	7.47	100
			90	81.25	59.45	63.54	7.44	100
			120	80.00	56.03	60.19	7.43	100

4 Conclusions

In this paper we have proposed a clustering algorithm based on copula functions. Being based on copulas, our proposal can account for complex dependence relationships between observations and it is able to group observations according to their underlying dependence structure. Moreover, the algorithm tested on simulated data, is able to *i*) identify the true number of clusters, *ii*) correctly allocate the *k*-plets and *iii*) discard irrelevant observations; hence, the new CoClust is a significant improvement with respect to the original version.

Our proposal can be further investigated by analyzing its performance when the data set is composed by more than one dependence structure. Moreover, the introduction of a model selection criteria for copula into the algorithm will be investigated.

References

1. Di Lascio, F.M.L., Giannerini, S.: A Copula-Based Algorithm for Discovering Patterns of Dependent Observations. *J. Classif.* **29(1)**, 50-75 (2012)
2. Joe, H., Xu, J.: The estimation method of inference functions for margins for multivariate models, Tech. Rep., **166**, Dept of Statistics, Univ. of British Columbia (1996)
3. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges, *Pub. Inst. Stat. Univ. Paris* **8**, 229-231 (1959)

Acknowledgements F. Marta L. Di Lascio acknowledges the support of the School of Economics and Management, Free University of Bozen-Bolzano, Italy.

Hierarchical market structure of Euro area regime dynamics

José G. Dias and Sofia B. Ramos

Abstract This paper analyzes country versus industry factors in European stock market dynamics. This dichotomy is framed as a hierarchical market structure by a two-step approach: first, time series indexes are filtered using a regime switching model; then, based on the Kullback-Leibler distance between posterior probabilities, the hierarchical market structure is revealed. Time series of 79 industry-country indexes from ten countries of the euro area show that industry factors are more important in explaining stock market heterogeneity.

Key words: clustering, hidden Markov model, stock markets

1 Introduction

Country versus industry factors has been subject of much research by academics and practitioners in the last decades due to the recognition of its importance in top-down diversification strategies [1, 2, 3]. This paper revisits this trade-off using a new methodology that takes into account the switching dynamics between regimes.

José G. Dias
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (UNIDE-IUL), Edifício ISCTE, Av. das Forças Armadas, 1649-026 Lisbon, Portugal e-mail: jose.dias@iscte.pt

Sofia B. Ramos
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (UNIDE-IUL), Edifício ISCTE, Av. das Forças Armadas, 1649-026 Lisbon, Portugal e-mail: sofia.ramos@iscte.pt

2 Methodology

Our methodology combines regime switching models in the modeling of longitudinal variations with cluster analysis that identifies groups of industries with similar profiles. This procedure extends regime-switching models (RSM), introduced by Hamilton [4]. RSMs have been quite influential and extensively used in empirical research, and finance research is no exception. There are good reasons to incorporate market regimes into the modeling and analysis of financial markets. We set a two-step approach: first, using a panel Markov regime switching model, the panel data (raw data) is transformed into a set of probabilities (posterior probabilities); then, a measure of distance between the posterior probabilities representing each time series is computed. Finally, a hierarchical clustering algorithm operating on the matrix of distances produces the dendrogram depicting the hierarchical structure of stock markets.

3 Data

We use Datastream stock market indexes for the ten original EMU countries: Austria, Belgium, Finland, France, Germany, Ireland, Italy, the Netherlands, Portugal, and Spain. To study country-industry indexes, we follow ICB industry classification and the industries analyzed are: Oil and Gas (OILGS), Basic materials (BMATR), Industrials (INDUS), Consumer Goods (CNSMG), Health Care (HLTCH), Consumer Services (CNSMS), Telecommunications (TELCM), Utilities (UTILS), Financials (FINAN) and Technology (TECNO). The panel is imbalanced because of different starting dates and because not all countries have listed companies for all industries. Because industry indexes have different starting dates and we need a common starting date we define the period from January 3, 1990 to December 30, 2012. The analysis is conducted in USD dollars to allow comparability of the stock markets and model estimation, before and after the euro launch.

4 Conclusion

Our results are summarized in Figure 1. This dendrogram shows that dissimilarity between industries is mainly determined by industries, where telecommunications and technology play an important role.

Acknowledgements Financial support from Fundação para a Ciência e Tecnologia is greatly acknowledged (PTDC/EGE-GES/103223/2008).

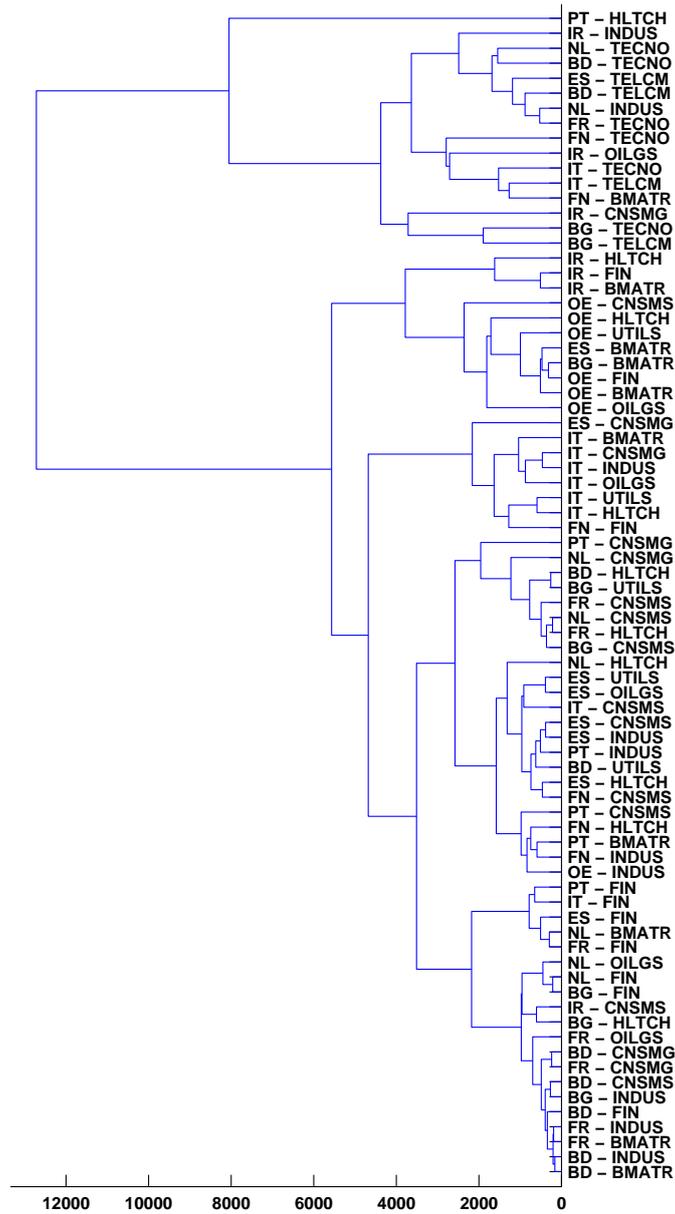


Fig. 1 European industry structure.

References

1. Ehling, P., Ramos, S.B.: Geographic versus industry diversification: Constraints matter. Journal of Empirical Finance. **13**(4-5), 396 – 416 (2006)

2. Eiling, E., Gerard, B., Hillion, P., De Roon, F.: International portfolio diversification: Currency, industry and country effects revisited. *Journal of International Money and Finance*. **31**(5), 1249–1278 (2012)
3. Ferreira, M. A., Ferreira, M. A.: The importance of industry and country effects in the EMU equity markets. *European Financial Management*. **12**(3), 341–373 (2006)
4. Hamilton, J. D.: A new approach to the economic-analysis of nonstationary time-series and the business-cycle. *Econometrica*. **57**(2), 357–384 (1989)

Consensus Community Detection: a Nonmetric MDS Approach

Drago Carlo and Balzanella Antonio

Abstract Community detection methods for the analysis of complex networks are increasingly important in modern literature. At the same time it is still an open problem. The approach proposed in this work is to adopt an ensemble procedure in order to obtain a consensus matrix, then to consider a non-metric MDS approach on the matrix and, finally, to classify the nodes on the new coordinate space. The simulation study offers some interesting insights from the procedure because it shows that it is possible to understand the key nodes and the stable communities by considering different algorithms. At the same time the procedure allows to have a first output about the identification of nodes which are in two different overlapped communities. The proposed approach is applied to real data related to a network of patents.

Key words: Complex Networks, Community Detection, Non-metric Multidimensional Scaling

1 Community Detection in Complex Networks

A network have a community structure if it can be divided in groups of sets of nodes [3]. These sets of nodes are particularly dense in terms of connections in the same group and sparse between the groups. In the case of non overlapping communities they are well defined and could be divided each other. Detecting communities in a network is particularly relevant for applicative reasons, in fact networks can usually be partitioned in community groups based on some attributes like location or occu-

Carlo Drago
Università degli Studi di Napoli "Federico II", Dipartimento di Scienze Economiche e Statistiche,
Via Cinthia 26 Napoli, 80126 e-mail: carlo.drago@unina.it

Antonio Balzanella
Seconda Università degli Studi di Napoli, Dipartimento di Scienze Politiche, Viale Ellittico 31
Caserta, 81100 e-mail: antonio.balzanella@sun.it

pation. At the same time there are important cases in which the communities can behave as independent departments in the network and they exhibit important functions [3]. So it is a very important task to detect the different communities which can occur in a network. Various algorithms and methods were proposed for accomplishing this task. In particular two different categories are considered usually (see for a review of the different classes of approaches [3, 6]). In an explorative framework where no apriori information is available on the communities in the network, the choice of the right algorithm can be unfeasible. Different methods can have different performances and biases in the community detection [4] so as, each method seems to be more appropriated in some specific network typologies and can provide partitions which can differ [5].

In order to deal with this challenge, we propose to use an ensemble of community detection algorithms and to find a consensus partition which allows to capture the most of information coming from the single community detection methods.

2 Community Detection Ensembles

We consider, as input, a network represented as an undirected graph $G = (V, E)$ where $V = (v_1, \dots, v_i, \dots, v_n)$ is the set of nodes of the network and E carries non negative values representing the presence of a connection between a pair of nodes. We consider the following set of methods in order to obtain a partitioning of the network into homogeneous communities: edge.betweenness, walktrap, fastgreedy, springlass, leading.eigenvector, multilevel, infomap, label.propagation, optimal modularity [2]. Finally we consider also blockmodeling ex-post as additional method.

We get, as output of each method, a partition $P^m = (C_1^m, \dots, C_k^m, \dots, C_K^m)$ where $m = 1, \dots, M$ is the index of the community detection method and C_k^m is the set of nodes included in the k -th community for the method m -th.

Similarly to [7] and [1], our ensemble method consists in building a consensus matrix $A = [a_{i,j}]$ (with $i, j = 1, \dots, n$) in which each cell $a_{i,j}$ (with $i \neq j$) records the number of times in which each couple of nodes is allocated to the same community of a local partition while the diagonal entries $a_{i,j} = M$ (with $i = j$), record the number of community detection methods. In this sense, a value of $a_{i,j}$ equals to the number M of methods in the ensemble indicates a full consensus in allocating the nodes (v_i, v_j) to the same community; on the contrary, the value 0 of $a_{i,j}$ indicates that no method allocates (v_i, v_j) to the same community. Finally, intermediate values of $a_{i,j}$ reveal that there is not a strong consensus in allocating the corresponding nodes to the same community as consequence of the differences among the methods in the ensemble.

3 Non Metric Multidimensional Scaling for community detection

In this section, we consider the obtained consensus matrix A as a similarity matrix. This is motivated by the following assumptions: 1) we can get a high value of $a_{i,j}$ only if a lot of community detection methods in the ensemble, consider the two corresponding nodes (v_i, v_j) so similar to be very often allocated to the same community; 2) if $a_{i,j}$ records a low value, the two corresponding nodes are considered very dissimilar by a lot of members of the ensemble; 3) intermediate values account for nodes which are considered similar by some algorithm and dissimilar by others. This involves an intermediate value of similarity.

The similarity matrix A can be transformed into a dissimilarity one D by $D = A - M$, in order to perform a non-metric multidimensional scaling algorithm. The latter provides an effective way to represent the nodes into a lower dimensional space. This is useful both, for getting a graphical representation of the network nodes and for discovering the consensus communities in the network. The latter can be obtained by running a k -means algorithm on the new coordinate system.

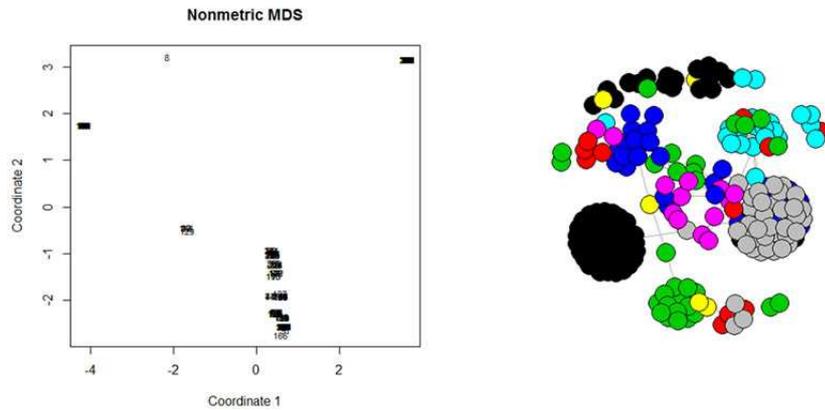
To interpret these results we consider a simulation study based on different synthetic networks.

4 Application

The application is related to a joint patent application network obtained by a new dataset of innovative firms operating in Italy. The source of data is the OECD REG-PAT database in which are collected data by edge-cut of the original network. So at the end of the procedure we obtain 216 nodes. Each node represents a single different company and each vertex represents a common patenting project of two nodes. So at this point, from the original matrix, we consider seven community detection methods: edge.betweenness, walktrap, fastgreedy, leading.eigenvector, multilevel, infomap, label.propagation. These methods detect the different communities which are collected in the consensus matrix. Then we start the non-metric MDS in order to find the distribution of the nodes in the axis X-Y identified by the procedure. Finally we use the clustering algorithm of K-Mean in order to find the stable communities. The number of chosen classes is 40. So the final interpretation of the results, is that we can detect some communities which have the structure of a node very central (representing firms very innovative) and the other nodes representing companies which participate to the common projects.

5 Conclusions

In this work we have considered an new approach in order to perform community detection. In particular this method is useful in order to detect the different tax-



onomies of nodes in a network. So we can have nodes which are particularly unstable (so they participate to different communities) and nodes which are particularly stable. This information is explored more in depth by analyzing the non metric MDS procedure which maps the different nodes in the space and provide a more simple way to interpret the original network and allows to identify the relevant patterns.

Acknowledgements The authors wish to thank Ivan Cucco for providing the data related to the joint patent application network.

References

1. Balzanella A., Verde R.: Summarizing and detecting structural drifts from multiple data streams. *Classification and Data Mining* (eds.) A. Giusti, G. Ritter, M. Vichi, XVIII, 26 p., 85. (2013).
2. Csardi G., Nepusz T.: The igraph software package for complex network research, *InterJournal, Complex Systems* **1695**. (2006).
3. Fortunato, S.: Community detection in graphs. *Physics Reports*, **486 (3)**, 75-174 (2010).
4. Leskovec, J., Lang, K. J., & Mahoney, M: Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web* (pp. 631-640) ACM. (2010).
5. Good, B. H., de Montjoye, Y. A., & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, **81(4)**, 046106 (2010).
6. Newman, M. E. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, **38 (2)**, 321-330 (2004).
7. Strehl, A. and Ghosh J. Cluster ensembles a knowledge reuse framework for combining multiple partitions, *Journal on Machine Learning Research* (2002).
8. Treviño III, S., Sun, Y., Cooper, T. F., & Bassler, K. E. (2012). Robust detection of hierarchical communities from *Escherichia coli* gene expression data. *PLoS computational biology*, **8(2)**, e1002391 (2012).

Clustering financial time series by measures of tail dependence

Fabrizio Durante, Roberta Pappadà and Nicola Torelli

Abstract We discuss two methods for clustering financial time series in extreme scenarios. The procedures are based on the calculations of two different measures of tail dependence, namely the (lower) tail dependence coefficient and the conditional Spearman's correlation. Performances of the proposed methodologies are compared via a simulation study.

Key words: Cluster analysis, Copula, Tail dependence.

1 Introduction

The clustering of a group of time series aims at finding sub-groups such that elements within a group have a similar stochastic dependence structure, while elements from distinct groups have different behaviour. In particular, such clustering techniques are particularly of interest in the case of financial time series. In fact, understanding the stochastic link between different financial time series (representing, for instance, the assets of a given portfolio) may be of interest for minimizing the total risk of a portfolio by adopting some diversification strategies.

Fabrizio Durante
School of Economics and Management, Free University of Bozen–Bolzano, Bolzano (Italy)
e-mail: fabrizio.durante@unibz.it

Roberta Pappadà
Department of Statistical Sciences, University of Padua, Padova (Italy)
e-mail: pappada@stat.unipd.it

Nicola Torelli
Department of Economic, Business, Mathematical and Statistical Sciences “Bruno De Finetti”,
University of Trieste - Trieste (Italy)
e-mail: nicola.torelli@econ.units.it

A number of approaches are available in the literature in order to create clusters of time series, which are based, for instance, on autoregressive distances [14, 16], Mahalanobis-like distances [3], variance ratio statistics [1], etc. Moreover, other procedures typically involve the choice of a convenient dissimilarity measure, derived for instance from the correlation matrix: see, for instance, [12] and [2, 13]. In fact, the main idea is that high positive correlation between time series may be interpreted in terms of some degree of similarity between them.

Following these lines of research, here we present some procedures aiming at grouping time series according to a different association measure that accounts for a kind of extreme (tail) dependence among the time series. Specifically, we aim at creating groups of time series such that elements of each group tend to comove (i.e. move together) when they are experiencing large losses.

The proposed methods are expected to provide an alternative to correlation-based clustering techniques, as also stressed by [4]. In fact, there is a growing consensus among financial institutions that classical correlation measures do not give an accurate indication risk exposures and, hence, clustering techniques that are tailored to risk management should adopt different procedures. In the following section the proposed methodology is briefly discussed. For more details see [6].

2 The methodology

Let $(x_{it})_{t=1,\dots,T}$ be a matrix of d financial time series ($i = 1, 2, \dots, d$) representing the returns of different assets and/or stock indices. In order to determine suitable groups among the considered time series according to their pairwise extreme association, we need to proceed in different steps.

First, we fit a suitable copula-based time series model to the financial time series in order to focus the attention to the link between the variables of interest without any bias by the heteroscedasticity effects [7]. Specifically, we assume that each univariate time series follows a specific ARMA-GARCH process with t -distributed innovations. The dependence among the time series is hence fully expressed by the knowledge of the copula coupling the different innovations. For more details, see [10, 15].

Once we have got the (empirical) copula expressing the dependence among the time series, we select a suitable way to measure the association between time series in the tails, i.e. we determine a criteria under which we may assign a number to the strength of the (positive) dependence between the time series in a given tail region of their joint distribution. To this end, two measures are adopted:

- The lower tail dependence coefficient that roughly corresponds to the probability that one variable exceeds a low threshold under the condition that the other variable exceeds a low threshold. For more details, see [8, 11].

- The conditional Spearman’s correlation, i.e. the Spearman’s correlation of the time series under consideration calculated conditionally on the fact that the assets are experiencing large losses (according to a given threshold). For more details, see [5].

The estimation of these quantities is obtained in a non-parametric way by using the information contained in the obtained empirical copula.

Then, we create a suitable dissimilarity matrix representing the extreme linkage among all the pairs extracted from the original vector of time series, i.e. we record the information in a way that is appropriate for the clustering procedure. Such a matrix is derived from the information contained in the previously introduced measures of association.

The dissimilarity matrix above defined could be used to determine clusters among the d time series of financial returns by means of the hierarchical agglomerative clustering techniques frequently used in practice. Specifically, among all the agglomerative strategies we may apply the three most common clustering procedures which differ in the computation of the distance between two groups: single linkage, complete linkage, average linkage (see, for instance, [9, 12]). Moreover, when suitable, the dissimilarity matrix could be further transformed in order to obtain a (Euclidean) distance matrix. Such a transformation could be helpful when one wants to adopt some specific cluster procedures (e.g. Ward method). This additional step is obtained by using, as done in [4], some multidimensional scaling procedure.

All these methodologies will be explained and compared via a simulation study. Moreover, the procedures will be applied to a specific financial dataset, showing its practical implementation.

3 Conclusions

We have presented some methods that can be applied in order to obtain clusters of financial time series that take into account the tail dependence of their joint distribution. The procedure assumes the existence of a suitable GARCH–copula model that could describe the joint behaviour of the time series. According to [4], we expect that the procedure could be applied in portfolio analysis, with particular emphasis on risk management strategies.

Acknowledgements The first author acknowledges the support of Free University of Bozen-Bolzano, School of Economics and Management, via the project “Risk and Dependence”.

References

1. Bastos, J., Caiado, J.: Clustering financial time series with variance ratio statistics. *Quantitative Finance*, in press (2013)
2. Bonanno, G., Caldarelli, G., Lillo, F., Micciché, S., Vandewalle, N., Mantegna, R.: Networks of equities in financial markets. *Eur. Phys. J. B* **38**(2), 363–371 (2004)
3. Caiado, J., Crato, N.: Identifying common dynamic features in stock returns. *Quant. Finance* **10**(7), 797–807 (2010)
4. De Luca, G., Zuccolotto, P.: A tail dependence-based dissimilarity measure for financial time series clustering. *Adv. Data Anal. Classif.* **5**(4), 323–340 (2011)
5. Durante, F., Jaworski, P.: Spatial contagion between financial markets: a copula-based approach. *Appl. Stoch. Models Bus. Ind.* **26**(5), 551–564 (2010)
6. Durante, F., Pappadà, R., Torelli, N.: Clustering of financial time series in risky scenarios. *Adv. Data Anal. Classif.*, under review (2013)
7. Forbes, K.J., Rigobon, R.: No contagion, only interdependence: measuring stock market comovements. *J. Financ.* **57**(5), 2223–2261 (2002)
8. Frahm, G., Junker, M., Schmidt, R.: Estimating the tail-dependence coefficient: Properties and pitfalls. *Insurance Math. Econom.* **37**(1), 80–100 (2005)
9. Härdle, W., Simar, L.: *Applied Multivariate Statistical Analysis*. Springer (2012). 3rd ed
10. Jaworski, P., Durante, F., Härdle, W.K. (eds.): *Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics - Proceedings*, vol. 213. Springer, Berlin Heidelberg (2013)
11. Joe, H.: *Multivariate models and dependence concepts, Monographs on Statistics and Applied Probability*, vol. 73. Chapman & Hall, London (1997)
12. Kaufman, L., Rousseeuw, P.: *Finding groups in data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons Inc., New York (1990)
13. Mantegna, R.: Hierarchical structure in financial markets. *Euro. Phys. J. B* **11**(1), 193–197 (1999)
14. Otranto, E.: Clustering heteroskedastic time series by model-based procedures. *Comput. Statist. Data Anal.* **52**(10), 4685–4698 (2008)
15. Patton, A.: Copula methods for forecasting multivariate time series. *Handbook of Economic Forecasting II*. Elsevier (2013). To appear
16. Piccolo, D.: A distance measure for classifying ARIMA models. *J. Time Ser. Anal.* **11**(2), 153–164 (1990)

Influence diagnostics for generalized linear mixed models: a gradient-like statistic

Marco Enea, Antonella Plaia

Abstract In the literature, many influence measures proposed for Generalized Linear Mixed Models (GLMMs) require the information matrix that can be difficult to calculate. In the present paper, a known influence measure is approximated to get a simpler form, for which the information matrix is no more necessary. The proposed measure is showed to have a form similar to the gradient statistic, recently introduced. Good performances have been obtained through simulation studies.

Key words: GLMM, outliers, diagnostics, gradient statistic

1 Introduction

Generalized Linear Mixed Models (GLMMs) [5] are useful extensions of both linear mixed models and generalized linear models in order to assess additional components of variability due to latent random effects. For this reason, these models have received growing attention during the past decades. Unfortunately, the model estimates may heavily depend on a small part of the dataset or even on a particular observation or cluster. Therefore, the identification of potentially influential outliers is an important step beyond estimation in GLMMs. There are two major approaches for detecting influential observations. The first one is **the local influence approach** [3] which develops diagnostic measures by using the curvature of the influence graph of an appropriate function. The second one, **the deletion approach** [1], develops a diagnostic measure by assessing a chosen quantity change that is induced by the exclusion of individual data points from an analysis. However, since the observed-data likelihood function in a GLMM involves intractable integrals, the development, as well as the evaluation, of deletion diagnostic measures on the basis of Cook's approaches is rather difficult.

In this paper, on the grounds of the measure suggested by Cook [3], we derive a diagnostic measure which does not require the information matrix, while maintaining the same large sample behaviour. In fact, as it will be shown later, the proposed measure is the analogue of the gradient statistic, recently introduced by Terrell [7] and further studied by Lemonte [4]. The performance of the proposed measure is assessed on a GLMM through well tested simulation studies.

Dipartimento di Scienze Economiche Aziendali e Statistiche, University of Palermo, Palermo, Italy; e-mail: marco.enea@unipa.it, antonella.plaia@unipa.it

2 Influence diagnostics

Let y_{ij} be the response of unit j of the i -th cluster, $j = 1, \dots, n_i$, $i = 1, \dots, N$, \mathbf{x}_{ij} and \mathbf{z}_{ij} the covariate arrays. The GLMM is written as: $g(\mu_{ij}) = g(E[y_{ij} | \mathbf{x}_{ij}, \mathbf{b}_i]) = \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i$, where $\boldsymbol{\beta}$ is a p -vector of fixed parameters and \mathbf{b}_i are assumed to be $N(\mathbf{0}, G)$. The three most used influence measures, computable for a single observation, cluster or more generally for its subset M_i , are:

the *log-likelihood distance*

$$LD_{M_i} = 2\{L(\mathbf{y}|\hat{\boldsymbol{\zeta}}) - L(\mathbf{y}|\hat{\boldsymbol{\zeta}}_{(M_i)})\}, \quad (1)$$

the *Cook's distance* [1]

$$CD_{M_i} = (\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}_{(M_i)})' \{-\ddot{L}\} (\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}_{(M_i)}), \quad (2)$$

and the *Cook's total influence measure* [3]

$$C_{M_i} = 2 |\boldsymbol{\Delta}'_{M_i} \ddot{L}^{-1} \boldsymbol{\Delta}_{M_i}|, \quad (3)$$

where \ddot{L} is the Hessian matrix of the log-likelihood relative to the parameter vector $\boldsymbol{\zeta} = (\boldsymbol{\beta}', \boldsymbol{\delta}')'$, with $\boldsymbol{\delta}$ corresponding to the variance components. Here $\boldsymbol{\Delta}'_{M_i} = \mathbf{s}_i - \mathbf{s}_{i(M_i)}$, where $\mathbf{s}_i = (\mathbf{s}'_i \boldsymbol{\beta}, \mathbf{s}'_i \boldsymbol{\delta})'$, is the subvector of the difference between the contribution to the score function of cluster i and the score function for such cluster without set M_i , for which details can be found in [6]. Of course, if interest is only in the influence of the i th cluster, it will be sufficient to consider $\boldsymbol{\Delta}'_{M_i} = \mathbf{s}_i$. Both \ddot{L} and $\boldsymbol{\Delta}'_{M_i}$ are calculated at $\boldsymbol{\zeta} = \hat{\boldsymbol{\zeta}}$. Notice that we use the ‘‘total’’, as opposed to the ‘‘local’’, influence measure in the sense that (3) is the deletion diagnostic subcase of [3], initially proposed to construct influence curves.

Now, let $\hat{\boldsymbol{\zeta}}_{(M_i)}$ be the estimate of $\boldsymbol{\zeta}$ when subset M_i is deleted. Since $\hat{\boldsymbol{\zeta}}_{(M_i)} \approx \hat{\boldsymbol{\zeta}} - [\ddot{L}_{(M_i)}(\hat{\boldsymbol{\zeta}})]^{-1} \boldsymbol{\Delta}_{(M_i)}(\hat{\boldsymbol{\zeta}})$, and by considering that $[\ddot{L}_{(M_i)}(\hat{\boldsymbol{\zeta}})]^{-1}$ can be approximated by $[\ddot{L}(\hat{\boldsymbol{\zeta}})]^{-1}$, as done by Zhu et al. [9], we have

$$\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}_{(M_i)} \approx \ddot{L}^{-1} \boldsymbol{\Delta}_{(M_i)}. \quad (4)$$

By pre-multiplying both members of (4) by $\boldsymbol{\Delta}'_{(M_i)}$, it becomes

$$\boldsymbol{\Delta}'_{(M_i)} (\hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}_{(M_i)}) \approx \boldsymbol{\Delta}'_{(M_i)} \ddot{L}^{-1} \boldsymbol{\Delta}_{(M_i)}. \quad (5)$$

Notice the similarity between the first member of (5) and the gradient statistic $\boldsymbol{\Delta}'_0(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$. Such a statistic is asymptotically χ^2 distributed, although it is not a quadratic form and might assume negative values for small sample sizes. By considering that $\sum_{i=1}^N \boldsymbol{\Delta}_{M_i} = \mathbf{0}$ and given that $\boldsymbol{\Delta}_{(M_i)} = \sum_{j \neq i} \boldsymbol{\Delta}_{M_j} = -\boldsymbol{\Delta}_{M_i}$, (5) becomes $\boldsymbol{\Delta}'_{M_i} (\hat{\boldsymbol{\zeta}}_{(M_i)} - \hat{\boldsymbol{\zeta}}) \approx \boldsymbol{\Delta}'_{M_i} \ddot{L}^{-1} \boldsymbol{\Delta}_{M_i}$. Finally, by substituting in (3) we have

$$C_{M_i} \approx C_{M_i}^a = 2 |\boldsymbol{\Delta}'_{M_i} (\hat{\boldsymbol{\zeta}}_{(M_i)} - \hat{\boldsymbol{\zeta}})|, \quad (6)$$

which is a measure of influence, because of the distance $\hat{\xi}_{(M_i)} - \hat{\xi}$, for which the use of the information matrix is no more necessary. Notice that if the M_i -th subset is influential, $\hat{\xi}_{(M_i)} - \hat{\xi}$ will be large and the accuracy of (4) will be likely to be lower. However that will no more needed as long as $C_{(M_i)}^a$ is “sufficiently large to draw our attention for further consideration.” [2, p.182].

3 Simulation studies

Following well-tested simulation schemes [8], a small-scale simulation study is performed from the following model: $y_{ij}|b_i \sim \text{Poisson}(\mu_{ij})$, $b_i \sim N(0, \sigma^2)$, $\log(\mu_{ij}) = x_{ij}\beta + b_i$, where $j = 1, \dots, n$, $i = 1, \dots, 10$, with equal sample size n in each cluster i . The single variable x_{ij} is chosen as j/n , while $\beta = 1$, $\sigma^2 = 0.1, 0.2, 1.0$ and $n = 30, 100, 200$; and both 100 and 1000 replications are considered for each combination of σ^2 and n . Both measures are calculated on models estimated adding an intercept. The aim is to investigate the performance of $C_{M_i}^a$ by assessing its concordance with C_{M_i} in terms of proportion of correct identification of: (a) the cluster with the largest C_i ; (b) the two clusters with the largest and the second largest C_i . Table 1 shows the results of the simulation. Observe that the proportions of correct identification are at least 83% (a) and 62.4% (b), which is a good result.

Table 1: Proportion of correct identification of (a) the cluster with the largest C_i and (b) the two clusters with the largest and the second largest C_i , using C_i^a .

		$n = 30, \sigma^2$			$n = 100, \sigma^2$			$n = 200, \sigma^2$		
		0.1 (%)	0.2 (%)	1.0 (%)	0.1 (%)	0.2 (%)	1.0 (%)	0.1 (%)	0.2 (%)	1.0 (%)
Replications	(a)	89.0	83.0	90.0	87.0	95.0	90.0	88.0	91.0	94.0
	(b)	64.0	67.0	73.0	66.0	74.0	71.0	73.0	75.0	80.0
1000	(a)	87.7	88.0	86.1	89.7	92.4	87.6	89.2	92.2	89.7
	(b)	62.4	68.9	68.9	67.8	73.9	72.3	64.9	75.8	75.9

A second and a third simulation study were carried out. By using the same parameters of the previous study, the second study was aimed at assessing the concordance among LR_i , CD_i , $C_i/2$ and $C_i^a/2$, on 100 “typical” datasets [8]. We generated 101 dataset, picked the one with median log-likelihood value and repeated the procedure 100 times. As a result, $C_i^a/2$ showed high concordance rates, above all with LR_i and $C_i/2$. Figure 1 shows the result from some “median” datasets by varying σ^2 and n .

In the third study, we assessed the concordance between C_i and C_i^a in a medium-large simulation study with artificially created outliers. Cluster-level and observation-level diagnostics were carried out showing very slight differences between C_i and C_i^a but, due to lack of space, these results are not reported here.

In conclusion, we have shown the gradient statistic to have an analogue, in the study of influence, which offers good performances. Although we have used the gradient-like measure in the context of the GLMMs, it can be used, as well as $LR_{(M_i)}$, $CD_{(M_i)}$, and $C_{(M_i)}$, to carry out an influence diagnostics on any model.

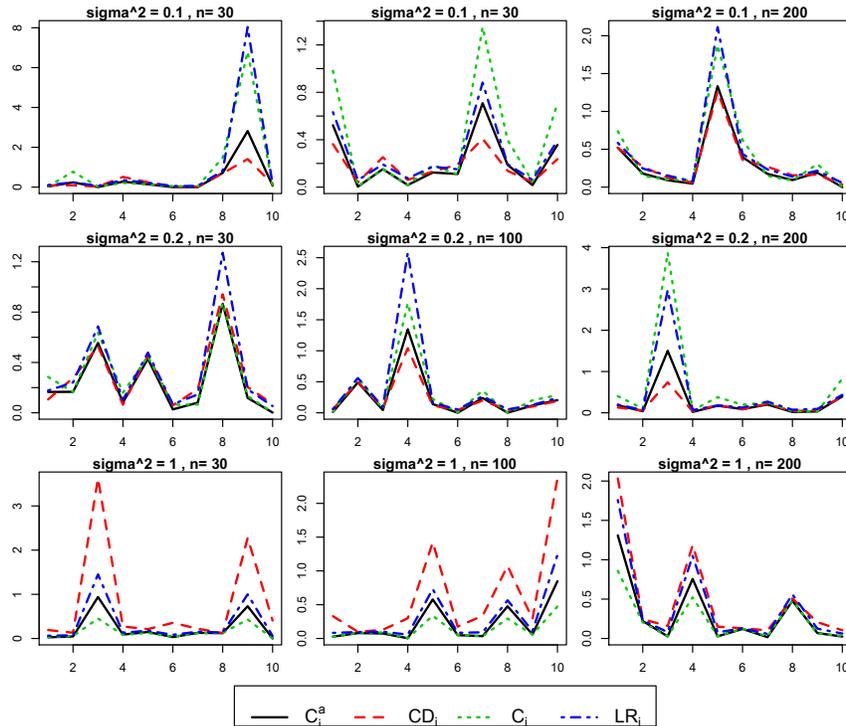


Fig. 1: Cluster-oriented diagnostics from the second simulation scheme.

References

- [1] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15–18
- [2] Cook, R. D., Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall.
- [3] Cook, R. D. (1986). Assessment of Local Influence. *J. Roy. Stat. Soc. B Met.* **4(2)**, 133–169
- [4] Lemonte, A.J. (2013). On the gradient statistic under model misspecification. *Statistics and Probability Letters*. **83**, 390–398
- [5] McCulloch, C.E., Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.
- [6] Ouwens, M. J. N. M., Tan, F. E. S., Berger, M. P. F. (2001). Local Influence to Detect Influential Data Structures for Generalized Linear Mixed Models. *Biometrics* **57(42)**, 1166–1172
- [7] Terrell, G.R.(2002). The gradient statistic. *Computing Science and Statistics* 34, **34**, 206-215
- [8] Xu, L., Lee, S., Poon, W. (2006). Deletion measures for generalized linear mixed models. *Comput. Stat. Data An.* **51**, 1131–1146
- [9] Zhu, H., Lee, S., Wei, B., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*. **88(3)**, 727–737

Joint estimation of poverty and inequality parameters in small areas

Enrico Fabrizi, Maria R. Ferrante and Carlo Trivisano

Abstract In this paper we consider the estimation of the at-risk-of-poverty rate, the relative median at-risk-of-poverty gap and the Gini index for the health districts of the administrative regions Emilia-Romagna and Tuscany, Italy. We use data from the EU-SILC survey, complemented with auxiliary information from administrative sources in the framework of small area estimation. We adopt a hierarchical Bayes approach to inference assisted by MCMC integration.

1 Introduction

In this paper we consider the estimation of the at-risk-of-poverty rate, the relative median at-risk-of-poverty gap and the Gini index for the health districts of the administrative regions Emilia-Romagna and Tuscany, Italy. The first two parameters are poverty indicators (Foster et al., 1984) while the third is an income inequality measure. They are all based on the individual distribution of the equivalent disposable income and are included in the set of common European statistical indicators on poverty and social exclusion endorsed by the Laeken council (see Marlier et al., 2007). Our main data source is given by the EU-SILC survey, wave 2010, complemented with data from administrative sources, as will be discussed later. The interest in estimating these parameters at the health district level is motivated by the role that this level of administration plays in the implementation of many social and health expenditure programmes related to the contrast of poverty and social exclusion.

¹ Enrico Fabrizi, DISES, Università Cattolica del S. Cuore; email:enrico.fabrizi@unicatt.it

Maria Ferrante, Dipartimento di Scienze Statistiche “Paolo Fortunati”, Università di Bologna; email: maria.ferrante@unibo.it

Carlo Trivisano, Dipartimento di Scienze Statistiche “Paolo Fortunati”, Università di Bologna; email: carlo.trivisano@unibo.it

The district-specific samples sizes available for estimating population descriptive parameters using ordinary design-based estimators are too small to allow adequately precise estimators in all but a few districts. In particular, direct estimates for the relative median at-risk-of-poverty gap, which is calculated only on the incomes of the poor have unacceptably large variances. This motivates the recourse to small area estimation methods. In particular we make use of area level models, in which the design-based estimates are improved with the help of models that establish a connection among the underlying area parameters and auxiliary information accurately known for each district. We consider a hierarchical Bayes approach to estimation assisted by MCMC integration. Application of this type of models to problems related to ours include Fabrizi et al. (2011), Hawala and Lahiri (2011). A small area estimation model works if the observed variation in the design-based estimators can be, at least in part, explained as function of area-level statistics calculated using auxiliary information. Our main challenge is that, in our application, this is true for the mean equivalised income (a parameter not of direct interest) and to a lesser extent for the Gini index, but not for the at-risk-of-poverty rate and the relative median at-risk-of-poverty gap. We propose a joint model in which the mean of the equivalized income is modelled assuming log-normality and other parameters are expressed as functions of the parameters characterizing this distribution.

2 The joint model

In this section we illustrate the basic feature of the joint hierarchical Bayes model we propose for the at-risk-of-poverty rate, the median poverty gap and the Gini index. For brevity, we omit to discuss the details concerning design-based estimation, the selection of auxiliary information, the variance smoothing algorithms and, in general all the modeling choices. We also omit prior specification for hyperparameters: as a general rule we specify diffuse normal priors for the regression coefficients and Gamma priors for the precision components.

2.1 *At-risk-of-poverty rate and relative median at-risk-of-poverty gap*

Let z_{ij} be the equivalent income for individual j ($j=1,\dots,n_d$) in district i ($i=1,\dots,m$). The relationship between different area parameters will be based on assuming log-normality of incomes:

$$z_{ij} \stackrel{ind}{\sim} \text{LogN}(\theta_i, \sigma_i^2) \quad (1)$$

For the area parameters θ_i we assume a linear linking model: $\theta_i = \alpha_o + \mathbf{x}'_i \beta_o + u_i$ where \mathbf{x}'_i is the $1 \times k$ vector of area-specific statistics based on

auxiliary information. v_i s are area-specific random effects for which we assume $u_i \stackrel{ind}{\sim} N(0, \sigma_u^2)$. Let's denote with p_i the design based estimate of the at-risk-of-poverty rate in district i . As $p_i \in (0, 1)$ we can assume $p_i \sim \text{Beta}(\alpha_i, \beta_i)$ with $\alpha_i = \pi_i(f_i - 1)$, $\beta_i = (1 - \pi_i)(f_i - 1)$ implying $E(p_i) = \pi_i$ and $V(p_i) = f_i^{-1} \{\pi_i(1 - \pi_i)\}$ where π_i is the underlying district level rate and f_i an effective sample size that can be calculated by smoothing the sampling variances estimated using a bootstrap algorithm. We specify a probit linking model for π_i :

Probit(π_i) = $\Phi^{-1}(\pi_i) = \mu_i + v_i$ where $v_i \stackrel{ind}{\sim} N(0, \sigma_v^2)$. To specify an equation for μ_i let's first denote with PT the poverty threshold and pt its logarithm. Under the assumption of log-normality of z_{ij} we have that the at-risk-of-poverty rate would be given by $\pi_i^{LN} = \Phi\{\sigma_i^{-1}(pt - \theta_i)\}$ so that $\Phi^{-1}(\pi_i^{LN}) = \sigma_i^{-1}(pt - \theta_i)$. We then assume that $\mu_i = \alpha_{\mu_i} + \beta_{\mu_i} \sigma_i^{-1}(pt - \theta_i)$. If log-normality exactly holds then $\alpha_{\mu_i} = 0$ and $\beta_{\mu_i} = 1$. We relax this assumption leaving the two parameters free and we assign them unconstrained priors. A point estimator of the unknown at-risk-poverty rate of district i may be obtained summarizing the posterior distribution using quadratic loss. i.e. $\pi_i^B = E(\pi_i | data)$. The parameter σ_i can be replaced by an estimate of the standard deviation based on the bootstrap algorithm. Nonetheless we will see that we can exploit the relationship between the standard deviation of the log-incomes and the Gini coefficient to model it.

Let ζ_i be the unknown area-specific relative median at-risk-of-poverty gap. We note that $\log \zeta_i = \text{Me}(y_{ij} | y_{ij} < pt)$ where $y_{ij} = \log z_{ij}$; under assumption (1) we have that $y_{ij} | y_{ij} < pt \sim TN(\theta_i, \sigma_i, l, u)$, a truncated normal distribution with $l = -\infty$, $u = pt$. Using standard formulas for this distribution we obtain $\zeta_i = \exp\left[\theta_i + \sigma_i \Phi^{-1}\{0.5\pi_i^{LN}\}\right]$. If we replace the log-normality based π_i^{LN} with π_i we obtain ζ_i as function of the parameters $\pi_i, \theta_i, \sigma_i$. The posterior mean $\zeta_i^B = E(\zeta_i | data)$ is a frequentistically biased estimator. For this reason we suggest the adoption of a constrained estimator, based on a benchmarking idea we took from Fay and Herriot (1979). Note that in deriving the small area estimator of median poverty gap we did not make use of direct estimators of the same quantity. These estimators are in fact very imprecise.

2.2 Gini index

As for the Gini index γ_i , we can observe that from (1) follows that $\gamma_i^{LN} = 2\Phi(2^{-1/2}\sigma_i) - 1$. Nonetheless, as the auxiliary variables have a good explanatory power for the direct estimates of the Gini index g_i , we prefer to specify a Beta-probit model in the line of Fabrizi et al. (2011): $g_i \sim \text{Beta}(\alpha_{g_i}, \beta_{g_i})$ where $\alpha_{g_i} = h_i^{-1}(1 - \gamma_i) - \gamma_i$, $\beta_{g_i} = h_i^{-1}(1 - \gamma_i)^2 \gamma_i^{-1} + \gamma_i - 1$ implying $E(g_i) = \gamma_i$ and $V(g_i) = \gamma_i^2 h_i$. The constant h_i can be estimated using smoothing model applied to sampling variances estimated using a bootstrap algorithm. Auxiliary information are included by means of the linking model $\text{Probit}(\gamma_i) = \Phi^{-1}(\gamma_i) = \alpha_\gamma + \mathbf{x}_i^t \beta_\gamma + w_i$ where $w_i \stackrel{ind}{\sim} N(0, \sigma_z^2)$ represents an area-specific random effect. Finally we may observe that from $\gamma_i^{LN} = 2\Phi(2^{-1/2}\sigma_i) - 1$ follows that $\sigma_i \simeq \gamma_i \sqrt{pi}$ where $pi = 3.14142\dots$. We use this relationship that expresses σ_i as a function of γ_i to improve the estimation of the at-risk-of-poverty rate π_i and the median poverty gap ζ_i that were expressed as functions of this parameter.

3 Concluding remarks

The adequacy of the proposed model has been assessed using posterior predictive checks. To evaluate the gains in efficiency achieved by small area estimators we compare coefficient of variations calculated using posterior moments and design-based estimates (variance estimated using a bootstrap algorithm). The coefficient of variations are reduced, more than 40% on average for all the three parameters. The improvements in precision are greater for districts with smaller sample sizes. The proposed estimators are design consistent; we empirically checked that the difference between model based and direct estimators go to zero as the domain specific sample size increases.

References

1. Fabrizi, E., Ferrante, M.R., Pacei, S., Trivisano, C.: Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Comp. Stat. Data An.*, 55, 1736–1747 (2011).
2. Fay, R.A., Herriot, R.E.: Estimates of income for small places: an application of James-Stein procedures to Census Data. *J. Amer. Statist. Assoc.* 74, 269–277 (1979).
3. Foster, J., Greer, J. and Thorbecke, E.: A class of decomposable poverty measures. *Econometrica*. 52, 761–766 (1984)
4. Hawala, S. and Lahiri, P.: Estimation of Poverty at the School District Level Using Hierarchical Bayes Modeling. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2832–2843 (2011).

5. Marlier, E., Atkinson, A.B., Cantillon, B., Nolan, B.: The EU and social inclusion. Facing the challenges. The Policy Press, Bristol (2007)

Macroeconomic Policy in DSGE and Agent-Based Models

Giorgio Fagiolo* and Andrea Roventini†

1 Introduction

At the dawn of 2008 a large number of contributions claimed that monetary – and, more generally, economic – policy was finally becoming more of a science (Mishkin, 2007; Galí and Gertler, 2007; Goodfriend, 2007; Taylor, 2007). Almost at the end of the Great Moderation, these authors argued that both the academic world and central banks had finally reached an overall consensus not only on the contingency rules to implement in alternative situations, but also on the fact that “the practice of monetary policy reflects the application of a core set of “scientific principles” (Mishkin, 2007, p.1). These scientific principles, in turn, derived from the so-called New Neoclassical Synthesis (Goodfriend, 2007; Woodford, 2009) grounded upon Dynamic Stochastic General Equilibrium (DSGE) models³. What is more, the available toolbox of economic policy rules was deemed to work exceptionally well not only for normative purposes, but also for descriptive ones. For example, Taylor (2007) argued that “while monetary policy rules cannot, of course, explain all of economics, they can explain a great deal” (p.1) and also that “although the theory was originally designed for normative reasons, it has turned out to have positive implications which validate it scientifically” (abstract). Given these Panglossian premises, scientific discussions on economic policy seemed therefore to be ultimately confined to either fine-tuning the “consensus” model, or assessing the ex-

* Sant’Anna School of Advanced Studies, Pisa, Italy. Mail address: Sant’Anna School of Advanced Studies, Piazza Martiri della Libertà 33, I-56127 Pisa, Italy. Tel: +39-050-883282. Fax: +39-050-883344. Email: giorgio.fagiolo@sssup.it

† University of Verona, Italy, Sant’Anna School of Advanced Studies, Pisa, Italy, OFCE, Sciences Po, Nice France. Mail address: Università di Verona, Dipartimento di Scienze Economiche, viale dell’Università 3, I-37129 Verona, Italy. Tel: +39-045-8028238. Fax: +39-045-8028529. Email: andrea.roventini@univr.it

³ For an introduction, see Clarida et al. (1999), Woodford (2003) and Galí and Gertler (2007). Cf. also Colander (2006) for an historical perspective.

tent to which elements of art (appropriable by the policy maker) still existed in the conduct of monetary policy (Mishkin, 2007)⁴.

Unfortunately, as it happened with two famous statements made, respectively, by Francis Fukuyama (1992) about an alleged “end of history”, and by many physicists in the recent debate on a purported “end of physics” (see, e.g., Lindley, 1994), these positions have been proven to be substantially wrong by subsequent events. The “perfect storm” which followed the bankruptcy of Lehman Brothers on September 15, 2008 brought financial markets on the edge of collapse causing in turn the worst recession developed economies have ever seen since the Great Depression. In 2012, the risks for the world economic system have not finished yet as the crisis is now menacing the sovereign debt of European countries and the very survival of the Euro.

What is worse, mainstream DSGE-based macroeconomics appear to be badly equipped to deal with the big turmoil we are facing. As Krugman (2011) points out, not only orthodox macroeconomists did not forecast the current crisis, but they did not even admit the possibility of such event and, even worse, they did not provide any useful advice to policy makers to put back the economy on a steady growth path (see also Stiglitz, 2011). On the same line, Delong (2011) reports that when the former U.S. secretary Lawrence Summers was recently asked what economics can offer to understand the crisis, he quoted the works of Bagehot, Minsky and Kindleberger, three dead men whose most recent book is 33 years old. This is so because the DSGE approach “has become so mesmerized with its own internal logic that it has begun to confuse the precision it has achieved about its own world with the precision that it has about the real one” (Caballero, 2010, p. 85).

In that respect, the Great Recession has revealed to be a natural experiment for economic analysis, showing the inadequacy of the predominant theoretical frameworks. Indeed, an increasing number of leading economists claim that the current “economic crisis is a crisis for economic theory” (Kirman, 2010; Colander et al., 2009; Krugman, 2009, 2011; Caballero, 2010; Stiglitz, 2011; Kay, 2011; Dosi, 2011; Delong, 2011). The basic assumptions of mainstream DSGE models, e.g. rational expectations, representative agents, perfect markets etc., prevent the understanding of basic phenomena underlying the current economic crisis.

In this paper, we argue that instead of performing Ptolemaic exercises (Stiglitz, 2011; Dosi, 2011; Caballero, 2010) trying to add additional “frictions” to fix the problems of DSGE models, economists should consider the *economy as a complex evolving system*, i.e. as an ecology populated by heterogeneous agents whose far-from-equilibrium interactions continuously change the structure of the system (more on that in Kirman, 2010; Dosi, 2011; Rosser, 2011). This is the starting point of agent-based computational economics (ACE, Tesfatsion, 2006; LeBaron and Tesfatsion, 2008). Bounded rationality, endogenous out-of-equilibrium dynamics, di-

⁴ At the opposite, according to Howitt (2011) “macroeconomic theory has fallen behind the practice of central banking” (p. 2). On the same camp, Mankiw (2006) thinks that macroeconomists should not behave as scientist but as engineers trying to solve practical problems. See also Summers (1991) for an extremely pessimistic view on the possibility of taking *any* economic model seriously econometrically. On these points see also Mehrling (2006).

rect interactions, are the tenets of ACE which allow to catch many of the features of the current crisis (e.g. asset bubbles, resilience of interbank network, self-organized criticality, financial accelerator dynamics, etc.).

On the normative side, due to the extreme flexibility of the set of assumptions regarding agent behaviors and interactions, ACE models (often called agent-based models, ABMs) represent an exceptional laboratory to perform policy exercises and policy design. Indeed, an increasing number of macroeconomic policy applications have been already devised and explored concerning fiscal and monetary policies, bank regulation and central bank independence.

Certainly, also in the ACE approach there are still open issues that should be addressed. The most important ones concern empirical validation, over-parametrization, estimation and calibration. Nevertheless, the success of ACE models in delivering policy implications while simultaneously explaining the observed micro and macro stylized facts are encouraging for the development of a new way of doing macroeconomic theory.

References

- Caballero, R. J. (2010), “Macroeconomics after the Crisis: Time to Deal with the Pretense-of-Knowledge Syndrome”, *Journal of Economic Perspectives*, 24: 85–102.
- Clarida, R., J. Galí and M. Gertler (December 1999), “The Science of Monetary Policy: A New Keynesian Perspective”, *Journal of Economic Literature*, 37: 1661–1707.
- Colander, D. (2006), “Post Walrasian Macroeconomics: Some Historic Links”, in D. Colander, (ed.), *Post Walrasian Macroeconomics*, Cambridge, Cambridge University Press.
- Colander, D., H. Folmer, A. Haas, M. D. Goldberg, K. Juselius, A. P. Kirman, T. Lux and B. Sloth (2009), “The Financial Crisis and the Systemic Failure of Academic Economics”, Technical Report, 98th Dahlem Workshop.
- Delong, J. B. (2011), “Economics in Crisis”, *The Economists’ Voice*, May.
- Dosi, G. (2011), “Economic Coordination and Dynamics: Some Elements of an Alternative “Evolutionary” Paradigm”, Technical Report, Institute for New Economic Thinking.
- Fukuyama, F. (1992), *The End of History and the Last Man*, London, Penguin.
- Galí, J. and M. Gertler (2007), “Macroeconomic Modelling for Monetary Policy Evaluation”, *Journal of Economic Perspectives*, 21: 25–46.
- Goodfriend, M. (2007), “How the World Achieved Consensus on Monetary Policy”, *Journal of Economic Perspectives*, 21: 47–68.
- Howitt, P. (2011), “What Have Central Bankers Learned from Modern Macroeconomic Theory?”, *Journal of Macroeconomics*, .
- Kay, J. (2011), “The Map is Not the Territory: An Essay on the State of Economics”, Technical Report, Institute for New Economic Thinking.

- Kirman, A. P. (2010), “The Economic Crisis is a Crisis for Economic Theory”, *CEsifo Economic Studies*, 56: 498–535.
- Krugman, P. (2009), “How did Economics Get it So Wrong?”, *New York Times Magazine*, : 36–44.
- Krugman, P. (2011), “The Profession and the Crisis”, *Eastern Economic Journal*, 37: 307–312.
- LeBaron, B. and L. Tesfatsion (2008), “Modeling Macroeconomies as Open-Ended Dynamic Systems of Interacting Agents”, *American Economic Review*, 98: 246–250.
- Lindley, D. (1994), *The End of Physics*, Basic Books.
- Mankiw, G. N. (2006), “The Macroeconomist as Scientist and Engineer”, *Journal of Economic Perspectives*, 20: 29–46.
- Mehrling, P. (2006), “The Problem of Time in the DSGE Model and the Post Walrasian Alternative”, in D. Colander, (ed.), *Post Walrasian Macroeconomics*, Cambridge, Cambridge University Press.
- Mishkin, F. S. (2007), “Will Monetary Policy Become More of a Science”, Working Paper 13566, NBER.
- Rosser, B. J. (2011), *Complex Evolutionary Dynamics in Urban-Regional and Ecologic-Economic Systems: From Catastrophe to Chaos and Beyond*, Springer: New York.
- Stiglitz, J. (2011), “Rethinking Macroeconomics: What Failed, and How to Repair It”, *Journal of the European Economic Association*, 9: 591–645.
- Summers, L. (1991), “The Scientific Illusion in Empirical Macroeconomics”, *Scandinavian Journal of Economics*, 93: 129–148.
- Taylor, J. (2007), “The Explanatory Power of Monetary Policy Rules”, Working Paper 13685, NBER.
- Tesfatsion, L. (2006), “ACE: A Constructive Approach to Economic Theory”, in L. Tesfatsion and K. Judd, (eds.), *Handbook of Computational Economics II: Agent-Based Computational Economics*, Amsterdam, North Holland.
- Woodford, M. (2003), *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton, NJ, Princeton University Press.
- Woodford, M. (2009), “Convergence in Macroeconomics: Elements of the New Synthesis”, *American Economic Journal: Macroeconomics*, 1: 267–279.

New Flexible Probability Distributions for Ranking Data

Salvatore Fasola and Mariangela Sciandra

Abstract Recently, several models have been proposed in literature for analyzing ranks assigned by people to some object. These models summarize the liking feeling for this object, possibly also with respect to a set of explanatory variables. Some recent works have suggested the use of the *Shifted Binomial* and of the *Inverse Hypergeometric* distribution for modelling the *approval rate*, while *mixture models* have been developed for taking into account the uncertainty of the ranking process. We propose two new probabilistic models, based on the *Discrete Beta* and the *Shifted-Beta Binomial* distributions, that ensure much flexibility and allow the joint modelling of the scale (approval rate) and the shape (uncertainty) of the distribution of the ranks assigned to the object.

Key words: Ranking data, Discrete Beta, Shifted-Beta Binomial

1 Introduction

Ranking data are frequently collected when individuals are presented with a set of alternatives, that in this work we will refer to as *items*, and are asked to order them from most to least preferred. Because of their applicability in many fields, several methods for ranking data have been proposed. Some recent works have suggested the use of the Shifted Binomial (D'Elia, 2000) and of the Inverse Hypergeometric distribution (D'Elia, 2003) for modelling the approval rate of objects, while mixtures of Discrete Uniform and Shifted Binomial random variables (MUB models)

Salvatore Fasola
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo,
e-mail: salvatore.fasola@unipa.it

Mariangela Sciandra
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo,
e-mail: mariangela.sciandra@unipa.it

have been proposed to deal with both the selection mechanism and the uncertainty in the ranking process (D'Elia and Piccolo, 2005).

2 Two new flexible distributions

When an individual ranks K items, the resulting observation is one of the possible permutations of the first K integers, under the assumption that ties cannot occur. Data arising from such situations are generally arranged in an $n \times K$ matrix of ranks $R = \{r_i^k\}$, where the generic entry r_i^k represents the rank assigned by the i -th individual to the k -th item. Supposing to be interested in summarizing the liking feeling towards item k , the response variable becomes univariate, and is denoted with R^k . Two new flexible probability distributions for R^k are now proposed, preserving both simplicity and interpretability of their parameters.

2.1 The Discrete Beta distribution

The first model proposed is intended to exploit the flexibility in shape of the Beta distribution to improve fitting performances. Clearly, as the support of the Beta distribution is continuous, it is necessary to consider a suitable transformation able to meet the discrete nature of ranks. Let X be a Beta random variable with parameters a and b , and indicate its *p.d.f* with $f_X(x; a, b)$. If K is the number of items, the idea is to divide the support of the variable X into K intervals of the same width and consider their respective (integrated) probabilities. Then, define a vector of $K - 1$ equally spaced threshold values $x_j = j/K$, $j = 1, 2, \dots, K - 1$. If we consider the discrete set of probabilities

$$P_j = F_X(x_j, a, b) - F_X(x_{j-1}, a, b) \quad j = 1, 2, \dots, K, \quad (1)$$

where $x_0 = 0$, $x_K = 1$ and $F_X(x, a, b)$ is the distribution function of X , it can be associated to ranks assuming

$$\text{Prob}(R_i^k = r_i^k) = P_{r_i^k}.$$

It is possible to show that the resulting discrete distribution reflects the shape of the original one, particularly when K is large.

2.2 The Shifted-Beta Binomial distribution

Let now R_i^k follow a *Shifted Binomial* (D'Elia, 2000) distribution

$$\text{Prob}(R_i^k = r_i^k) = \binom{K-1}{r_i^k-1} \psi_k^{r_i^k-1} (1-\psi_k)^{K-r_i^k}, \quad (2)$$

where ψ_k is the *disliking indicator*. A possible alternative is to use a rather natural generalization of equation (2) in which ψ_k is assumed to be the realization of a Beta random variable $\Psi_k \sim B(a, b)$. From this assumption, a *Shifted-Beta Binomial* model follows as

$$\text{Prob}(R_i^k = r_i^k) = \binom{K-1}{r_i^k-1} \frac{B(a+r_i^k-1, b+K-r_i^k)}{B(a, b)}. \quad (3)$$

Note that, when $a = 1$, it reduces to the *Inverse Hypergeometric* model (D'Elia, 2003).

3 Model estimation

For the two models, estimates of a and b can be obtained via numerical maximization of the likelihood function:

$$L(a, b; \mathbf{r}^k) = \prod_{i=1}^n \prod_{j=1}^K \text{Prob}(R_i^k = r_i^k)^{I(r_i^k=j)}. \quad (4)$$

As the expressions of the mean and the variance for the two proposed distributions are quite complex and difficult to treat mathematically, we propose to use some known results from the underlying continuous distributions. More precisely, a proper mathematical manipulation of the parameters of the Beta distribution allows to define

$$E(X) = E(\Psi) = \frac{a}{a+b}$$

as a *disliking indicator*, and

$$\frac{E(X)[1-E(X)]}{V(X)} - 1 = \frac{E(\Psi)[1-E(\Psi)]}{V(\Psi)} - 1 = a+b$$

as an *accuracy indicator*. An attractive reparameterization assumes

$$\eta = \text{logit} \left(\frac{a}{a+b} \right), \quad \gamma = \log(a+b); \quad (5)$$

it allows the joint modelling of the scale (*disliking indicator*) η and the shape (*accuracy indicator*) γ of the rank distribution. For both parameters, a set of explanatory variables can be considered. For example, if we assume that two (possibly equal) vectors \mathbf{x}_i and \mathbf{z}_i influence, respectively, η and γ linearly, than:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}, \quad \gamma_i = \mathbf{z}_i^T \boldsymbol{\alpha}. \quad (6)$$

4 Application

The two models introduced have been applied to the ranks assigned by $n = 5738$ members of the American Psychological Association (APA) to the 5 candidates during the election of the president in 1980. The complete dataset is reported in Diaconis (1989); our attention is focused on preferences expressed towards the third candidate, due to the particular shape assumed by the corresponding observed rank distribution (Fig. 1). The discrete beta model gives $\hat{a} = 0.64$ and $\hat{b} = 0.70$ (AIC=18146.95), the shifted beta binomial model gives $\hat{a} = 0.55$ and $\hat{b} = 0.61$ (AIC=18143.55), while the MUB model (D'Elia and Piccolo, 2005) gives $\hat{\xi} = 0.99$ and $\hat{\pi} = 0.10$ (AIC=18261.22). The goodness of fit is much better for the proposed models, thanks to the shape flexibility of the distributions derived from the *Beta*; more specifically, the MUB model performs very well for concave, monotonic and uniform distributions, but can not fit convex distributions.

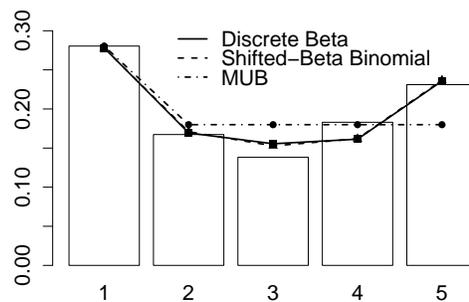


Fig. 1 Observed and fitted distributions for the ranks assigned to the third candidate for the election of the president of APA in 1980.

Acknowledgements We thank doctor Federico Torretta for his useful comments and suggestions.

References

1. D'Elia, A.: A shifted binomial model for rankings. In: Nunez-Antn, V., Ferreira, E. (Eds.), Proceedings of the 15th International Workshop on Statistical Modelling, Servicio Editorial de la Universidad del Pais Vasco, Bilbao. pp. 412–416 (2000).
2. D'Elia, A.: Modelling ranks using the inverse hypergeometric distribution. *Statistical modelling*, **3**, 65–78 (2003).
3. D'Elia, A., Piccolo, D.: A mixture model for preference data analysis. *Computational Statistics and Data Analysis*, **49**, 917–934 (2005).
4. Diaconis, P.: A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, **17**, 949–979 (1989).

A new fuzzy clustering algorithm with entropy regularization

Maria Brigida Ferraro and Paolo Giordani

Abstract The Fuzzy k -Means (FkM) algorithm is a tool for clustering n objects into k homogeneous groups. FkM is able to detect only spherical shaped clusters, hence it may not work properly when clusters have different shapes. For this purpose a variant of FkM is the Gustafson-Kessel (GK) algorithm, which can recognize the shapes of the clusters by the computation of the covariance matrix for each cluster. The fuzziness of the FkM and GK partitions is tuned by the so-called parameter of fuzziness which is an artificial device lacking a physical meaning. In order to avoid this inconvenience a fuzzy clustering algorithm with entropy regularization can be used. The idea consists in tuning the amount of fuzziness of the obtained partition by the concept of entropy. Unfortunately, such a clustering algorithm can identify only spherical clusters. In this respect, we introduce a GK-like algorithm with entropy regularization capable to discover non-spherical clusters.

Key words: Fuzzy clustering, Entropy regularization, Gustafson-Kessel algorithm

1 Introduction

Clustering consists in shaping classes from a set of objects, based on knowing some of their properties represented by a set of features (variables), X_1, \dots, X_r , observed on n objects. In this context there are different sources of uncertainty [2]: (i) sampling variability, (ii) number and shape of clusters, (iii) assignment of objects to

Maria Brigida Ferraro
Department of Statistical Sciences, Sapienza University of Rome
High Performance Computing and Networking Institute, National Research Council, Napoli
e-mail: mariabrigida.ferraro@na.icar.cnr.it

Paolo Giordani
Department of Statistical Sciences, Sapienza University of Rome
e-mail: paolo.giordani@uniroma1.it

clusters, (iv) imprecision/vagueness of observed features. The most known procedure used to cope with source (iii) is the Fuzzy k-Means (FkM) [1] algorithm. FkM can be written as

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{H}} J_{FkM} &= \sum_{i=1}^n \sum_{g=1}^k u_{ig}^m d^2(\mathbf{x}_i, \mathbf{h}_g), \\ \text{s.t. } u_{ig} &\in [0, 1], \quad \sum_{g=1}^k u_{ig} = 1, \quad i = 1, \dots, n, \quad g = 1, \dots, k, \end{aligned} \quad (1)$$

where u_{ig} is the membership degree of object i to cluster g and $\mathbf{h}_g = [h_{g1}, \dots, h_{gr}]$, the g -th row of the prototype matrix \mathbf{H} of order $(k \times r)$, is the prototype of cluster g . The membership degrees u_{ig} , collected in the matrix \mathbf{U} of order $(n \times k)$, indicate the extent to which an object belongs to a cluster. Finally, $d^2(\mathbf{x}_i, \mathbf{h}_g)$ is the dissimilarity measure (usually the squared Euclidean distance) between object i and prototype g and $m > 1$ is a coefficient tuning the amount of fuzziness in the obtained partition. Unfortunately, FkM algorithm cannot detect non-spherical shape clusters. In order to overcome this drawback, a variant of FkM, known as the Gustafson-Kessel (GK) algorithm, has been introduced [3]. The GK procedure recognizes clusters of different shapes by computing the covariance matrix for each cluster. The GK algorithm can be expressed as in (1) replacing $d^2(\mathbf{x}_i, \mathbf{h}_g)$ with the Mahalanobis distance $d_M^2(\mathbf{x}_i, \mathbf{h}_g) = (\mathbf{x}_i - \mathbf{h}_g)' F_g (\mathbf{x}_i - \mathbf{h}_g)$, being F_g a symmetric and positive definite based on the covariance matrix for the g -th cluster, namely $S_g = \frac{\sum_{i=1}^n u_{ig}^m (\mathbf{x}_i - \mathbf{h}_g)(\mathbf{x}_i - \mathbf{h}_g)'}{\sum_{i=1}^n u_{ig}^m}$.

2 GK-like algorithm with entropy regularization

The main limitation of the FkM and GK algorithms involves the fuzziness parameter m . In fact, it is an artificial device lacking a physical meaning. For this purpose, Li and Mukaidono [4, 5] propose a different fuzzy clustering algorithm obtained by adding an entropy regularization term to the objective function. Such a regularization term plays the role of measuring the overall fuzziness of the partition. The clustering process produces a fuzzy partition of the objects ensuring the maximum of compactness of the obtained clusters. The entropy-based fuzzy clustering algorithm can be formulated as

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{H}} J_{ent} &= \sum_{i=1}^n \sum_{g=1}^k u_{ig} d^2(\mathbf{x}_i, \mathbf{h}_g) + p \sum_{i=1}^n \sum_{g=1}^k u_{ig} \log u_{ig}, \\ \text{s.t. } u_{ig} &\in [0, 1], \quad \sum_{g=1}^k u_{ig} = 1, \quad \sum_{g=1}^k u_{ig} = 1, \quad i = 1, \dots, n, \quad g = 1, \dots, k, \end{aligned}$$

where p is the degree of fuzzy entropy. The parameter p is also called the ‘‘temperature’’ in statistical physics.

Such an entropy-based variant of FkM is able to discover only clusters of spherical shape. Therefore, as for FkM, it may fail when clusters of different shapes exist.

For this reason, we propose a GK-like algorithm involving an entropy regularization term. Differently from the existing techniques, the new clustering procedure can detect clusters of different shapes and does not involve the use of the artificial parameter m . The new clustering algorithm can be written as:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{H}, F_1 \dots F_k} J_{GK-ent} &= \sum_{i=1}^n \sum_{g=1}^k u_{ig} d^2(\mathbf{x}_i, \mathbf{h}_g) + p \sum_{i=1}^n \sum_{g=1}^k u_{ig} \log u_{ig}, \\ \text{s.t. } u_{ig} &\in [0, 1], \quad \sum_{g=1}^k u_{ig} = 1, \quad \sum_{i=1}^n u_{ig} = 1, \quad i = 1, \dots, n, \quad g = 1, \dots, k. \end{aligned} \quad (2)$$

We refer to the problem in (2) as the GK-like algorithm with entropy regularization. By using the Lagrange multipliers the iterative solution can be found as

$$u_{ig} = \frac{\exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{h}_g)}{p}\right)}{\sum_{g'=1}^k \exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{h}_{g'})}{p}\right)}, \quad \mathbf{h}_g = \frac{\sum_{i=1}^n u_{ig} \mathbf{x}_i}{\sum_{i=1}^n u_{ig}}, \quad F_g^{-1} = \frac{S_g}{(\det(S_g))^{1/r}}.$$

An interesting property of the proposed algorithm is that the prototypes are computed as weighted means using the u_{ig} 's as system of weights, rather than the u_{ig}^m 's as in FkM and GK.

3 Real case-study

The aim of this section is to cluster the EU27 countries on the basis of the unemployment rates in 2011. In particular, the features analyzed in this study are (source Eurostat): the total unemployment rate, the unemployment rate for young people (aged 15-24) and the long-term unemployment share (percentage of unemployment persons who have been unemployed for 12 months or more). The correlation matrix of these variables is

$$\begin{bmatrix} 1 & 0.92 & 0.58 \\ 0.92 & 1 & 0.54 \\ 0.58 & 0.54 & 1 \end{bmatrix}.$$

Thus, we can see that the three variables are highly correlated. It is therefore reasonable that such a correlation structure is also recovered in the clustering process. Nonetheless, this is not the case for the standard FkM. In fact, FkM recognizes three spherical clusters characterized, respectively, by low, medium and high values of the unemployment rates and long-term unemployment shares. On the contrary, in order to detect the existing correlation structure characterizing the clusters, the GK-like algorithm with entropy regularization allows us to recognize clusters having different shapes. We get the following prototype matrix:

$$\mathbf{H} = \begin{bmatrix} 9.40 & 23.19 & 36.74 \\ 5.91 & 14.72 & 39.07 \\ 15.48 & 33.21 & 51.93 \end{bmatrix}$$

Therefore, Cluster 1 is characterized by medium total and young unemployment rates but low long-term unemployment shares, denoting a dynamic labour market. Cluster 2 represents a static labour market, in details, it is composed by countries with low total and young unemployment rates but medium long-term unemployment shares. Finally, high total and young unemployment rates and high long-term unemployment shares are the characteristics of Cluster 3 highlighting a critical situation. In the hard clustering sense (i.e., membership degrees higher than 0.50), the obtained clusters are: {Bulgaria, Croatia, Cyprus, Denmark, Finland, France, Hungary, Iceland, Poland, Slovakia, Sweden, Turkey, UK} (dynamic labour market), {Austria, Belgium, Czech Republic, Germany, Italy, Luxembourg, Malta, Netherlands, Norway, Romania, Slovenia, Switzerland} (static labour market) and {Estonia, Ireland, Greece, Latvia, Lithuania, Portugal, Spain} (critical situation). For the sake of completeness, note that the standard GK-like algorithm (without entropy regularization) produces essentially the same partition in the hard clustering sense, but, in general, the membership degrees are fuzzier.

4 Concluding remarks

In this work we proposed a GK-like algorithm with an entropy regularization term able to recognize non-spherical shaped clusters. This helped us to avoid the artificial parameter m used in the standard FkM algorithm tuning the fuzziness of the partition. We checked the adequacy of our proposal by means of a real-case study.

References

1. Bezdek J.C.: Pattern recognition with fuzzy objective function algorithm, Plenum Press, New York (1981)
2. Coppi, R.: Management of uncertainty in statistical reasoning: the case of regression analysis. *Int. J. Approx. Reason.* **47**, 284305 (2008)
3. Gustafson E., Kessel W.: Fuzzy clustering with a fuzzy covariance matrix. In: Proc. of IEEE CDC (1979)
4. Li R.P., Mukaidono M.: A Maximum-Entropy Approach to Fuzzy Clustering. In: Proceedings of the Fourth IEEE Conference on Fuzzy Systems (FUZZ-IEEE/IFES 95), 2227–2232 (1995)
5. Li R.P., Mukaidono M.: Gaussian clustering method based on maximum-fuzzy-entropy interpretation. *Fuzzy Sets and Systems* **102**, 253–258 (1999)

Mobility measures for the dairy farms in Lombardy

Ferretti C., Ganugi P., Pieri R.

Abstract We provide an analysis of mobility of dairy farms in Lombardy. Using the AGEA dataset about the two Provinces of Cremona and Lodi, we subdivide the farms into classes, according to the milk production (tons), and we construct the corresponding six-years transition matrices. The degree of mobility is measured by means of two indices, which are evaluated as functions of the same matrices. We apply here a new index able to grasp the prevailing direction in the evolution of farms, towards up- or downsizing. Results show that the turbulence in Lodi is higher than Cremona, whereas Cremona has an higher tendency towards the upsizing. The problem of entry-exit of farms is also introduced.

Keywords Mobility indices - Transition Matrices - Firm size analysis

1 Introduction

Milk production is a very relevant sector of Lombardy economy, given its weight in terms of Employment, Added Value and its key role in the supply chain of the regional dairy production. Its particular Industrial Organization can be grasped by three sharp features: 1) the skewness of firm size distribution; 2) a massive entry and exit process; 3) the continuous and relevant upsizing of the dairy farms. In this paper we want to focus on the process of upsizing using two mobility indices. The first is the *trace index* ([3]), while the second is the new *directional index* proposed in [1], which is used in one particular version. We finally introduce the problem to extend the same directional index to the measurement of entry and exit.

¹

Camilla Ferretti, Univ. Cattolica del Sacro Cuore (PC). camilla.ferretti@unicatt.it

Piero Ganugi, Univ. degli Studi di Parma (PR). piero.ganugi@unipr.it

Renato Pieri, Univ. Cattolica del Sacro Cuore (PC). renato.pieri@unicatt.it

1.1 Data and main features of Milk Production

We have at disposal the yearly amount of produced milk (tons), from 1995 to 2011. Data regard 1710 dairy farms in the Province of Cremona (CR), and 617 farms in the Province of Lodi (LO). In Table 1 we produce the main statistics for the distribution of farms in 1995, 2000, 2005 and 2010.

Table 1: Summary statistics for the milk production (tons) in 1995, 2000, 2005 and 2010.

Year	CR				LO			
	1995	2000	2005	2010	1995	2000	2005	2010
I ^o quartile	246	310.2	459.3	575.8	365.5	445.3	542.1	612.7
Median	502	648.5	871.8	1056.9	699.7	838.7	976.2	1039.1
III ^o quartile	923.4	1103.9	1428.1	1662.8	1107.2	1279.6	1448.6	1586.4
Mean	656.6	824.9	1080.4	1301.8	849.9	1001.0	1160.7	1287.0
SE	574.6	728.2	884.3	1080.5	706.2	798.8	969.0	1089.6
Skewness	2.3	2.2	1.8	1.9	2.6	2.0	2.6	2.6

From Table 1 it is evident the upsizing of dairy farms in the two Provinces: as an example the mean production in 2010 is about 200% of the production in 1995 in Cremona, and 150% in Lodi. The upsizing is a particular feature which is unknown in Italian Manufacturing Industry. We estimate also the skewness of the distribution. The positive values reveal the presence of a right tail composed by the biggest farms.

Another important topic regards the not-negligible presence of the entry-exit phenomena. Indeed, Table 2 shows the number of newborn and expired farms in 1995, 2000, 2005 and 2010¹.

Table 2: Newborn and expired farms in 1995, 2000, 2005 and 2010.

Year	CR				LO			
	1995	2000	2005	2010	1995	2000	2005	2010
Newborn	44	15	48	14	15	5	24	7
Expired	46	58	83	39	18	12	22	16
Total	1412	1148	961	851	503	424	376	331

As we can see the mortality rate is higher than natality, causing a decrease in the number of active farms. Using the whole data from 1995 to 2011 we can calculate the distribution of the farms lifetime, shown in Table 3:

Table 3: Lifetime distribution of farms (percentages).

Lifetime (years)	1	2	3	4	5	6	7	8	9
CR	5.1	8.7	4.5	4.0	6.4	4.3	3.3	4.4	4.2
LO	5.4	8.4	4.7	3.6	3.7	3.9	2.8	2.1	4.9
Lifetime (years)	10	11	12	13	14	15	16	17	
CR	2.6	3.4	4.8	2.2	2.1	2.2	2.6	35.7	
LO	3.4	2.9	4.2	3.4	2.3	2.6	3.1	38.7	

¹ "Newborn" stands for farms producing milk in the (t+1)-th but not in the t-th year, whereas "expired" stands for farms producing milk in the t-th but not in the (t+1)-th year.

2 The analysis of Mobility

The tool we use to measure the mobility is a new index which has been recently proposed in [1]. Such index is able to grasp the prevailing direction towards a growth or a reduction in the milk production, and it is defined as a function of the transition matrix, as in the following formula:

$$I_{\omega, \nu}(P) = \frac{1}{Z} \sum_{i=1}^k \omega_i \sum_{j=1}^k p_{ij} \text{sign}(j-i) \nu(|j-i|)$$

where $P = \{p_{ij}\}_{i,j=1,\dots,k}$ is the transition matrix, $\omega = (\omega_1, \dots, \omega_k)$ is a vector of weights to be attributed to the states, $\text{sign}(x)$ is the sign function, equal to -1 if $x < 0$, +1 if $x > 0$ and 0 if $x = 0$, and ν is a function to measure the magnitude of the jumps from the i -th to the j -th state. Z is a normalizing constant.

The directional is compared to the trace index proposed in [3]:

$$I_{tr}(P) = \left(k - \sum_{i=1}^k p_{ii} \right) / (k-1)$$

To construct the transition matrices, we subdivide the farms in classes, according to the milk production (see [2]), as in Table 4:

Table 4: Classes based on the milk production (tons).

Class	I	II	III	IV	V	VI	VII	VIII	IX
Tons	0-10	10-20	20-50	50-100	100-200	200-500	500-1000	1000-2000	>2000

From Table 3 we find out that the median lifetime of farms is equal to 11 years in Cremona and 13 years in Lodi. In this light we consider transition matrices covering six years, that is half of the median lifetime. Transition matrices are built on consecutive and overlapped spans of time: 1995-2001, 1996-2002, 1997-2003, ..., etc.

For every period we consider a closed panel, formed by that farms which are active both in the initial and in the final year.

Table 5 shows the measures of mobility for every span of time. The directional index is evaluated using the distribution of farms in the initial year as weight ω , and setting $\nu(|j-i|) = |j-i|$

Table 5: Mobility indices (percentages) evaluated on the 6-years transition matrices.

Span of time	CR			LO		
	Trace index	Directional index	Nr. of farms	Trace index	Directional index	Nr. of farms
1995-2001	26.7	6.3	1013	51.9	6.0	389
1996-2002	27.5	8.2	995	56.2	9.7	395
1997-2003	28.9	8.7	963	54.0	5.3	393
1998-2004	36.4	9.8	906	58.3	6.7	369
1999-2005	28.9	11.2	875	58.3	5.8	353

Table 5: continued on next page

Table 5: continued from previous page

Span of time	CR			LO		
	Trace index	Directional index	Nr. of farms	Trace index	Directional index	Nr. of farms
2000-2006	29.5	13.1	816	59.6	0.4	330
2001-2007	26.8	12.5	814	60.2	-0.2	319
2002-2008	28.0	6.8	770	58.9	0.4	295
2003-2009	25.8	10.9	752	58.5	5.9	285
2004-2010	26.0	6.5	741	59.9	3.7	284
2005-2011	22.9	5.0	728	61.7	-1.1	278

The trace index measures the turbulence in the dynamics, revealing that the mobility is higher in Lodi (around two times than the mobility in Cremona). The positive sign of our directional index instead confirms the tendency to the upsizing of dairy farms, and its absolute value reveals that such tendency is higher in Cremona than in Lodi. In addition we see that the same tendency reaches a peak in 2000-2006 in Cremona, and a minimum in the same period in Lodi, and that it tends to become lower in the last years. In Lodi the last period is even denoted by a slight downsizing.

3 Open Problems

In applying the directional index we note mainly two open problems. Firstly we aim to provide a measure able to consider the presence of not-equally spaced classes. Indeed, with $v(|j-i|) = |j-i|$, we are assuming that a jump from the first to the second class requires the same effort of a jump from the second-last to the last class, which is not true in the case of exponentially growing classes. The choice $v(|j-i|) = \exp(|j-i|) - 1$ should be a solution, but it is too susceptible to little fluctuation in the transition probabilities $\{p_{ij}\}_{i,j=1,\dots,k}$.

Secondly we note that the entry-exit dynamics represents an important feature in the analysis of economic phenomena (see [2]). Then we aim to extend the mobility analysis, for example by comparing the mobility of incumbents and newborns. Further research will be focused on this topic.

References

1. Ferretti, C., Ganugi, P.: A new Mobility Index for Transition Matrices, Stat. Methods Appl. (2013) DOI 10.1007/s10260-013-0232-9
2. Lanciotti, C., Mambriani, D.: I cambiamenti in atto e le proiezioni al 2004 e 2008 della struttura delle aziende da latte in Italia. In: La struttura della zootecnia da latte in Italia e in Europa, Pieri, R. and Rama, D. eds. (2002)
3. Shorrocks, A.F.: The measurement of Mobility, Econometrica, 46, 1013-1024 (1978)

Model averaging and ensemble methods for risk corporate estimation

Silvia Figini, Marika Vezzoli

Abstract When many competing models are available for estimation, model averaging represents an alternative to model selection. Despite model averaging approaches have been present in statistics for many years, only recently they are starting to receive attention in applications especially in credit risk modelling ([3] and [4]). In this paper we compare Bayesian (see e.g. [8] and the references therein) and classical model averaging approaches, like Random Forest [1], with the aim of pointing out that aggregated models work better than single ones. Using a real data set, we estimate the corporate risk of a set of Small and Medium Enterprises (SME) in Germany comparing the results obtained with multiple and single models.

Key words: Bayesian model averaging, Random Forests, corporate risk

1 Introduction

Many papers have argued that combining predictions from alternative models often improves upon forecasts based on a single best model. In an environment where individual models are subject to structural breaks and misspecified by varying degrees, a strategy that pools information from the many models typically performs better than methods that try to select the best forecasting model.

To use this strategy, the forecaster faces two basic choices: which models to include in the pool of models, and how to combine the model predictions.

A vast body of research has investigated optimal model combination, yet have repeatedly found that a simple average of the forecasts produced by individual predictors is a difficult benchmark to beat, and commonly outperforms more sophisticated

Silvia Figini,
University of Pavia, e-mail: silvia.figini@unipv.it

Marika Vezzoli,
University of Brescia, e-mail: marika.vezzoli@med.unibs.it

weighting schemes that rely on the estimation of theoretically optimal weights. This is the forecast combination puzzle.

While there is a large literature examining model combination weights, Capistran et al. [2] point out that little research has focused on how to choose the models to combine. A correlated research issue concerns what to combine, since we may focus on weak learners (e.g. trees) in the spirit of ensemble methods, or dealing with aggregated models also mixing parametric together with non parametric models.

In order to provide a response to these issues, in this preliminary study we compare Bayesian model averaging and ensemble methods with respect to their correspondent single models (Logistic Regression and Tree, respectively). For identifying the best predictors (that we will combine in future researches) we look at common assessment measures employed in real application (i.e. the predictive capability and the discriminatory power). This paper is organized as follows: Section 2 reports a brief description of the multiple models employed in this study and the measures of performance used for comparing the predictions obtained from each model. Section 3 reports the empirical evidence at hand using a real data set provided by a credit rating agency and some conclusions.

2 Methodological approach

In this section we report a description of the averaging techniques employed in this paper: Bayesian Model Averaging (BMA) for parametric logistic regression models and Random Forest (RF) for non parametric models based on trees. In the parametric framework, let us suppose that a researcher has q possible models in mind, $h = 1, \dots, q$. This implies that there are q different estimates of the effect of interest depending on the model considered, say $(\hat{\beta}_1, \dots, \hat{\beta}_q)$. The key idea of model averaging is to consider and estimate all the q candidate models and then report a weighted average as the estimate of the effect of interest. The model averaging estimate $(\hat{\beta}_{MA} = \sum_{h=1}^q \omega_h \hat{\beta}_h)$, where ω_h is the weight associated to model h .

Given q variables we obtain 2^q models, M_1, \dots, M_{2^q} . The posterior for the parameters calculated using M_j (with $j = 1, \dots, 2^q$) is written as:

$$g(\beta^j|y, M_j) = \frac{f(y|\beta^j, M_j)g(\beta^j|M_j)}{f(y|M_j)}. \quad (1)$$

For each model we have at hand a likelihood $f(y|\beta^j, M_j)$ and a prior $g(\beta^j|M_j)$ distribution. The model weights in BMA is represented by the posterior model probability $p(M_j|y)$ computed as $p(M_j|y) = \frac{f(y|M_j)p(M_j)}{f(y)}$.

Given the Bayesian framework based on parameter distributions, when applying BMA both estimation and inference come naturally together from the posterior distribution that provides inference about β that takes into account of model uncertainty. In the non parametric framework, the ensemble learning techniques (e.g. [1], [5]) combine poor predictors, like trees, in order to obtain robust fore-

casts. Schapire [7] showed that weak learner could always improve its performance by training two additional predictors on filtered versions of the input data, while Breiman [1] generated multiple predictors combining them by simple averaging (regression) or voting (classification). In this study, we focus our attention on RF [1] where every weak learner is obtained by growing a non-pruned tree on a training set which is a different bootstrap sample drawn from the data. An important issue of RF is about the use of Out-Of-Bag (OOB) predictions, where for each observation $z_i = (\mathbf{x}_i, y_i)$ the algorithm computes the predictions by averaging only those trees grown using a training set not containing z_i . For improving the accuracy, the injected randomness has to maximize the differences between the trees. For this reason, in each tree node a subset of predictors is randomly chosen. The RF obtained provides an accuracy level that is in line with Boosting algorithm [5], while it is faster. In sum, RF can handle high dimensional data using a large number of trees in the ensemble, also selecting the variables so as to generate a set of very different trees. In order to assess if averaged models (BMA and RF) are better with respect to their correspondent single models (Logistic Regression and Trees) we shall compare them using out of sample measures of performance (more precisely, their ability in predicting a particular event in the validation sample). In order to detect the predictive ability of a model, we employ the confusion matrix and related measures of interest with different types of cut-off. The cut-off point can be selected taking into account the false negative classification and the a priori incidence of the dependent variable (P opt), or the value where sensitivity and specificity are equal (P fair), or maximising the Kappa statistics (P kappa). A further predictive performance measure that we shall consider is the Receiver Operating Characteristic curve (ROC) and the correspondent Area Under the Curve (AUC) (see e.g. [6]).

3 Application to risk measurement and conclusions

Our empirical analysis are based on a data set provided by Creditreform, one of the major rating agencies in Germany for SME. The data contains annual observations (from 1996 to 2004) concerning 742 SME. When handling bankruptcy data, it is natural to label the categories as success (healthy) or failure (default) and to assign them the values 0 and 1 respectively. Therefore, our data set consists of a binary response variable (default) and 10 financial ratios used as covariates. As mentioned before, on this data set we compute: Bayesian Model Averaging, Random Forest, Logistic Regression and a single Tree. In order to compare the ability in making predictions, for each method we compute some measures of performance and we report them in Table 1 and 2.

On the basis of the results obtained, we confirm that averaged methods performs better with respect to the correspondent single models. Starting from this simple conclusion, in our future research agenda we will consider the possibility of combining Bayesian Model Averaging and ensemble methods (not only Random Forest) in order to further increase the ability of making predictions. A possible way of com-

Model	AUC	Confidence Interval (Bootstrap)
Bayesian Model Averaging	0.87	(0.80 ; 0.94)
Random Forest	0.81	(0.75 ; 0.86)
Logistic Regression	0.77	(0.67 ; 0.87)
Tree	0.67	(0.57 ; 0.77)

Table 1 Model Comparison based on AUC

Model	Measures	P opt	P fair	P kappa	p=0.5
Bayesian Model Averaging	Cut-off	0.35	0.13	0.35	0.50
	Sensitivity	0.55	0.75	0.55	0.27
	Specificity	0.96	0.76	0.96	0.98
	Correct classification (%)	0.91	0.76	0.91	0.89
Random Forest	Cut-off	0.43	0.14	0.41	0.50
	Sensitivity	0.40	0.75	0.43	0.32
	Specificity	0.97	0.74	0.96	0.97
	Correct classification (%)	0.89	0.74	0.89	0.89
Logistic Regression	Cut-off	0.81	0.09	0.19	0.50
	Sensitivity	0.07	0.76	0.65	0.20
	Specificity	0.98	0.76	0.85	0.96
	Correct classification (%)	0.86	0.76	0.82	0.86
Tree	Cut-off	0.50	0.06	0.17	0.50
	Sensitivity	0.17	0.48	0.27	0.17
	Specificity	0.98	0.82	0.95	0.98
	Correct classification (%)	0.87	0.78	0.86	0.87

Table 2 Confusion matrix performance measures

binning these forecasts is to weight each of them taking into account their predictive accuracy. Hence, the final aim of our research will be of obtaining "predictors rated A⁺" using a wide variety of tools.

References

1. Breiman, L.: Random forests. *Mach. Learn.* **45**(1),
2. Capistran, C., Timmermann, A., Aiolfi, M.: Forecast Combinations. Technical Report (2010)
3. Figini, S., Fantazzini, D.: Random Survival Forests Models for SME Credit Risk Measurement. *Methodol. Comput. Appl.* **11**, 29–45 (2009)
4. Figini, S., Giudici, P.: Credit risk predictions with Bayesian model averaging. *DEM Working Paper Series*, **34** (2013)
5. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kaufman, San Francisco, 148–156 (1996)
6. Krzanowski, W.J., Hand, D.J.: ROC curves for continuous data. CRC/Chapman and Hall (2009)
7. Schapire, R.E.: The strength of the weak learnability. *Mach. Learn.* **5**(2), 197–227 (1990)
8. Steel, M.F.J.: Bayesian Model Averaging and Forecasting. *Bulletin of E.U. and U.S. Inflation and Macroeconomic Analysis*, 30–41 (2011)

New proposals for clustering based on trimming and restrictions

Luis Angel García-Escudero
Alfonso Gordaliza
Carlos Matrán
Agustín Mayo-Iscar

**Departamento de Estadística e Investigación Operativa. Universidad de Valladolid.
Valladolid and IMUVA (Instituto de Investigación en Matemáticas UVa), Spain**

Abstract TCLUS is a model-based clustering methodology that uses trimming and restrictions to get robustness and to avoid spurious solutions. It is available in the `tclust` package at the CRAN website. TCLUS methodology also includes graphical tools to assist to the users in choosing the input parameters. Extensions of TCLUS modelling include clustering around linear subspaces, mixture model estimation or fuzzy clustering. We are right now interested in other extensions of this model, like the common principal components model, the cluster weighted model or the skew-normal models. In all these cases, trimming and restrictions are applied in a natural way. Theoretical and robustness properties for the corresponding estimators can be obtained.

1 TCLUS methodology

TCLUS is a model-based clustering methodology that uses trimming and restrictions to get robustness and to avoid spurious solutions (see García-Escudero et al, 2008). We assume that the sample comes from to a known number of normal populations with possibly different scatterings and weights. Restrictions are based on the ratio between the maximum and the minimum eigenvalues of the groups scatter matrices. The input parameters for applying this model are the number of populations, the level of trimming and the level of the restrictions. It is a well-posed estimation problem, for which the existence of the estimator is guaranteed as well as its consistency to the population solution. As expected, the estimator works under the presence of a proportion of contaminating observations in the sample.

Robustness properties of TCLUS methodology related to the influence function and the breakdown point can be found in Ruwet et al (2012 y 2013)

The `tclust` package for the R environment for statistical computing (R Development Core Team 2010) implements the algorithm for obtaining the solution of TCLUS methodology. This package is available at <http://CRAN.R-project.org/package=tclust>. The package was presented in Fritz et al (2012). Details about the algorithm, the FAST TCLUS, can be found at Fritz (2011). It is a classification EM algorithm which is adapted to address the restricted maximization problem included in the estimator definition. This is done by additionally evaluating an explicit function at $2kp+1$ values within each M step. This algorithm is a substantial improvement over the initial release based on Dijkstra algorithm.

The TCLUS_T package also contains graphical tools in order to assist the users in choosing the number of groups and the level of trimming. Additionally, users can find measures for the strength of the group-assignments. All of these add-ins are described in García-Escudero et al (2011).

2 TCLUS_T extensions

Trimming methodology also works for fitting observations around linear subspaces in presence of contamination. It was applied in two ways: in the sense of assuming orthogonal errors (García-Escudero et al., 2009) or in the sense of linear regression errors (García-Escudero et al., 2010). This last approach included, apart from the typical trimming of observations far away to the model, a second trimming in order to avoid the influence of outliers in the explanatory variables.

We are also interested in estimating clusters having common principal components in a robust way. This model appears in Biometry and it corresponds to applying restrictions to the eigenvectors. These restrictions can be added to the eigenvalues ratio restrictions in a natural way.

A fuzzy extension of TCLUS_T methodology, based on the implementation of trimming and restrictions, can be found in Fritz (2013). In this case, additional input parameters, for controlling the level of fuzziness and the proportion of observations with hard assignment, appear with the ordinary ones in TCLUS_T.

We are also interested in the extension of TCLUS_T methodology to normal mixture models. Initially, our interest focused on applying eigenvalues ratio restrictions in order to avoid spurious solutions (McLachlan and Peel 2000) in the estimation of this model (García-Escudero et al, 2012). In this case, our proposal is to solve the restricted estimation problem for different levels of the restrictions determined by a grid of restriction factors in a very wide range and, after that, to choose a sensible solution among the very reduced set of essentially different obtained estimations. An extension of trimming and restrictions for estimating normal mixture models in a robust way can be found in García-Escudero et al (2013). In this case, we add trimming to the previous proposal in order to get a robust estimator. That paper includes empirical evidences of the estimator performance under contamination and theoretical proofs for the existence of this estimator and its corresponding consistency to the population solution.

Another model interesting to be incorporate to TCLUS_T methodology is the skew normal distribution. The skew-normal distribution adds flexibility to the typical clustering modelling based on the normal model. In order to work with it, the representations for the skew normal model appeared in Lee and McLachlan (2013) are very useful. The estimation of this model by using trimming and restrictions can be a competitor of estimators based on skew *t* modelling.

Cluster-weighted modelling (CWM) introduced in Gershfeld (1997) and developed in Ingrassia et al (2012a & 2012b) and Ingrassia et al (2013) appears as very interesting methodology to be modified by applying trimming and restrictions. An ongoing collaboration with Ingrassia's research group is aimed to study the properties of this modified proposal.

New research is necessary in the TCLUS framework in order to choose the number of groups, the level of trimming and the level of the restrictions in an automatic and sensible way. The research group of Parma is leading a trial in this direction.

References

1. Fritz, H.; García-Escudero, L.A. and Mayo-Iscar, A. (2012). "tclust: An R package for a trimming approach to Cluster Analysis". *Journal of Statistical Software*, Vol. 47, <http://www.jstatsoft.org/v47/i12>
2. Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2012). A fast algorithm for robust constrained clustering. *Computational Statistics & Data Analysis*.
3. Fritz, H.; García-Escudero, L.A. and Mayo-Iscar, A. (2013), "Robust Constrained Fuzzy Clustering". Preprint
4. García-Escudero, L.A.; Gordaliza, A.; San Martín, R., Van Aelst, S. and Zamar, R. (2009), "Robust Linear Clustering". *Journal of the Royal Statistical Society Ser. B*, Vol. 71, Pag. 301-319
5. García-Escudero, L.A.; Gordaliza, A.; San Martín, R; Mayo-Iscar, A. (2010) "Robust Clusterwise linear regresión through trimming". *Computational Statistics and Data Analysis*. Vol. 54, Pag. 3057-3069.
6. García-Escudero, L.A.; Gordaliza, A.; Matrán, C. and Mayo-Iscar, A. (2008), "A General Trimming Approach to Robust Cluster Analysis". *Annals of Statistics*, Vol.36, pp. 1324-1345.
7. García-Escudero, L.A.; Gordaliza, A.; Matrán, C. y Mayo-Iscar, A. (2011) "Exploring the number of groups in robust model-based clustering". *Statistics and Computing* Vol. 21, Pag. 585-599 .
8. García-Escudero, L.A.; Gordaliza, A.; Matrán, C. y Mayo-Iscar, A. (2012) "Avoiding Spurious Local Maximizers in Mixture Modeling". Preprint
9. García-Escudero, L.A.; Gordaliza, A.; y Mayo-Iscar, A. (2012). Preprint
10. Gershfeld, N. (1997). Nonlinear Inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences*, 808(1), 18-24.
11. Ingrassia S., Minotti S.C., Punzo A. (2013). Model-based clustering via linear cluster-weighted models, *Computational Statistics and Data Analysis*, in press.
12. Ingrassia S., Minotti S.C., Incarbono G. (2012). An EM Algorithm for the Student-t Cluster-Weighted Modeling, in "Gaul W., Geyer-Schulz A., Schmidt-Thieme L., Kunze J. (Eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*", Springer-Verlag, Berlin, 2012, 13-21.
13. Ingrassia S., Minotti S.C., Vittadini G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions, *Journal of Classification*, 29, n.3, 363-401
14. Lee, S.X. and McLachlan, G.J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*. To appear.
15. McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, John Wiley Sons, Ltd., New York.
16. R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
17. Ruwet, C.; García-Escudero, L.A.; Gordaliza, A. and Mayo-Iscar, A. (2012). "The influence function of the TCLUS robust clustering procedure". *Advances in Data Analysis and Classification*. Vol. 6, N. 2, Pages 107-130 "
18. Ruwet, C., García-Escudero, L. A., Gordaliza, A., & Mayo-Iscar, A. (2013). On the breakdown behavior of the TCLUS clustering procedure. *TEST*, Vol. 6, N.3, 466-487.

Formal Diagnostics for Graph Clustering: The Role of Graph Automorphisms

Andreas Geyer-Schulz and Fabian Ball

Abstract The randomized greedy (RG) family of graph clustering algorithms currently is a state-of-the-art approach for modularity maximization. These algorithms produce either a randomly selected locally maximal cluster partition of the graph or a set of these partitions. In graph clustering the symmetries of a graph indicate the presence of permutation groups which lead to graph automorphisms. In the following we show how graph automorphisms can be detected in a set of locally optimal solutions and that this information leads to a characterization of clusters and cluster elements of the optimal partitions with regard to the information revealed by the clustering in the optimal partitions.

Key words: graph clustering, automorphisms, evaluation

1 Introduction

The family of randomized greedy algorithms (RG) introduced in [8] and [6] for modularity clustering (see [5]) is a state-of-the-art approach for graph clustering. The RG algorithm (see e.g. [12]) and its ensemble variant (see [3] and [7]) won both the Quality and the Pareto challenge in the 10th DIMACS Implementation Challenge (2012).

However, since the solutions of algorithms from the family of randomized greedy graph clustering algorithms (or from any other randomized graph clustering family as e.g. label propagation ([10] and [2])) are randomly selected local graph partitions

Andreas Geyer-Schulz
Informationsdienste und elektronische Märkte, Karlsruhe Institute of Technology (KIT), Karlsruhe
e-mail: andreas.geyer-schulz@kit.edu

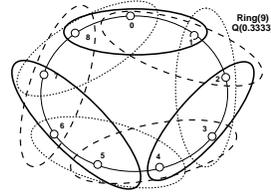
Fabian Ball
Informationsdienste und elektronische Märkte, Karlsruhe Institute of Technology (KIT), Karlsruhe
e-mail: Fabian.Ball@kit.edu

only, a further analysis of the quality of these graph partitions is necessary and, because of the effects of graph symmetries, also possible.

2 Graph Automorphisms and Modularity Maximization

Figure 1 serves as an illustration of the the role of automorphisms in graph clustering. A randomized graph clustering algorithm finds an arbitrary one of the three optimal solutions. The clusters found **depend on the starting point**, the automorphism subgroup shown in figure 1 has nine shifts on the circle.

Fig. 1 Three clusterings. (3 shifts on the circle).



Formally, we use permutation groups (see [13]) to describe this situation: Ω is a set, its elements are α, β, \dots . The symmetric group $Sym(\Omega)$ is the set of all permutations of Ω . For a finite set with n elements, we write S_n for the symmetric group. The group operation of $Sym(\Omega)$ is the composition of permutations. A permutation group on Ω is a subgroup of $Sym(\Omega)$.

The permutation $g : \Omega \rightarrow \Omega$ is a bijection (a function which is one to one and onto). We write permutations from the right: the image β of α under the permutation g is αg . We compose from left to right: The composition of g and h is gh , so that $\alpha(gh) = (\alpha g)h$ holds.

The “shifting” of the partitions over the circle with nine vertices in figure 1 is modelled by a group action on a set (and on a partition): Let G be a group and Ω a nonempty set. For each element $\alpha \in \Omega$ and each $g \in G$ we define the element α^g : $(\alpha, g) \rightarrow \alpha^g$ is a function of $\Omega \times G$ into Ω . This defines an action of G on Ω (G acts on Ω) if the following holds:

1. $\alpha^1 = \alpha \quad \forall \alpha \in \Omega$ with 1 the identity element of the group,
2. $(\alpha^g)^h = \alpha^{gh} \quad \forall \alpha \in \Omega$ and all $g, h \in G$.

A group G acting on the set Ω also acts on the the power set 2^Ω , the partitions $P(\Omega), \dots$ (see e.g. [1]).

When a group G acts on a set Ω , a typical point α is “moved” by the elements of G to other points. The set of these images is called the orbit of α under G :

$$\alpha^G = \{\alpha^g \mid g \in G\}$$

The action of a group G on a set Ω induces an equivalence relation on the set: for $\alpha_1, \alpha_2 \in \Omega$, let $\alpha_1 \sim \alpha_2$ iff there exists a $g \in G$ so that $\alpha_1^g = \alpha_2$ (α_1 and α_2 are on the same orbit). The orbits of a group partition the set Ω . All elements of an orbit are isomorphic. Applied to our motivating example, this implies that the three shifted partitions belong to a group orbit and that the membership of a vertex to a cluster reveals no information. As a consequence, a formal condition for a partition “to reveal the hidden structure” in a graph is that the automorphism group of the partition is trivial (contains only the identity element). This also implies that all known measures for comparing partitions which use only two partitions (e.g. the RAND index) fail to detect arbitrary group automorphisms.

$\Gamma = (V, E)$ is an undirected, loop-free graph with vertex set $V = \{1, \dots, n\}$, and edge set $E \subseteq V \times V$. P is a partition of V . Its automorphism group $Aut(\Gamma)$ is $\{\gamma \in S_n \mid \Gamma^\gamma = \Gamma \text{ and } P^\gamma = P\}$ where $\Gamma^\gamma = (V^\gamma, E^\gamma)$, $E^\gamma = \{(x^\gamma, y^\gamma) \mid (x, y) \in E\}$. The automorphism group of a graph contains all permutations of vertices that map edges to edges and vertices to vertices, and it is a subgroup of $Sym(V)$ (see [4]).

Next, we consider the effects of the group action on the cluster criterium of the graph clustering algorithm, namely the modularity $Q(\Gamma, C) = \sum_{i=1}^p (e_{ii} - a_i^2)$ where $C = \{C_1, \dots, C_p\}$ is a partition of V , and p the number of clusters. Let $x, y \in V$. The adjacency matrix M of Γ is $m_{xy} = m_{yx} = \begin{cases} 1 & \text{if } (x, y) \in E \\ 0 & \text{otherwise} \end{cases}$. Then $e_{ij} = \frac{\sum_{x \in C_i} \sum_{y \in C_j} m_{xy}}{\sum_{x \in V} \sum_{y \in V} m_{xy}}$ is $P(\text{random edge between } C_i \text{ and } C_j)$ and e_{ii} is $P(\text{random edge in } C_i)$. $a_i^2 = (\sum_j e_{ij})^2$ is $P(\text{random edge in } C_i)$ for a random graph with nodes with the same degree as Γ .

Hans Rademacher [9] introduced the following definition: A function is said to belong to a group if it is invariant under the transformations of that group. A trivial example is the function $\sin(z)$: The function $\sin(z)$ is periodic with a period of 2π : $\sin(z) = \sin(z + 2\pi) = \sin(z + 2k\pi)$. We write $z' = z + 2k\pi$ and then $\sin(z) = \sin(z')$. The sine function does not change when its argument is transformed by a linear transformation from the infinite group $z' = z + 2k\pi$.

We show that $Q(\Gamma, C)$ belongs to the automorphism group $Aut(\Gamma, C)$: It is invariant under the group actions of $Aut(\Gamma, C)$. This is a direct consequence from the isomorphism of all elements on a group orbit.

However, $Q(\Gamma, C)$ is also invariant with regard to certain congruences: For example, Q is invariant for a ring of subgraphs with the same connection to the ring (outer structure) and the same number of edges in the subgraph (inner structure). The subgraphs may have a different inner structure.

In addition, there exists the following correspondence of permutations and partitions (see e.g. [11, p. 158]): Any permutation h of S_n generates a subgroup H of S_n , and, through the orbits of the subgroup H a partition of n : The number of parts of the partition is the number of summands of the partition. Two elements of α, β of the partition lie in a block of the partition if and only if $\beta = h(\alpha)$ for some $h \in H$. We give several examples how this correspondence can be used in the analysis of randomized hierarchical clustering algorithms.

3 Types of Optimal Solutions of a Graph Clustering Problem

The set O of locally optimal partitions is the result of several runs of a randomized graph clustering algorithm. We distinguish the following solution types:

- A unique solution: a single partition with the maximal value of Q .
- More than one partition with the maximal value of Q :
 - The partitions result from a graph automorphism. We determine the degree of the automorphism group and the information content of the partitions.
 - The partitions do not result from a graph automorphism. We have found multiple solutions which may have a different discipline specific interpretation.

References

1. Thomas Beth, Dieter Jungnickel, and Hanfried Lenz. *Design Theory*. Cambridge University Press, Cambridge, 1993.
2. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):1–13, 2008.
3. Andreas Geyer-Schulz and Michael Ovelgönne. The randomized greedy modularity clustering algorithm and the core groups graph clustering scheme. In Wolfgang Gaul, Andreas Geyer-Schulz, Akinori Okada, and Yasumasu Baba, editors, *German-Japanese Interchange of Data Analysis Results*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 15–34, Heidelberg, 2013. Springer.
4. Gordon James and Adalbert Kerber. *The Representation Theory of the Symmetric Group*, volume 16 of *Encyclopaedia of Mathematics and Its Applications*. Addison-Wesley, Reading, 1981.
5. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, Feb 2004.
6. Michael Ovelgönne and Andreas Geyer-Schulz. Cluster cores and modularity maximization. In *ICDMW '10. IEEE International Conference on Data Mining Workshops*, pages 1204 – 1213, Piscataway, 2010.
7. Michael Ovelgönne and Andreas Geyer-Schulz. An ensemble learning strategy for graph clustering. In David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner, editors, *Graph Partitioning and Graph Clustering*, volume 588 of *Contemporary Mathematics*, pages 187–205, Providence, 2013. American Mathematical Society.
8. Michael Ovelgönne, Andreas Geyer-Schulz, and Martin Stein. Randomized greedy modularity optimization for group detection in huge social networks. *4th ACM SNA-KDD Workshop on Social Network Mining and Analysis.*, 2010:1–9, 2010.
9. Hans Rademacher. *Functions Belonging to Groups*, chapter 10, pages 115 – 124. Birkhäuser, Boston, 1983.
10. Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106, Sep 2007.
11. Ambar N. Sengupta. *Representing Finite Groups : A Semisimple Introduction*. Springer New York, New York, NY, 2012.
12. Martin Stein and Andreas Geyer-Schulz. A comparison of five programming languages in a graph clustering scenario. *Journal of Universal Computer Science*, 19(3):428 – 456, 2013.
13. Helmut Wielandt. *Finite Permutation Groups*. Academic Press, New York, 1964.

An overview on multiple regression models based on permutation tests

Massimiliano Giacalone¹ and Angela Alibrandi²

Abstract

When the population, from which the samples are extracted, is not normally distributed, or if the sample size is particularly reduced, it becomes preferable to use nonparametric statistical tests. Within the regression models, it is possible to use permutation tests, considering their ownerships of optimality, when the normal distribution of the response variables is not guaranteed especially in the multivariate context. In the literature there are numerous permutation tests applicable to the estimation of regression models. In this paper we focused our attention on the permutation tests of the independent variables, proposed by Oja, and other methods of nonparametric inference, in the regression models context.

1 Permutation test in regression models

In many cases, when the classical conditions of regression models are not respected, it's possible to use the permutation tests, considering their ownerships of optimality, especially in the multivariate context. The evaluation of the parameters' significance is an inferential procedure, based on randomization tests (if the same experimental plan justifies them) or permutation tests (if the observed samples are random, so that the analyzed samples justify the calculations) (Kempthorne and Doerfler, 1969). Through the use of permutation tests, we assess the null hypothesis of casualness: in fact, it suggests that, if the examined phenomenon has a certain tendency, confirmed by a model that appears as gives, it is a purely accidental effect of the observations in casual order. We proceed choosing a useful S statistic test to

¹ Massimiliano Giacalone, lecturer with annual appointment, Bologna University, email: massimilia.giacalone@unibo.it

² Angela Alibrandi, Department of Economical, Business, Environmental Sciences and Quantitative Methods, Messina University, email: aalibrandi@unime.it

measure the entity of the phenomenon of interest in relation to the observed data and we compare the observed s statistic test value of S and the distribution of S_r obtained by casually rearranging the data. The test is based on the following principle: if the null hypothesis were true, then all the possible arrangements of the observations would have equal probability to verify, that is the order of the observed data is one of the possible equally probable arrangements and s appears as one of the possible values of the randomization distribution of S . If s is a significant value, then the null hypothesis is rejected, then, for implication, the alternative hypothesis is considered more reasonable. The significance level of s is, the percentage of the values that are great or equal to s in the randomization distribution. It represents a measure of evidence strength against the null hypothesis.

2 The Oja permutation test of the independent variables

The experimental plan presented by Oja (1987) considers n subjects to which a treatment variable x is assigned in order to study their effects on a response variable Y . In addition, for each k subject, further Z explanatory variables (covariates) are considered. The non-parametric permutation tests proposed by Oja are relative to a completely permuted plane: in fact, they are based on the assumption that the treatment values are randomly assigned to the subjects. Therefore, the permutation distribution used to verify the significance of a relationship between X and Y , taking into account the effects of the Z covariates, is obtained by permuting the X values to the n statistical units. In formal terms, this is a regression plan model where the results can be generalized to multiple regression. The model can be expressed as:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i \quad (1)$$

with $i=1, \dots, n$, where α , β and γ are unknown parameters, X is the explicative variable of the plane such that $\sum_i^n X_i = 0$, z is the explanatory covariable and $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed random errors with zero mean. The attention is focused on the β parameter; therefore the null hypothesis is expressed by $H_0: \beta=0$; α and γ are nuisance parameters. Let's suppose that Y , X and Z are given for all i : the X variable is considered as a realization of the random permutation x^* of x . Then, the corresponding y values, which have not been realized, are $y^* = y + \beta(x^* - x)$, from which we can easily obtain $y^* - \beta(x^* - x) = y - \beta x$. The test statistic proposed by OJA to assess the null hypothesis is:

$$T = \sum_{i < j < k} \Delta_{ijk}^y \Delta_{ijk}^{x^*} \quad \text{where} \quad \Delta_{ijk}^y = \begin{vmatrix} 1 & 1 & 1 \\ y_i & y_j & y_k \\ z_i & z_j & z_k \end{vmatrix} \quad (2)$$

with $i < j < k$ and similarly for $\Delta_{ijk}^{x^*}$.

This statistic is not easily calculable; so, Oja proposed an alternative form of this test in order to facilitate the calculations:

$$T = \sum_i \hat{y}_i x_i^* \quad \text{where} \quad \hat{y}_i = \sum_{j < k} \Delta_{ijk}^y \delta_{jk}^{x^*} \quad \text{with} \quad \Delta_{jk}^{x^*} = \begin{vmatrix} 1 & 1 \\ z_j & z_k \end{vmatrix} \quad (3)$$

where $j < k$. Collins (1987) tried to approximate the permutation distributions proposed by Oja, with other distributions. In particular Oja has suggested a standardized normal approximation or, equivalently, to square the test statistic proposal and compare the

result with the critical value of a χ^2 . Of course there is no certainty that these distributions provide adequate approximation to the corresponding distributions of permutation. Collins (1987) has proposed, in his work, a reformulation of the Oja statistics, using easier methods of calculation, to obtain the explicit formulas of the moments of permutation and especially to have the advantage of being able to recognize a beta distribution as an approximation of the exact distribution of null permutation. These procedures presented by Oja and Collins have not had much success and development because, by permuting the independent variables, they violate the principle of ancillarity, according to which the plan should be subject to maintain the collinearity between the explanatory variables (Kennedy, 1995).

3 Some alternative approaches of permutation tests in regression

Other methods of nonparametric inference, in a regression model are:

- **The residual permutation test of the complete model**, proposed by Ter Braak (1992), that is analogous to a bootstrap test and it consists of permuting the residual samples of a multiple regression in order to produce a distribution that can be compared with the value sample of a statistical test. In effects, this test is not a permutation test in the traditional sense, because the data are transformed to get the residues, before their exchange happens. Moreover, it is hybrid between a permutation test and a bootstrap test and its justification can be derived from both value b^* around the true b value in the bootstrap samples. Similarly the variability of F_{obs} to test $\beta = \beta_0$ are similar to the variability of F^* to test $\beta = b$. These appreciable properties are also justified because the F used statistic test is asymptotically pivotal; whatever the distribution of errors, the F asymptotic distribution doesn't depend on the parameters in the model. This test is well applicable with great samples because the variability of b around the truth β is similar to the variability of the resampled that are not tested (Levin and Robbins, 1983; Gail, Tan and Piantadosi, 1988; Kennedy and Cade, 1996).
- **The permutation test of the dependent variable**, used to verify the null hypothesis $H_0: \beta = 0$. It can be performed by comparing the values of the F test statistic with the distribution obtained by permuting the Y observations, to casually assign to the sets of the observations of X and Z independent variables. Manly (1991) proposes three possible motives to justify this type of permutation approach: in first place, the n observations can be a casual sample from a population of possible observations, where the Y variable could be independent from the X and Z explanatory variables; in according to place, the values of the experimental variables X and Z can casually be assigned to n statistical units and, therefore, the values of the Y response variable can be observed (Y would not to be influenced by the X and Z variables). Besides, if the variable Y and the explanatory variables X and Z are independent, all the possible joining among every value Y and every values X and Z are equally probable in relation to a potential mechanism that generates the data;
- **The exact restricted permutation tests for partial regression models**, fundamentally developed by Brown and Maritz (1982), furnishes an exact permutation test for a partial regression model, within the regression plain. The

proposed scheme, united to a suitable experimental plan, is used for the inference on the regression coefficient β of X , when exists another explanatory Z variable that influences the Y response. The X coefficient is therefore a disturb parameter.

4 Final remarks

In this paper we revisited the use of permutation tests to evaluate non parametric inference in a regression model. Comparing the randomization and permutation tests with the conventional test for inference in a regression model, we can underline some aspects. First of all, the randomization and permutation tests have two important advantages: they are valid and opportunely applicable without casual samples and they allow to select a statistic test appropriated for a particular considered situation. Nevertheless, it's not possible to generalize the conclusions of a randomization test to the whole population of interest. In fact a randomization test identifies the probability that a phenomenon of interest is casual. The concept of population from which to extract samples of observations is not fundamental and this is the reason for which the casual sampling is not required. In the other hand, the generalization of results of the conventional tests to the whole population is based on the assumption, not always verifiable, that the observed samples are equivalent to a casual sample or that the data are available for the whole population of interest (but this last condition is practically unrealizable). So, randomization and permutation tests represent a methodologically adequate solution in a large number of practical experimental contexts in which the samples are not random. These methods seem to be appropriate for particular conditions, alternatively to conventional tests whose assumptions are too restricted.

References

1. Brown, B. M., Maritz, J. S.: Distribution- free methods in regression. In: Australian Journal of Statistics, 24, pp. 318-333 (1982).
2. Collins, M. F.: A permutation test for planar regression. In: Australian Journal of Statistics, 29, pp. 303-308 (1987).
3. Gail, M. H., Tan, W.Y. , Piantadosi, S.: Tests for no treatment effect in randomized clinical trials. In : Biometrika, 75, pp. 57-64 (1988).
4. Kempthorne, O., Dorfler, T.E.: The behavior of some significance tests under experimental randomization. In: Biometrika, 56, pp. 231-248 (1969).
5. Kennedy, P.E.: Randomization tests in econometrics. In: Journal of Business and Economic Statistics, 13, pp. 85-94 (1995).
6. Kennedy, P.E, Cade, B.S.: Randomization tests for multiple regression. In: Communication in Statistics – Simulation and Computation, 25, pp. 923-936 (1996).
7. Levin, B., Robbins, H.: Urn models for regression analysis, with applications to employment discrimination studies. In: Law and Contemporary Problems, 46, pp. 247-267 (1983).
8. Mainly, B. J. F.: Randomization and Monte Carlo methods in biology. Chapman and Hall. London (1991).
9. Oja, H.: On permutation tests in multiple regression and analysis of covariance problems. In: Australian Journal of Statistics, 29, pp.91-100 (1987).
10. Ter Braak, C. J. F.: Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Bootstrapping and Related Techniques, (K. H. Jockel, G. Rothe and W. Sendler, Eds.) New York, Springer Verlag, pp. 79-85 (1992).

The determinants of Italian students' reading scores: a Quantile Regression analysis

Francesca Giambona and Mariano Porcu

Abstract In recent years the measurement of students' achievement has received a good deal of attention. Empirical studies have found that students' characteristics, family background, school attended and territorial differences have a prominent role in affecting students' performances. Through a Quantile Regression approach we analyse Italian students' reading achievement using the OECD-PISA 2009 survey data. Results, compared to the usual OLS mean regression, show a strong and differentiated effect of some of the selected predictors on students' achievement, also highlighting different paths of reading achievement for different quantiles (i.e., for different types of students).

Key words: reading achievement, quantile regression, OECD-PISA

1 Introduction

Education is crucial both for individuals' life and for the economic development of the countries in order to foster economic growth, enhancing productivity, improve social development and to reduce social inequality. Higher education is associated with markedly higher earnings, lower unemployment, higher labour force participation and to a longer healthy life. The key role of education in social and economic policies has highlighted the need to control and to monitor the educational processes into countries. Since Nineties international surveys have been addressed to quantify students' performances in different fields of knowledge and to compare the performances of the different educational systems worldwide. Since the year 2000 the

Francesca Giambona
Università degli Studi, Cagliari, e-mail: francesca.giambona@unica.it

Mariano Porcu
Università degli Studi, Cagliari, e-mail: mrporcu@unica.it

OECD carries on the Program for International Student Assessment (PISA). It is administered every three years to provide comparisons of 15-16-year-old students' achievement in reading, mathematics and science among the participating countries, and with a major focus on one of the three competencies. PISA 2009 marks the beginning of a new round with a return to a focus on reading and offers the most comprehensive and rigorous international measurement of students' reading skills. In order to explain the differences in students' performances, PISA survey collects sociodemographic information at students, families and schools levels. In this paper, PISA 2009 data regarding Italy have been considered. Many studies have analyzed the determinants of students' achievement using the standard regression methods based on the Ordinary Least Squares (OLS) to estimate the effect of predictor variables on the students' achievement (the response variable, measured, usually, as a test score). Since OLS estimators disclose the effect of predictor variables at one point of the distribution of the dependent variable (the conditional mean) the information gathered by OLS regression is limited to this specific point of the distribution. In terms of achievement this can lead to imprecise, or at least incomplete, findings as it is possible that the effects of the predictors are different at other points of the distribution, i.e. at different quantiles. To this aim, using the last PISA 2009 survey, we will perform a Quantile Regression (QR) model to estimate the effects on the Italian students' reading score of some predictors related to some features already highlighted in previous studies (see, for example, [2, 1]), in order to describe the effect of predictor variables along the entire students' achievement scores distribution; i.e., we are interested in assessing the changes of the predictor influence in the different points of the distribution, compared with the usual OLS conditional mean regression.

2 The Quantile Regression

Originally, QR was suggested by Koenker [3] as a "robust" technique alternative to Ordinary Least Squares (OLS) when the errors are not normally distributed. Contrary to the usual OLS mean regression, quantile regression aims at estimating a selected conditional quantiles of the response variable (i.e. the median). QR allows to estimate the whole of quantiles of the conditional distribution of the response variable and to assess the influence of predictors on the shape of the distribution.

The basic QR model specifies the conditional quantile as a linear function of predictors. The τ -th regression quantile ($0 < \tau < 1$) of y is get minimizing the sum of the absolute deviations residuals:

$$\min_{\beta^\tau} \sum_{\sigma_k < 0} \tau |y_k - X_k \beta^\tau| + \sum_{\sigma_k > 0} (1 - \tau) |y_k - X_k \beta^\tau| \quad (1)$$

where τ determines the selected conditional quantile of interest. Hence, any one of the components of the QR coefficients, $\beta(\tau)$, provides an estimate of the marginal effect of the corresponding independent variable on the dependent one for the τ -th

quantile, controlling the effects of the remaining variables. The important feature of this method is that the marginal effects of the independent variables, given by $\beta(\tau)$, may vary over quantiles.

3 Results

For the analysis we use as response variable the five Plausible Values (or PVs) provided by PISA dataset as indicators of students' reading achievement. PVs are a recent innovation in item response theory and they are increasingly used in surveys on student achievement. Plausible values are multiple estimates of students' achievement needed as PISA sample units did not take the full battery of assessment items (each student was given a subset of items). For PISA 2009, five plausible values were computed for each student respondent, indicating possible "true" values of the students score on the underlying conceptual dimension (reading, mathematics, science), by considering a plausible distribution according to the results of the test. Consequently, the standard error calculation has to take into account the sampling variance in the estimate of the response variable. Then, we consider the average coefficients got using all the five PVs in reading available for each student. Furthermore we use the *final student weight* to obtain estimates representative of the population and, thus, to make generalizations to the national population of the 15-16-year-old students represented by PISA 2009. Moreover, to take into account for sampling design we use the 80 replicate weights available in the PISA 2009 *student file* to obtain unbiased standard errors (for a detailed description see [4]).

Following the main findings of the literature, we consider as predictor variables: i) students' characteristics, ii) some students' family background (related to the economic, social and cultural status and home educational resources), iii) the familiarity with ICT (Information and Communication Technology), iv) two control variables to take into account for regional differences and for the school program attended. We estimate a linear regression function at different quantiles from ($\tau = 0.1$) to ($\tau = 0.9$), with step equal to 0.1, and we examine if there is homogeneity in the effect of the predictors. For brevity, we consider here only three quantiles in order to assess the effect of the predictors on reading achievement for the the higher-performer, the median-performer and lower-performer students (for sake of brevity, the Table with estimates is omitted).

Considering the family background, students' reading performance is better for the students who have parents graduated and with a higher occupational status. Nevertheless, the effect of the educational level of parents is different in the two opposite quantiles of the distribution (0.1, 0.9); the positive effect for the lower-performer students is stronger than for higher-performer. Family cultural possessions (based on having classical literature books, books of poetry and works of art at home), home educational resources and books in the home have positive effects on reading score whilst the index of family wealth shows a negative sign. This evidence highlights the role of family possessions directly related to education rather than a general avail-

ability of goods like cellular phones, televisions, computers, cars, etc.. The positive effect of home educational resources on reading score for lower-performer students is stronger. This could mean that lower-performer students need to study with educational support at home (e.g., a desk and a quiet place to study, technical reference books, dictionary, etc.), while, on the contrary, the availability of books at home improves the achievement of higher-performer rather than the lower. Female students perform better than males, but the gender effect is more considerable for male lower-performer students than the higher; the same if the student has anticipated the entrance to school (at 5 instead of 6 years of age); while if the student has not repeated at least a year at the lower secondary education level her/his reading achievement is better with not relevant change considering the different quantiles. Results suggest a significant positive effect of having a computer at home, with a stronger effect for the lower-performer students. Although students from the “Liceo” perform better than the others, the effect related to the school attended change along the distribution, thus the higher-performer students enrolled in a Vocational program achieve better than the lower. Furthermore, a better disciplinary climate at school improves the achievement, especially for the lower-performer students.

Finally, the regional variable shows the presence of noteworthy territorial effects in reading achievement. Students in Northern regions have better reading scores: the average scores differ strongly among the Northern and the Southern regions producing a wide North-South literacy-divide.

Moving from South to North, the effect of region variable differs significantly along the distribution. In Southern regions the effect increases over the quantiles (increasing slope). Nonetheless, the inverse U-shaped form of the distribution in Northern regions highlights that the effect is approximatively the same for the two extreme types of students. In particular, in these regions we observe a different slope before and after the median value: increasing and decreasing, respectively.

Results suggest that the relationships between achievement and its predictors differ at different points of the conditional distributions, suggesting: i) QR approach is more appropriate than OLS, ii) the need of different types of policies, for lower-performer students or higher-performer or both (i.e., when the effects are homogeneous in the entire achievement distribution).

References

1. Agasisti, T. and Vittadini, G.: Regional Economic Disparities as Determinants of Students' Achievement in Italy. *Research in Applied Economics*. **41**, 33–54 (2012)
2. Bratti, M. and Checchi, D. and Filippin, A.: Geographical Differences in Italian Students' Mathematical Competencies: Evidence from PISA 2003. *Giornale degli Economisti e Annali di Economia*. **663**, 299–333 (2007)
3. Koenker, R. and Bassett, G. Jr.: Regression Quantiles. *Econometrica*. **461**, 33–50 (1978)
4. OECD: PISA 2009 Technical Report. PISA, OECD Publishing (2012) doi: 10.1787/9789264167872-en

The R Package ThreeWay

Paolo Giordani, Henk A.L. Kiers and Maria Antonietta Del Ferraro

Abstract The R package `ThreeWay` is presented and its main features are illustrated. The aim of `ThreeWay` is to offer a suit of functions for handling three-way arrays. In particular, the most relevant available functions are `T3` and `CP`, which implement, respectively, the Tucker3 and Candecomp/Parafac methods. They are the two most popular tools for summarizing three-way arrays in terms of components. After briefly recalling both techniques from a theoretical point of view, the functions `T3` and `CP` are described by considering two real life examples.

Key words: multiway analysis, Tucker3, Candecomp/Parafac, R, `ThreeWay`

1 Introduction

In statistics, data generally refer to the observations of some variables on a set of units and are stored in a (two-way) matrix, say \mathbf{X} of order $(I \times J)$, where I and J denote the number of units and variables, respectively. However, in several situations, the available information consists of some variables collected on a set of units on different occasions and is usually stored in a (three-way) array, say $\underline{\mathbf{X}}$ of order $(I \times J \times K)$, with generic element x_{ijk} , $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$ and $k \in \{1, \dots, K\}$ where K denotes the number of occasions. The two most popular techniques for

Paolo Giordani
Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, Italy e-mail:
paolo.giordani@uniroma1.it

Henk A.L. Kiers
University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands e-mail:
h.a.l.kiers@rug.nl

Maria Antonietta Del Ferraro
Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, Italy e-mail: mariaantoni-
etta.delferraro@yahoo.it

performing component models on three-way data are the Tucker3 (T3) and Candecomp/Parafac (CP) models. The aim of this work is to illustrate the R [6] package `ThreeWay` [3] for performing a complete three-way analysis [5]. The paper is organized as follows. In Section 2, we introduce the T3 and CP models. Then, Section 3 is devoted to the main features of `ThreeWay` with particular reference to the implementation of T3. Finally, in Section 4 some concluding remarks are given.

2 The Tucker3 and Candecomp/Parafac models

The Tucker3 (T3) model [7] is a multi-linear model summarizing $\underline{\mathbf{X}}$ by extracting different components for every mode. The Tucker3 model with P components for the unit mode, Q for the variable mode and R for the occasion mode, can be formalized as

$$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}. \quad (1)$$

where a_{ip} , b_{jq} and c_{kr} express the component scores of the i -th unity on the p -th component for the unit mode, of the j -th variable on the q -th component for the variable mode and of the k -th occasion on the r -th component for the occasion mode, respectively. Furthermore, g_{pqr} is the generic element of the so-called core array giving the interaction among the components of the three modes and e_{ijk} is the error term. In the T3 model, limited numbers of components for *all* the three modes are sought. The obtained solution is not unique. Such an indeterminacy can be used in order to obtain simple structure solutions. The Candecomp/Parafac (CP) model [2, 4] aims at reducing $\underline{\mathbf{X}}$ by extracting the same number of components, say S , for every mode. It can be written as

$$x_{ijk} = \sum_{s=1}^S a_{is} b_{js} c_{ks} + e_{ijk}, \quad (2)$$

with obvious notation. Differently from T3, the CP solution is unique under mild conditions. In `ThreeWay`, the CP model is implemented in the function `cp` and the T3 model in the function `T3`.

3 The Function `T3` of the R package `ThreeWay`

We apply the T3 model to the ‘Learning to read’ data [1]. The data set refers to the process of learning to read of seven pupils ($I = 7$). Five tests ($J = 5$) are used to evaluate the learning process: each test measures different reading aspects. The pupils are tested weekly from week 3 to week 47 except for 10 holidays weeks, hence $K = 37$. The aim of the study is to investigate the learning process and whether

the performances of the pupils are equal over time. In the following, we are going to summarize the most relevant steps for carrying out a three-way data analysis pipeline. Thus, some steps are omitted for the sake of brevity. We load the data and run the function `T3`.

```
R> library("ThreeWay")
R> data("Bus")
R> t3bus=T3(BusN)
```

The first relevant point to be addressed concerns preprocessing. In fact, prior to fitting a model to the data, it is fundamental to decide how to preprocess the data. Preprocessing can be done by centering within a mode or a combination of modes and normalizing across a mode. In the package `ThreeWay` this can be done using functions `cent3` and `norm3`, respectively. These functions are automatically implemented when launching `T3`. In our analysis we decide to summarize the data using two components for the pupil mode ($P = 2$), one for the test mode ($Q = 1$) and two for the time occasion mode ($R = 2$). In this way the fit of the model is very high (96.26%). Note that the `T3` solution is obtained in `T3` calling the function `T3func`. The next step of `T3` concerns how to simplify the solution taking into account the `T3` indeterminacy. The resulting output (an object of class `list` called `t3bus`) is very easy to interpret. To analyze the dynamics of the occasion component scores, we represent them in Figure 1. The first component can be interpreted as the ‘General performance level’ because the scores are close to 0 in the beginning and close to 1 at the end of the testing time. The second component is more complex due to the negative scores in the second half of the occasions. It is interpreted approximately as the ‘Learning rate’ but the negative values do not indicate that the learning rate decreases in the end: it is due only to the rescaling procedure. The component for

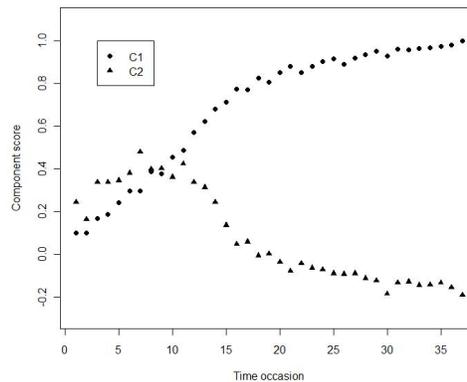


Fig. 1 Component scores for the occasion mode.

the 5 tests (scores not reported) is connected with the ‘Difficulties of the items’. Tak-

ing into account that the core is simplified to the identity matrix (because $P=QR$), we can deduce that the first and second components for the time occasions are related, respectively, to the first and second components for the pupils. Therefore, the pupils whose component scores are high are those who have a performance level (first component) and a learning rate (second component) above average.

```
R> print(round(t3bus$A, 2))
R>      A1      A2
R> n.1  1.06  -0.42
R> n.2  0.96  -0.30
R> n.3  0.99  -0.38
R> n.4  1.28   1.00
R> n.5  1.16   0.19
R> n.6  1.09  -0.01
R> n.7  0.89  -0.42
```

4 Final remarks

The most relevant features of the R package `ThreeWay` have been introduced by an example. `ThreeWay` offers a suite of about fifty functions for handling three-way arrays. Such functions carry out an interactive three-way analysis calling several additional functions to further extract relevant information from the data under investigation. The need for an interactive analysis arises because all the steps of a three-way analysis should not be done automatically. Nonetheless, the package also contains non-interactive functions useful for simulation studies.

References

1. Bus, A.G.: A longitudinal study in learning to read. Paper presented at the 9th World Congress on Reading, July 26–30, Dublin, Ireland (1982)
2. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n -way generalization of Eckart-Young decomposition. *Psychometrika* **35**, 283–319 (1970)
3. Del Ferraro, M.A., Kiers, H.A.L., Giordani, P.: `ThreeWay`: Three-Way Component Analysis. R package version 1.1, <http://CRAN.R-project.org/package=ThreeWay> (2013)
4. Harshman, R.A.: Foundations of the Parafac procedure: models and conditions for an “explanatory” multimodal factor analysis. *UCLA Working Papers in Phonetics* **16**, 1–84 (1970)
5. Kroonenberg, P.M.: *Applied Multiway Data Analysis*. John Wiley & Sons, Hoboken, NJ (2008)
6. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/> (2012)
7. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311 (1966)

Co-occurrence Network from Semantic Differential Data

Giuseppe Giordano and Ilaria Primerano

Abstract The aim of this paper is to propose a method to transform semantic differential data in a network structure, whose graph representation is interpreted as an empirical *adjectival graph*. The derived graph is constituted by the adjectives of the semantic differential task as nodes, two nodes are linked depending on the scoring assigned by the set of respondents. Semantic differential data are handled by means of a peculiar coding that induces a weighted adjacency matrix. This relational data structure allows us to realize, through a graph representation of the adjectives network, an *adjectival graph* of the concept under study. The weighting system is given by the co-occurrence of respondents' scoring. In this way, the cohesive part of the network (the core) is constituted by the set of the adjectives chosen by the most part of respondents, whereas those signifiers (adjective-nodes) less tied to the underlying concept, will locate in the peripheral part of the network. The proposed approach to *adjectival graph* aims at defining the concept-stimulus by looking at the edges between the various adjectives that characterize it. A case study is presented to show the significance of the proposed approach, while simulations will aid to validate the results.

Key words: Adjectival Network, Graph Component, Network Density, Social Network Analysis, Weighted Adjacency Matrix.

Giuseppe Giordano

Dept. of Statistics and Economics, University of Salerno, Via Giovanni Paolo II, 132, e-mail: ggiordan@unisa.it

Ilaria Primerano

Dept. of Statistics and Economics, University of Salerno, Via Giovanni Paolo II, 132, e-mail: iprimerano@unisa.it

1 Semantic Differential and Data Structure

The Semantic Differential [5] is a technique of psychometric assessment, designed in the '50s by the American psychologist Charles Osgood in order to operationalize the measure of the implied meaning of linguistic terms. Osgood proposed the Semantic Differential technique aiming to detect the meanings that concepts take on for the interviewed.

According to Osgood, concepts are multidimensional and their dimensions define the so-called semantic space whose structure is stable, while the placement of the concept in this space varies among subjects. This technique is widely used for the exploration of the dimensions of *meaning*. It is based on the assumption that the different aspects contribute to the overall meaning of the stimulus.

The classical analysis of the Semantic Differential consists of the summary description of the induced stimulus evaluated through its average profile, i.e. the average ratings of respondents.

Typically, due to the specific structure of the Semantic Differential data, researchers' interest turned to the variability of the scores given by each respondent on each pair of adjectives. We distinguish between two sources of variability, the *variability for each individual*, i.e. the difference between the pairs of adjectives and the *variability for each pair of adjectives*, i.e. the difference between the scores attributes by individuals. The analysis may focus on both sources of variability, defining from one hand the profile of the respondent (within), and on the other one the average profile of the whole set of respondents (between).

To exploit the multidimensionality features of the Semantic Differential technique, the characterization of meaning has been addressed in the framework of exploratory multidimensional data analysis by means of factorial techniques and Multidimensional Scaling [3] that allows to identify the semantic dimensions that underpin the individual evaluations. In order to verify the existence of such dimensional structure, it is also used the Principal Components Analysis, in which each pair of adjectives constitutes a scale that corresponds to a variable in the data matrix. The main purpose is to identify the direction and the intensity of the opinion expressed by respondents indicating the position between the two adjectives of the continuum, according to their perception of the stimulus being analyzed. Furthermore, a Hierarchical Clustering on the factorial results is useful to verify how the different respondents cluster together. Recognizing such clusters is useful in detecting homogeneous meanings given to the stimulus.

In this paper we focus on a different way to look how respondents are able to highlight similarity among adjectives. The main idea is introducing in this framework the concept of relational ties as defined in Social Network Analysis (SNA) [6]. We present the SNA terminology and give the suitable interpretation of the emerging adjectival network [4]. In the next section we give the basic definition and discuss the proposed method to build the adjectival network from Semantic Differential data.

2 Network Data from adjectives co-occurrences

An important facet of Semantic Differential is the possibility to describe the meaning of the underlying stimulus where adjectives take part with different importance. We exploit this feature by considering a relational tie between any two adjectives. The tie indicates a common role played in the definition of meaning. In other words, we aim at defining an emerging network of adjectives where the nodes are the initial set of adjectives used for the Semantic Differential and the ties depend on the occurrence of positive scores given by a set of respondents.

Indeed, the most important step of our analysis is to consider each adjective in relation to each other, depending on the intensity of the scores provided by a group of respondents considered as a whole. Since, in a Semantic Differential task, any pair of adjectives represent opposed meanings, particular importance is devoted to the ability of each adjective to polarize the scale. Therefore, we will study the role and the position assumed by each adjective in the network describing how it influences the definition of the stimulus.

Semantic Differential data are handled in an original way by proposing an ad-hoc encoding able to define an Incidence Matrix [2] (respondents per adjectives). This dataset is transformed into a graph structure representing the adjectival network of the respondents' perception.

Let $\mathbf{X}(n,p)$ be the matrix holding the scores given by n individuals on p pairs of bipolar adjectives in a Semantic Differential task. The columns of the matrix \mathbf{X} are doubled [1], allowing one column for each adjective. These doubled scores are centered so that the negative values indicate low importance in defining the stimulus, conversely, the positive ones indicate the existence of a direct meaning, whose intensity depends on the score. Since we are not interested in scoring the lack of importance, a value of zero to all negative scores is assigned.

The matrix containing the doubling coded scores ($\bar{\mathbf{X}}$) is read in the scope of network analysis as a weighted Incidence Matrix. Namely, we need to measure how much the co-occurrence of any pair of adjectives is positively scored by respondents. So we derive the adjective-by-adjective adjacency matrix as $\mathbf{W} = \bar{\mathbf{X}}'\bar{\mathbf{X}}$. It shows all possible co-occurrences between the adjectives and their weight in the definition of the stimulus. The matrix \mathbf{W} is a symmetric weighted adjacency matrix, in which the weights correspond to the cross-product of the scores given by all respondents to any pairs of adjectives. The main diagonal is the scoring sum of squares get by any single adjective, the larger the entry, the more important is the adjective in the stimulus definition (since it has been positively scored by a large part of respondents).

To derive normalized weights in $[0,1]$, the elements of the adjacency matrix \mathbf{W} are divided by their theoretical maximum value, given by the product of two quantities: the squared maximum score of the Semantic Differential scale times the number of respondents. This normalization proportionally rescale the importance of each adjective (main diagonal). Whereas, the derived co-occurrences (off-diagonal elements) depend both on the magnitude of the scores assigned, and on the number of respondents that positively rated any pair of adjectives. As a result, two adjectives

co-occur in the definition of the concept when they have been jointly chosen by a large part of respondents.

The scaled adjacency matrix is here denoted as $(\tilde{\mathbf{W}})$, as such it represents a Graph $\mathcal{G}(A, E)$ where A is the adjectives node-set of size n and E is the edge-set. This type of visualization enlightens relationships between adjectives. The considered relationship is given by the common choice of two adjectives so that the graph network is an image of the emerging definition of the concept-stimulus.

The matrix $(\tilde{\mathbf{W}})$ will be analyzed in the framework of Network Analysis. Since this adjacency matrix, by construction, likely tends to form a complete graph, an important step of our analysis concerns the choice of a cutting threshold allowing to highlight an emerging *core network*, it is defined as the cohesive set of adjectives in the giant component of the network. Indeed, the threshold value influences both the network density and the number of graph components. A simulation study is designed to assess the trade-off between the network density and the number of components that emerge when destroying edges.

In order to show and discuss the proposed method also on empirical data, we will present a case study in the field of road safety. The aim of this study is to assess young people attitude in road safety by using a concept-stimulus indirectly related to the topic analyzed, but close to their own life-style, that is the *Saturday Evening*. The units of our analysis are 283 students enrolled in the final year of high school, with age in the range 17 – 21 years-old, novice or going to obtain a driving license.

The use of some typical tools of social network analysis applied to the derived network data facilitate the emerging of the main signifiers (e.g. core-periphery structures, size and number of cliques, etc.) together with a more intuitive interpretation of the semantic meaning induced by the stimulus.

References

1. Greenacre M., Hastie T.: The Geometric Interpretation of Correspondence Analysis. Journal of the American Statistical Association, Vol. **82**, Issue **398** (1987)
2. Hanneman, R.A., Riddle, M.: Introduction to social network methods. University of California. Riverside (2005)
3. Kruskal, J.B., Wish, M.: Multidimensional Scaling. Sage Publications. Beverly Hills (1978)
4. Lin, C.: Semantic Network for Vocabulary Teaching. Journal of Taiwan Normal University: Humanities & Social Science, Vol. **42** (New Version), 43-54 (1997)
5. Osgood, C.E., Suci, G., Tannenbaum, P.: The measurement of meaning. University of Illinois Press. Urbana, IL (1957)
6. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)

Financial risk data analysis

Paolo Giudici

Department of Economics and Management

University of Pavia

giudici@unipv.it

August 1, 2013

1 Introduction

The guidelines proposed by the Basel Committee on Banking Supervision for the banking sector (www.bis.org) encourage financial institutions to use mathematical and statistical approaches for the computation of capital charges covering operational risks. On the other hand, non financial institutions are motivated to measure operational risk by the need of having under control the quality of their operational processes. In this context, operational risk measurement could be seen as a predictive tool, aimed at prioritising interventions and triggering actions, aimed at improving the quality of operations and, therefore, of products and services of a given institution, be it private or public.

Operational risks are usually classified in event types, according to the type of risk involved; and in business lines, according to the activities of an institution that are mostly affected by the risk events. To measure the operational risk, for each business line and for a given event type, the most employed approach is the actuarial model, based on quantitative loss data (see e.g. Cruz, 2002, Alexander 2003, Cornalba and Giudici, 2004, Dalla Valle and Giudici, 2008, Figini et al., 2012). Such an approach suggests to estimate the severity and the frequency probability distributions of the random losses associated with each risk, and to integrate them, obtaining the operational loss distribution, through a Monte Carlo simulation. The main output of the actuarial model is then the extraction of functionals of interest from the loss distribution, such as the Value at Risk and the Expected Shortfall. The calculation of such functionals is

required by Basel regulations to quantify the equity capital required to protect a financial institution (and, therefore, its stakeholders: shareholders, bondholders, depositors and, ultimately, taxpayers) against possible "unexpected" losses.

However, when data are available not only at a quantitative level, but also in an ordinal scale, as it is common for non financial companies, a pure quantitative approach is not possible. In this talk we show how operational risk measurement is possible also in this case.

The main contributions of this talk are: a) to introduce novel measures of operational risk, aimed at solving modelling problems in complex settings, arising from the presence of ordinal or heterogeneous data sources; b) to show how these measures can be effectively employed in two rather different contexts: risk management of a telecommunication company; quality assessment of academic research output. For lack of space, in this written contribution we shall focus on the methodological part, leaving applications to the actual presentation. Some of these applications are described in length in Cerchiello et al. (2010), for the measurement of teaching quality; in Cerchiello and Giudici (2013), for the measurement of research quality; in Figini and Giudici (2013) for the measurement of telecommunication risks. Here we shall focus on the modelling of ordinal risk data, as in Figini and Giudici (2013); methodologies for the integration of ordinal data with quantitative data are illustrated in Figini and Giudici (2011, 2012) and Cerchiello and Giudici (2012).

2 Proposal

In this section we present a general framework to measure risks, on the basis of ordinal variables. The framework will be presented, without loss of generality, with reference to operational risks.

Operational data for risk measurement are typically summarised in a matrix composed of I event types (the columns of the matrix) and J business lines (the rows of the matrix).

Let E_{ij} be a risk event, in the i -th event type ($i = 1, \dots, I$) and in the j -th business line ($j = 1, \dots, J$). For each combination of event type and business line, we have two different measures of risk: the frequency (how many risk events have appeared in that combination) and the severity (the mean loss of the events in that combination).

In the Basel framework for financial institutions, the severity is a continuous random variable. In the context of non financial companies, instead, the severity is generally expressed in an ordinal scale, characterised by S distinct levels ordered according to the corresponding magnitude (for example $S=3$, with H=high severity; M=medium severity and L=low severity). In this context, in order to summarise the frequency and the severity in a location measure, we may structure a loss contingency table, which counts, for each event type - business line and a given severity level, the absolute frequency.

More formally, let n_{H11} be the number of times for which the first event type in the first business lines appears with high severity; n_{M11} the number of times for which the first event type appears in the first business line with medium severity and n_{L11} the number of times for which the first event appears in the first business line with low severity. In general, let n_{Hij} , n_{Mij} and n_{Lij} be the number of times for which high, medium or low severity occur for the event type $i = 1, \dots, I$ in the $j = 1, \dots, J$ business line.

The above counts can be represented in a contingency table composed of J rows, representing the business lines (BL_1, \dots, BL_J) and $I \times S$ columns, equivalent to the number of event types multiplied by the levels of severity S under analysis (in our running example three: high=H, medium=M and low=L).

The location measure we propose is based on the cumulative distribution function of each cell variable in the loss contingency table. The latter can be calculated from each relative frequency $p_{hij} = n_{hij}/N$. Let F_{hij} be the cumulative distribution function calculated in each cell of the loss contingency table. On the basis of data structured as before, we propose to estimate the operational risk in each business line / event type combination through a Stochastic Dominance Index (SDI) measure which is defined as follows:

$$SDI_{ij} = \sum_{h=1}^S \frac{F_{hij}}{S} = \sum_{h=1}^S \sum_{l=1}^h p_{lij}, \quad (1)$$

for $i = 1, \dots, I$, $j = 1, \dots, J$ and where F_{hij} for $h = 1, \dots, K$ are the cumulative frequencies of each event type / business line combination.

Note that SDI_{ij} is a novel proposal, in the context of risk assessment; it has been used in the context of quality assessment (see Cerchiello et al., 2010, and the references therein), where it has proved a good approach to measure quality on the basis of an ordinal scale. Note that $SDI_{ij} = 0$ when the risk event never appears; and $SDI_{ij} = 1$ when the risk event is concentrated only on values with highest severity.

Furthermore, it is easy to show that SDI_{ij} is bounded between 0 and 1. On the basis of the well known properties of the cumulative distribution function, the proposed index could be employed to compare events of interest and business lines, producing an ordering among risks. These results may be very useful for the data owner, to prioritise interventions, also in terms of improvement of the related operational controls.

Furthermore, since our index is based on the cumulative distribution function, we are able to introduce an ordering criteria through the stochastic dominance approaches for ordinal variables, thus generating an order of preference on the set of $I \times J$ distribution functions involved in the analysis (see e.g. Shaked et al., 1994).

Stochastic dominance approaches are a good solution to compare the ordinal loss distributions corresponding to different event type/business lines combinations. In this way the decision maker can compare operational risks and prioritise interventions. We also remark that comparing loss distributions in terms of

stochastic dominance is a good approach to take into account the whole distributions and not only particular location measures or quantiles, as done in the standard actuarial approach.

A further practical advantage of the introduced measure is that it can be used to combine risk measures in a simple way that, besides, may preserve stochastic dominance.

For example, in order to aggregate event type risks over different business lines, we can use the following equation, that expresses the overall business line risk as a geometric mean of the measures of risk associated with each event type in that business line:

$$SDI_j = \left(\prod_{i=1}^I SDI_{ij} \right)^{1/I}. \quad (2)$$

Similarly, in order to aggregate business line over different event type risks, we can use the following equation, that expresses the overall event type risk as a geometric mean of the measure of risk associated with each business line in that event type risk:

$$SDI_i = \left(\prod_{j=1}^J SDI_{ij} \right)^{1/J}. \quad (3)$$

The above expressions show that risks over different units may be assumed to interact in a multiplicative way, being the units of an integrated system. Jean (1980) has shown that the geometric mean is a necessary condition to preserve stochastic dominance ranking when aggregating distribution functions. This because the geometric mean can be expressed as an arithmetic mean of logarithms and, since the logarithmic function is monotone, the corresponding geometric means are ordered. We remark that it is not true, in general, that the geometric mean is a sufficient condition to preserve stochastic dominance ranking.

The geometric mean has also a number of practical motivations. For example, many risk management problems require to combine multiplicatively the different components of risk; in operational risk the distribution of a financial loss is obtained multiplying the frequency distribution with the severity distribution.

The relationship between the geometric mean and the stochastic dominance framework is central to provide a simple and mathematically coherent approach to the construction of effective operational risk measures. Specifically, it can be demonstrated that the logarithm of the geometric mean is a coherent measure of risk (see e.g. Artzner et al. 1999).

So far we have discussed a methodology to derive a point estimate of operational risk, through the SDI measure. As in risk management we are typically interested in percentile values, it is of interest to derive confidence interval risk measures for ordinal variables. One possible approach is to derive a Bayesian

distribution for the SDI measures assuming that the $I \times J$ observed frequencies p_{hij} , $h = 1, \dots, S$ independently follow a Multinomial Distribution with prior parameter $\theta_{ij} = \theta_{hij}$, $h = 1, \dots, S$ and that θ_{ij} follows a Dirichlet prior distribution with prior parameters $\alpha_{ij} = \alpha_{hji}$, $h = 1, \dots, S$. From standard prior to posterior Bayesian analysis, we know that the posterior distribution of θ_{ij} is a Dirichlet with parameters $\alpha_{ij}^* = \alpha_{hji} + n_{hij}$ $h = 1, \dots, S$, from which confidence intervals on SDI can be derived approximately by means of Markov Chain Monte Carlo simulations.

An important final remark that operational risk modelling concepts and methods can be fruitfully applied to many quality assessment contexts, including those that may seem far from it. A noticeable example is the measurement of research quality of scientists. Hirsch (2005) proposed a simple method to assess and order such quality: the H-index. The Hirsch definition is that "a scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each".

In a recent contribution (Cerchiello and Giudici, 2013) that will be discussed in the talk we shall show how the (random) research impact of a scientist can be assimilated to a (random) operational loss, with the number of produced papers and their individual impact modelled in analogy with the frequency and severity arising in operational risk modelling. In this way a stochastic-based H-index can be derived, and inferential measurements, such as confidence intervals, attached to it. To achieve this aim we shall employ extreme value ordinal distributions, such as the Zipf-Mandelbrot distribution.

3 Acknowledgments

We thank the European VI research programme Project MUSING: Multy industry semantic based business intelligence (2006-2010), and the Italian Ministry of Research project MISURA: Multivariate Statistical models for risk assesment (2013-2015) , for financial and scientific support.

4 References

1. Alexander C. (2003). Operational Risk: Regulation, Analysis and Management, Prentice Hall.
2. Artzner, P., Delbaen, J. and Heat, D. (1999): Coherent measures of risk, *Mathematical Finance*, vol. 9, pp. 203-228.
3. Cerchiello, P., Dequarti E., Giudici, P., Magni C. (2010): Scorecard models to evaluate perceived quality of academic teaching. *Statistica e Applicazioni*, vol. 2 pp. 145-156.

4. Cerchiello, P., Giudici, P. (2012) Bayesian credit rating assessment. To appear in *Communications in Statistics: theory and methods*.
5. Cerchiello, P., Giudici, P. (2013) H-index: a statistical approach. Submitted.
6. Cornalba, C., Giudici, P. (2004) Statistical models for operational risk management. *Physica A: Statistical Mechanics and its applications*, 338, pp.166-172.
7. Cruz, M. (2002): *Modeling, Measuring and Hedging Operational Risk*, Wiley.
8. Dalla Valle, L., Giudici, P. (2008). A Bayesian approach to estimate the Marginal loss distributions in Operational Risk management. *Computational Statistics and data analysis*, 52, 3107-3127
9. Figini, S, Gao, L., Giudici, P: (2012) Bayesian efficient estimation of capital at risk. Submitted.
10. Figini, S. Giudici, P.(2011) Statistical merging of rating models, *Journal of the operational research society*, vol- 62, pp. 1067-1074.
11. Figini, S., Giudici, P. (2012) Credit risk predictions with model averaging. Submitted.
12. Figini,S., Giudici, P. (2013) A risk measure for ordinal variables. *The journal of operational risk*,vol.8, n.2.
13. Hirsch, J. E., 2005, An index to quantify an individual's scientific research output: *Proceedings of the National Academy of Sciences of the United States of America*, v. 102, p. 16569-16572.
14. Jean, W.H. (1980): The geometric mean and stochastic dominance, *Journal of Finance*, vol. 35, pp. 151-158.
15. Shaked, M. and Shanthikumar, G.J. (1994): *Stochastic Orders and Their Applications*, Academic Press, Boston.

A Comparison between SEM and Rasch model: the polytomous case

Silvia Golia and Anna Simonetto

Abstract The aim of the paper is to apply Rating Scale Model and Structural Equation Model to the same polytomous data in order to highlight the differences and similarities between the two models, developing a simulation study involving different situations. Moreover, we present a real case regarding the analysis of the quality of work.

Key words: Rating Scale Model, SEM, job satisfaction, procedural fairness

1 Introduction

There are two main approaches in the analysis of ordinal data coming from a survey: the Underlying Variable Approach (UVA), which assumes that the observed categorical outcomes are incomplete observations of unobserved continuous variables, and the Item Response Theory (IRT), which does not imply any assumptions regarding the nature of the observed variables [3]. Two of the most widespread models belonging to UVA and IRT are Structural Equation Model (SEM) and Rasch Model (RM). The comparison between SEM and IRT models is not new in the literature, but it is usually performed by comparing SEM with multidimensional IRT models, such as Random Coefficient Multinomial Logit Model [6]. In this work, we decide to apply the RM, which is a unidimensional model, because it produces objective measures of the latent trait. Our interest is thus focused on the comparison between RM and SEM, with particular attention to the estimated measures. We will develop a simulation study involving different situations and present a real case regarding the analysis of the quality of work. Given that our application context refers to polytomous data, as reference model belonging to the family of RM, we will use the

Silvia Golia, Anna Simonetto

Department of Economics and Management, University of Brescia, C.da S.Chiera 50, 25122 Brescia, Italy, e-mail: golia@eco.unibs.it; simonett@eco.unibs.it

Rating Scale Model (RSM) [1]. This family of measurement models converts raw scores into linear and reproducible measurement and its distinguishing characteristics are: separable person and item parameters, sufficient statistics for the parameters and conjoint additivity; prerequisites are unidimensionality and local independence. If the data fit the model, then the obtained measures are objective. The RSM theorizes that the log-odds ratio of two adjacent categories $k - 1$ and k is given by the difference between the ability of person s (θ_s), the difficulty of item i (δ_i) and the threshold j (τ_k): $\ln \left[\frac{P_{sik}}{P_{si(k-1)}} \right] = \theta_s - \delta_i - \tau_k$. Next to the RSM, we also investigate SEM. It is a family of models that had a wide spread in contexts very different between them. We will focus on reflective models based on the covariance matrix (Muthén's approach, [4]). We model the direct effects of a construct to its measure. The measurement model refers to the relationship between latent variables (θ) and their indicators: $Y = \Lambda \theta + \varepsilon$. The Structural Model is the set of latent variables in the model, together with the direct effects connecting them: $\theta = B\theta + \zeta$. B is the coefficient matrix for the effects (β) of constructs on each other. We allow latent variables to have simultaneous effects on each other (so that the B matrix has nonzero elements both above and below the diagonal). The following section reports the results of comparative study between RSM and SEM and the application to the real case.

2 Results

The first part of this section is devoted to the results of the simulation study. We will compare the findings and the measures obtained applying both RSM and SEM to a set of databases simulated according to different situations of interest. The second part of the section will take into account a real data set related to two sections of the survey on the quality of work, carried out in 2013 at the municipal district of Brescia (Italy), and useful to measure the worker Job Satisfaction (JS) and the Perceived Procedural Fairness (PPF). The data used into the simulation study are generated as follows. A sample of 1000 abilities θ is drawn from a standardized normal distribution, and they represent the true abilities. Two separate sets of 15 and 10 difficulty parameters δ are drawn from a continuous uniform distribution on the interval from -1.9 to 1.9 and then transformed so that the sum of the parameters is equal to zero. We consider two sets of threshold parameters τ , [-1, -0.5, 0, 0.5, 1] and [-1.2, -0.6, 0, 0.6, 1.2], one for each of the two different sets of δ . They imply six response categories. The response given by the subject s , with ability θ_s , to the item i , which has difficulty δ_i , is obtained as follows. For each category, the corresponding response probability is computed making use of the RSM formula. Then, the response probability cumulative sum is calculated and it is compared with a random number (rn) drawn from a uniform distribution on the interval [0,1]. The response category, corresponding to the first value of the cumulative sum which is equal or larger than rn , is assigned as the response of the subject s to the item i . This procedure is repeated for all the items in the test in order to simulate the response record of each of the 1000 subjects forming the simulated sample. In order to induce a second dimension

in the data, we define a group of items that is based onto a different level of ability. Following Smith [5], this different ability θ_a is drawn from a normal random variable and it depends on θ and on a second standardized normal random variable Z , independent of θ , according to the relation $\theta_a = (a \cdot \theta + \sqrt{1-a^2} \cdot Z) \cdot S_a + M_a$, where M_a and S_a are respectively the mean and the standard deviation of the θ_a distribution. In our simulations these, these two parameters will be set equal to 0 and 1, respectively. The θ_a is correlated with θ at the specific level a , with $a \in (0, 1)$. The responses to the items related to second dimensions are simulated as previously described. The Rasch analysis was performed using the Winsteps 3.72.3 computer program, whereas for the SEM analysis we used Mplus 5.21.

The simulation design considers four situations of interest and, for each of them, 100 data sets were simulated and analyzed using both RSM and SEM.

In the first case, the simulated databases come from an unidimensional 15-item questionnaire. Both RSM and SEM identify this unique dimension and the comparison between the two models has shown that the two methods build equivalent standardized measures: the mean correlation between the same abilities estimated by RSM and SEM is equal to 0.996 (sd. 0.0007). For all the data sets, the null hypothesis underlying the two-sample Kolmogorov-Smirnov test, that is $\hat{\theta}_{RSM}$ and $\hat{\theta}_{SEM}$ are drawn from the same underlying continuous population, is accepted. Moreover, both measures are coherent with the true ability (mean $Cor(\theta, \hat{\theta}_{RSM}) = 0.963$; mean $Cor(\theta, \hat{\theta}_{SEM}) = 0.962$).

The second condition allows the case of a 15-item questionnaire with two reference abilities: θ for the first 12 items and θ_a for the last 3 items. Two levels of a are considered: $a = 0.2$ and $a = 0.7$. In all cases, the RSM recognizes the presence of the extra dimension which must be deleted. SEM identifies the same structures with two dimensions, but it underestimates the correlation between θ and θ_a , probably owing to the well known effect of attenuation due to error of measurement.

In the third situation, we generate the 15-item questionnaire so that the first ten items have reference ability $ra = \theta$, three items belong to a second dimension with $ra = \theta_{0.6}$, the 14-th item has $ra = \theta_{0.1}$, and the last item has $ra = \theta_{0.2}$. Rasch analysis identifies an unwanted dimension composed by the three items with reference ability $\theta_{0.6}$ and two misfitting items, which are the ones with reference abilities $\theta_{0.1}$ and $\theta_{0.2}$ [2]. Applying an Exploratory Factor Analysis (EFA) with SEM, we find that data present three underlying latent constructs. The first 10 items refer to the first latent variable, the following three items to the second latent constructs and the last two items do not present significant loadings for any dimension, therefore they represent a noise detected by the model through the third dimension. If we eliminate these noise-items, the model indicates the presence of only two dimensions, each of which measured by the correct set of items, as for the RSM.

In the last simulation condition, we take into account two separate databases, referred to two distinct abilities (θ measured with 15 items and θ_a measured with 10 items) and we fixed two levels of a : 0.2 and 0.7. Rasch analysis considers these two tests separately, producing two distinct measures, one for each dimension, with $Cor(\hat{\theta}, \hat{\theta}_{0.2}) = 0.170$ (sd. 0.0002) and $Cor(\hat{\theta}, \hat{\theta}_{0.7}) = 0.614$ (sd. 0.0001). SEM jointly estimates the two latent constructs. The EFA results confirm the presence

of the two latent dimensions, associating the items to the correct constructs. The estimated $Cor(\hat{\theta}, \hat{\theta}_{0,2})$ is 0.185 (sd. 0.015) and the estimated $Cor(\hat{\theta}, \hat{\theta}_{0,7})$ is 0.675 (sd. 0.011). Even in this case, the two models provide analogous results: the mean correlation between the measures estimated with RSM and SEM from the same data sets reaches 0.99.

In the last part of our work we apply the Rasch and SEM analysis to a real data set, coming from the survey on the quality of work attended by 398 workers of the municipal district of Brescia. The considered latent traits are the worker JS and PPF. The original questionnaires are composed by 11 and 8 items, respectively, on a 10 points Likert scale. The Rasch analysis is performed on the two tests separately. A preliminary analysis has identified disordered thresholds for both JS and PPF, so the categories have been conveniently merged together. Looking at the JS test, the items *Pay*, *Workplace Environment*, and *Working Hours* are misfitting and so they have been deleted from the analysis; therefore the JS measure is constructed using 8 items. For the PPF questionnaire, the items *Acquire New Technologies* and *Transparency of Actions towards Users* are misfitting and so deleted from the analysis; therefore the PPF measure is based on 6 items. Through the SEM analysis, we reach analogous results, without merging categories. The same items have been deleted as a result of the EFA or merging the goodness of fit and a qualitative interpretation of the items. As seen in the last simulation, the estimated correlation between the measured obtained with SEM and RSM are high: for JS $Cor(\hat{\theta}_{SEM}, \hat{\theta}_{RSM}) = 0.956$ and for PPF $Cor(\hat{\theta}_{SEM}, \hat{\theta}_{RSM}) = 0.98$. Moreover, the estimated correlation between $\hat{\theta}_{JS}$ and $\hat{\theta}_{PPF}$ is equal to 0.516 using RSM and to 0.613 using SEM. The reason for this difference can be explained by the fact that the SEM is designed to consider the measurement error associated with the observed variables. This allows to reduce the effect of attenuation, due to the presence of measurement error on the items, for which the estimate of the correlation between the two latent variables should be less biased with SEM compared to RSM.

References

1. Andrich, D.: A rating formulation for ordered response categories. *Psychometrika* **43**, 561–573 (1978)
2. Bond, T.G., Fox, C.M.: *Applying the Rasch Model. Fundamental measurement in the human sciences*, 2nd edition, Lawrence Erlbaum Associates, Mahwah NJ (2007)
3. Cagnone, S., Mignani, S., and Moustaki, I.: Latent variable models for ordinal data. In Monari, P., Bini, M., Piccolo, D., and Salmaso, L., editors, *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, Contributions to Statistics, 17–28. Physica-Verlag HD (2010)
4. Muthén, L. K., Muthén, B. O.: *Mplus user's guide* (6th ed.) [Computer software manual]. Los Angeles, CA: (1998-2010)
5. Smith, Jr.E.V.: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J. Appl. Meas.* **3**, 205–231 (2002)
6. Wolfe, E.W., Singh, K.: A comparison of structural equation and multidimensional Rasch modeling approaches to Confirmatory Factor Analysis. *J. Appl. Meas.* **12**, 212–221 (2011)

Some considerations on VCUB models

Anna Gottard

Abstract CUB models are useful for studying ordinal data in rating surveys. The response probabilities are modelled as a convex combination of a shifted Binomial and a discrete Uniform distribution, to capture respondent psychological behaviour of being only partially able to quantify his/her feeling toward a specific item. A recent extension of CUB models, called the VCUB model, replaces the Uniform component with a variety of distributions, allowing for a more flexible description of respondent uncertainty. The comparison of different options can be interesting also from a psychological point of view. This work studies the necessity of multiple comparison adjustments, whenever several specifications seem adequate.

Key words: Multiple hypothesis testing, non-nested hypotheses, VCUB models

1 Introduction

In studies on several frameworks, people are asked to express their judgement on items or services by means of ratings. Usually, ordinal responses are analysed in terms of transformations of cumulative probabilities of multinomial distributions (see Agresti, 2010 for an extended presentation). Alternatively, one can adopt a CUB model, as formalized by Piccolo (2003) and later extended in several directions (see, for example, Iannario and Piccolo, 2012). The main idea in CUB models is that, when choosing rating responses, an individual is driven by a psychological mechanism that combines in a single value both the personal feeling towards the item and an inherent uncertainty in the choice of this value. This mechanism is formalized in the CUB models by a non-standard mixture of two distributions. The first component of the mixture, that is usually called *Feeling*, is assumed to be related to the respondent quantification of his/her judgement. This component is

Anna Gottard
DISIA, University of Florence, e-mail: gottard@disia.unifi.it

suitably modelled as a shifted Binomial distribution, derived by shifting the ordinary Binomial distribution to have support in $\{1, 2, 3, \dots, m\}$. The second component of the mixture, called *Uncertainty*, accounts for the respondent's inability to precisely quantify his/her judgement. The discrete Uniform random variable is considered a benchmark for graduating the maximum level of indecision or vagueness.

A recent extension of CUB models (Gottard *et al.*, submitted) introduces a varying uncertainty component. There are several circumstances that may cause a varying degree of uncertainty: respondents can sometimes refrain from using extreme values, or, on the contrary, people can systematically avoid the smallest or the largest categories, being influenced by a general optimistic or pessimistic climate. Several studies have found that some respondents prefer to endorse the middle value to indicate indifference or a lack of caring, reluctance to reveal their personal opinion or unclear interpretation of the item. In these circumstances, the discrete Uniform distribution should be replaced by different specifications.

Denoting with R the ordinal response with support $\{1, 2, \dots, m\}$, for a given m , a VCUB model can be formalized as

$$P(R = r) = \pi b_r(\xi) + (1 - \pi) V_r, \quad r = 1, 2, \dots, m \quad (1)$$

where π is the mixing proportion, $b_r(\xi) = \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1}$ is the shifted Binomial distribution and V_r is a known discrete probability distribution, called *Varying Uncertainty component*, whose specification is chosen *a priori*. Gottard *et al.* (submitted) proposed several specifications for V_r , adequate for most of the common psychological behaviours that can arise in specific circumstances. Selecting an adequate specification for the varying uncertainty component is an important task from a statistical point of view and, at the same time, it can be a topic of interest in some behavioural or psychological studies.

Two VCUB models with different varying uncertainty components are nonnested, or, in other words, they belong to separate families of distributions. Gottard *et al.* (submitted) proposed to utilize, for comparison, the Vuong's test (Vuong, 1989), a statistical test for strictly nonnested models.

In this work, the performance of the Vuong's test for VCUB models is evaluated and some considerations are provided on problem of multiple comparisons.

2 Considerations on multiple comparisons adjustments for the Vuong's test

Choosing an appropriate statistical model is complex but crucially important task. In particular for a VCUB model, an inadequate specification of the varying uncertainty component leads to biased estimates, as shown in Gottard *et al.* (submitted).

The Vuong's test is based on the Kullback–Leiber Information criteria (Kullback and Leibler, 1951) and is aimed to compare two competing nonnested models, say \mathcal{M}_0 and \mathcal{M}_1 . Under H_0 the test assumes that the two models are equally close to

the true, unknown, model ($\mathcal{M}_1 \approx \mathcal{M}_2$). The alternative hypothesis is divided in two parts of size $\alpha/2$, one in favor of the first model (\mathcal{M}_1), whenever the Vuong's test statistic takes an unexpectedly large positive value (H_1^+), the other in favor of the second model (\mathcal{M}_2), in case of unexpectedly large negative values (H_1^-).

Whenever the number of possibly adequate V_r specifications is greater than two, Gottard *et al.* (submitted) remark that an adequate correction for p -values could be necessary, still preventing to increase the test tendency to be conservative, as shown in their simulation study.

To study the Vuong's test performance in choosing among several VCUB models, we conduct a Monte Carlo study on 4 different specifications for the V_r component, depicted in Figure 1. Call $\mathcal{M}_T, \mathcal{M}_U$ (the ordinary CUB model), \mathcal{M}_1 and \mathcal{M}_2 the corresponding models, that are nonnested whenever $\pi \neq 1$. For various sample sizes, 1 000 samples have been generated from \mathcal{M}_T , and the Vuong's test is performed of each models versus all the others. Table 1 reports the percentage of times in which a specific model was not discarded, that is H_0 or H_1^+ has been selected. As can be ascertain from simulations results, the introduction of a correction for multiple comparisons seems not relevant if the interest is in not rejecting a true model. Even with small sample sizes, the true model, \mathcal{M}_T , is rejected only the 0.2-0.3% of the times. The closest model to the true one, \mathcal{M}_1 , is the second one not rejected. Only for very large samples, the Vuong's test is able to distinguish between \mathcal{M}_T and \mathcal{M}_1 . As a matter of fact, a correction to improve the power of the test, particularly for small samples, should be more useful than the ordinary control for false discovery rate or family-wise error rate.

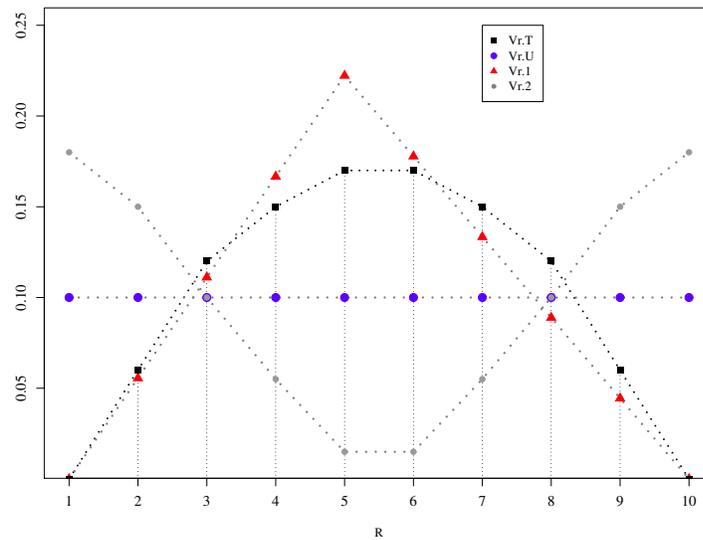


Fig. 1 Specifications of the V_r component in VCUB models adopted in the simulation scheme

Table 1 Percentage of model not rejection according to the Vuong's test ($\alpha = 0.05$)

n	50	100	1 000	5 000	50 000
	No correction				
\mathcal{M}_T	99.7	99.7	100.0	100.0	100.0
\mathcal{M}_U	74.7	38.7	0.0	0.0	0.0
\mathcal{M}_1	97.5	96.2	79.2	27.1	0.0
\mathcal{M}_2	76.5	37.9	0.0	0.0	0.0
	Holm (1979)'s correction for family-wise error rate				
\mathcal{M}_T	99.8	99.8	100.0	100.0	100.0
\mathcal{M}_U	87.1	62.9	0.0	0.0	0.0
\mathcal{M}_1	98.2	96.6	79.2	27.1	0.0
\mathcal{M}_2	87.6	63.1	0.0	0.0	0.0
	Hommel (1988)'s correction				
\mathcal{M}_T	99.8	99.7	100.0	100.0	100.0
\mathcal{M}_U	83.8	55.4	0.0	0.0	0.0
\mathcal{M}_1	98.0	96.4	79.2	27.1	0.0
\mathcal{M}_2	85.1	56.0	0.0	0.0	0.0
	Benjamini and Hochberg (1995)'s control of false discovery rate				
\mathcal{M}_T	99.8	99.7	100.0	100.0	100.0
\mathcal{M}_U	82.6	51.2	0.0	0.0	0.0
\mathcal{M}_1	97.0	96.4	79.2	27.1	0.0
\mathcal{M}_2	82.7	51.2	0.0	0.0	0.0

References

1. Agresti, A.: Analysis of ordinal categorical data. Wiley (2010)
2. Benjamini, Y., Hochberg, Y. : Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300 (1995)
3. Gottard, A., Iannario, M., Piccolo, D.: Varying Uncertainty in CUB Models, submitted.
4. Holm, S. : A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979)
5. Hommel, G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386 (1988)
6. Iannario, M., Piccolo, D.: CUB models: Statistical methods and empirical evidence. In: Kenett, R., Salini, S. (eds.), *Modern Analysis of Customer Surveys: with applications using R*, pp. 231–258. Wiley & Sons, (2012)
7. Kullback, S., Leibler, R.: On information and sufficiency. *The Annals of Mathematical Statistics*. **22**, 79–86 (1951)
8. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* **5**, 86–104 (2003)
9. Vuong, Q.: Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. **57**, 307–333 (1989)

Data driven EM constraints for mixtures of factor analyzers

Francesca Greselin and Salvatore Ingrassia

Abstract Mixtures of factor analyzers are becoming more and more popular in the area of model based clustering of high-dimensional data. In this paper we implement a data-driven methodology to maximize the likelihood function in a constrained parameter space, to overcome the well known issue of singularities and to reduce spurious maxima in the EM algorithm. Simulation results and applications to real data show that the problematic convergence of the EM, even more critical when dealing with factor analyzers, can be greatly improved.

Key words: Mixture of Factor Analyzers, Model-Based Clustering, Constrained EM algorithm.

1 Introduction and motivation

Finite mixture distributions, dating back to the seminal work of Newcomb and Pearson, have been receiving a growing interest in statistical modeling all along the last century. Along the lines of Ghahramani and Hilton (1997) we assume that the data have been generated by a linear factor model with latent variables modeled as Gaussian mixtures. Our purpose is to improve the performances of the EM algorithm, giving practical recipes to overcome some of its issues. Following Ingrassia (2004), in this paper we introduce and implement a procedure for the parameters estimation of mixtures of factor analyzers, which maximizes the likelihood function in a constrained parameter space, having no singularities and a reduced number of spurious local maxima. Within the Gaussian Mixture (GM) model-based approach to density estimation and clustering, the density of the d -dimensional random variable \mathbf{X} of interest is modeled as a mixture of a number, say G , of multivariate normal densities in some unknown proportions π_1, \dots, π_G ,

Francesca Greselin
Department of Statistics and Quantitative Methods
Milano-Bicocca University
Via Bicocca degli Arcimboldi 8 - 20126 Milano (Italy), e-mail: francesca.greselin@unimib.it

Salvatore Ingrassia
Department of Economics and Business
University of Catania
Corso Italia 55 - Catania (Italy), e-mail: s.ingrassia@unict.it

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{g=1}^G \pi_g \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

where $\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the d -variate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, we postulate a finite mixture of linear sub-models $\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \mathbf{e}_{ig}$ with probability π_g ($g = 1, \dots, G$) for $i = 1, \dots, n$, for the distribution of the full observation vector \mathbf{X} , given the (unobservable) latent factors \mathbf{U} , where $\boldsymbol{\Lambda}_g$ is a $d \times q$ matrix of *factor loadings*, the *factors* $\mathbf{U}_{1g}, \dots, \mathbf{U}_{ng}$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ distributed independently of the *errors* \mathbf{e}_{ig} , which are independently $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ distributed, and $\boldsymbol{\Psi}_g$ is a $d \times d$ diagonal matrix ($g = 1, \dots, G$). We suppose that $q < d$, which means that q latent factors are jointly explaining the d observable features of the statistical units. Under these assumptions, $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$ ($g = 1, \dots, G$). The parameter vector $\boldsymbol{\theta}_{MGFA}(d, q, G)$ now consists of the elements of the component means $\boldsymbol{\mu}_g$, the $\boldsymbol{\Lambda}_g$, and the $\boldsymbol{\Psi}_g$, along with the mixing proportions π_g ($g = 1, \dots, G-1$).

2 The likelihood function and the EM algorithm for MGFA

In this section we summarize the main steps of the EM algorithm for mixtures of Factor analyzers, see e.g. McLachlan *et al.* (2003) for details. Let \mathbf{x}_i ($i = 1, \dots, n$) denotes the realization of \mathbf{X}_i . Then, the complete-data likelihood function for a sample $\tilde{\mathbf{X}}$ of size n can be written as

$$L_c(\boldsymbol{\theta}; \tilde{\mathbf{X}}) = \prod_{i=1}^n \prod_{g=1}^G [\phi_d(\mathbf{x}_i | \mathbf{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) \phi_q(\mathbf{u}_i) \pi_g]^{z_{ig}}. \quad (1)$$

Due to the factor structure of the model, we consider the alternating expectation-conditional maximization (AECM) algorithm, consisting of the iteration of two conditional maximizations, until convergence. There is one E-step and one CM-step, alternatively i) considering $\boldsymbol{\theta}_1 = \{\pi_g, \boldsymbol{\mu}_g, g = 1, \dots, G\}$ where the missing data are the unobserved group labels $\mathbf{Z} = (\mathbf{z}'_1, \dots, \mathbf{z}'_n)$ and ii) considering $\boldsymbol{\theta}_2 = \{(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g), g = 1, \dots, G\}$ where the missing data are the group labels \mathbf{Z} and the unobserved latent factors $\mathbf{U} = (\mathbf{U}_{11}, \dots, \mathbf{U}_{nG})$. In the First Cycle, after updating the $z_{ig}^{(k+1)}$ in the E-step, the M-step provides new values for $\pi_g^{(k+1)}, \boldsymbol{\mu}_g^{(k+1)}, n_g^{(k+1)}$. In the Second Cycle, after writing the complete data log-likelihood, some algebras lead to the following estimate of $\{(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g), g = 1, \dots, G\}$

$$\hat{\boldsymbol{\Lambda}}_g = \mathbf{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{(k)'} [\boldsymbol{\Theta}_g^{(k)}]^{-1} \quad \hat{\boldsymbol{\Psi}}_g = \text{diag} \left\{ \mathbf{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g^{(k)} \mathbf{S}_g^{(k+1)} \right\},$$

where

$$\mathbf{S}_g^{(k+1)} = (1/n_g^{(k+1)}) \sum_{i=1}^n z_{ig}^{(k+1)} (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'$$

$$\boldsymbol{\gamma}_g^{(k)} = \boldsymbol{\Lambda}_g^{(k)'} (\boldsymbol{\Lambda}_g^{(k)} \boldsymbol{\Lambda}_g^{(k)'} + \boldsymbol{\Psi}_g^{(k)})^{-1} \quad \text{and} \quad \boldsymbol{\Theta}_g^{(k)} = \mathbf{I}_q - \boldsymbol{\gamma}_g^{(k)} \boldsymbol{\Lambda}_g^{(k)} + \boldsymbol{\gamma}_g^{(k)} \mathbf{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{(k)'}$$

3 Likelihood maximization in constrained parametric spaces

Along the lines of Ingrassia (2004) let us consider the constrained parameter space

$$\Theta_c = \{(\pi_1, \dots, \pi_G, \mu_1, \dots, \mu_G, \Sigma_1, \dots, \Sigma_G) \in \mathbb{R}^{k[1+d+(d^2+d)/2]} : \\ \pi_g \geq 0, \pi_1 + \dots + \pi_G = 1, a \leq \lambda_{ig} \leq b, \quad g = 1, \dots, G \quad i = 1, \dots, d\}. \quad (2)$$

Applying the eigenvalue decomposition to the square $d \times d$ matrix $\Lambda_g \Lambda_g'$ we can find Γ_g and Δ_g such that $\Lambda_g \Lambda_g' = \Gamma_g \Delta_g \Gamma_g'$ where Γ_g is the orthonormal matrix whose rows are the eigenvectors of $\Lambda_g \Lambda_g'$ and $\Delta_g = \text{diag}(\delta_{1g}, \dots, \delta_{dg})$ is the sorted diagonal matrix of the eigenvalues of $\Lambda_g \Lambda_g'$, i.e. $\delta_{1g} \geq \dots \geq \delta_{qg} \geq 0$, and $\delta_{(q+1)g} = \dots = \delta_{dg} = 0$. Further, applying now the singular value decomposition to Λ_g , we get $\Lambda_g = \mathbf{U}_g \mathbf{D}_g \mathbf{V}_g'$. This yields $\Lambda_g \Lambda_g' = (\mathbf{U}_g \mathbf{D}_g \mathbf{V}_g') (\mathbf{V}_g \mathbf{D}_g' \mathbf{U}_g') = \mathbf{U}_g \mathbf{D}_g \mathbf{D}_g' \mathbf{U}_g'$ hence $\text{diag}(\delta_{1g}, \dots, \delta_{qg}) = \text{diag}(d_{1g}^2, \dots, d_{qg}^2)$. We can now modify the EM algorithm in such a way that the eigenvalues of the covariances Σ_g (for $g = 1, \dots, G$) are confined into suitable ranges. To this aim we exploit the following inequalities

$$\begin{aligned} \lambda_{\min}(\Lambda_g \Lambda_g' + \Psi_g) &\geq \lambda_{\min}(\Lambda_g \Lambda_g') + \lambda_{\min}(\Psi_g) \geq a \\ \lambda_{\max}(\Lambda_g \Lambda_g' + \Psi_g) &\leq \lambda_{\max}(\Lambda_g \Lambda_g') + \lambda_{\max}(\Psi_g) \leq b \end{aligned}$$

which enforce (2) when imposing the following constraints

$$d_{ig}^2 + \psi_{ig} \geq a \quad i = 1, \dots, d \quad (3)$$

$$d_{ig} \leq \sqrt{b - \psi_{ig}} \quad i = 1, \dots, q \quad (4)$$

$$\psi_{ig} \leq b \quad i = q + 1, \dots, d \quad (5)$$

for $g = 1, \dots, G$, where ψ_{ig} denotes the i -th diagonal entry of Ψ_g . In particular, we note that condition (3) reduces to $\psi_{ig} \geq a$ for $i = (q + 1), \dots, d$.

It is important to remark that the resulting EM algorithm is monotone, once the initial guess, say Σ_g^0 , satisfies the constraints. Further, as shown in the case of gaussian mixtures in Ingrassia and Rocci (2007), the maximization of the complete log-likelihood is guaranteed. On another note, a data driven method to gauge the bounds a and b is needed.

4 Numerical studies

A brief numerical study is presented here, to compare the performance of the constrained vs unconstrained EM algorithm. More simulations have been performed, also with real datasets are available in (see Greselin and Ingrassia, 2013). A sample of $N = 150$ data has been generated with weights $\alpha = (0.3, 0.4, 0.3)'$ with parameters such that $\max_{i,g} \lambda_i(\Sigma_g) = 4.18$. We run 100 times both the unconstrained and the constrained AECM algorithms (for different values of the constraints a, b) using a common random initial clusterings. The unconstrained algorithm attains the right solution in 24% of cases; summary statistics about the misclassification error, over

the 100 runs, are reported in Table 1. To compare how a and b influences the performance of the constrained EM, different pairs of values has been considered, and Table 2 shows the more interesting cases. Further results are reported in Figure 1, where the boxplots of the distribution of the misclassification errors show the poor performance of the unconstrained algorithm compared to its constrained version. For all values of the upper bound b , the third quartile of the misclassification error is steadily equal to 0. Indeed, for $b = 6, 10$ and 15 we had no misclassification error, while we observed very low and rare misclassification errors only for $b = 20$ and $b = 25$ (respectively 3 and 11 not null values, over 100 runs). Moreover, the robustness of the results with respect to the choice of the upper constraint is apparent. A data driven method to select the bounds can be derived from the observed results, by running the constrained EM for increasing values of the upper bound, till a decrease in the final likelihood. The value of b before the decrease, observed over a series of run, will be chosen as upper value for the constrained estimation.

Table 1 Summary statistics for the Misclassif Error over 100 runs of the unconstrained EM alg

min	Q_1	Q_2	Q_3	max
0%	17%	36%	45.3%	60%

Table 2 Percentage of convergence to the right maximum of the constrained EM algorithm for $a = 0.01$ and some values of the upper constraint b

$b : +\infty$	6	10	15	20	25
24%	100%	100%	100%	97%	89%

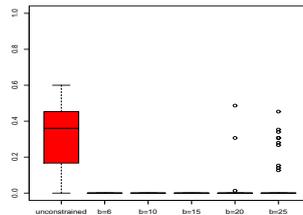


Fig. 1 Boxplots of the misclassification error. From left to right, boxplots refer to the unconstrained algorithm, then to the constrained algorithm, for $a = 0.01$ and $b = 6, 10, 15, 20, 25$.

References

- Ghahramani, Z. and Hilton, G. (1997). The EM algorithm for mixture of factor analyzers. *Tech. Rep. CRG-TR-96-1*.
- Greselin, F. and Ingrassia, S. (2013) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers, <http://arxiv.org/abs/1301.1505>.
- Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Stat. Meth. & Appl.*, **13**, 151–166.
- Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Comp. Stat. & Data Anal.*, **51**, 5339–5351.
- McLachlan, G., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Comp. Stat. and Data Anal.*, **41**, 379–388.

Predicting students' academic performance: a challenging issue in statistical modelling

Leonardo Grilli, Carla Rampichini and Roberta Varriale

Abstract We discuss some issues related to the statistical modelling of students' academic performance, with special emphasis on the role of predictors such as high school marks and pre-enrolment tests. After a brief review of the literature, we outline the strategies we devised in the analysis of data on the freshmen of the University of Florence.

Key words: binomial mixture model, concomitant variables, excess zeroes, hurdle model, regression chain graph

1 Introduction

Predicting students' academic performance is a key step in order to improve the efficiency of university systems. Indeed, delays or failures are costly for both the students and the administration. Therefore, it is of primary importance to determine the factors associated with the performance in order to plan actions such as guidance, restrictions to the access, and tutoring. To this end, universities can typically rely on information about the high school career, such as the type of school and various measures of proficiency. However, the results at high school are not fully appropriate to predict the academic performance due to several issues, including the possible mismatch between the competencies evaluated at high school and the com-

Leonardo Grilli
Department of Statistics, Informatics, Applications 'G. Parenti', University of Florence, e-mail: grilli@disia.unifi.it

Carla Rampichini
Department of Statistics, Informatics, Applications 'G. Parenti', University of Florence, e-mail: rampichini@disia.unifi.it

Roberta Varriale
Istat, Rome, e-mail: varriale@istat.it

petencies required for a given degree program, and the heterogeneity in the criteria for awarding marks (usually, there is substantial variability across types of schools and across geographical regions). To overcome those issues, some universities devise a pre-enrolment assessment test tailored on the needs of each degree program. However, a quick look at pre-enrolment tests around the world reveals a lack of commonly accepted guidelines and a shortage of empirical evidence about the predictive ability.

The literature on the empirical research about predicting students' academic performance is scattered in various journals, ranging from Psychology to Economics. Some noteworthy papers are Murray-Harvey (1993), Wedman (1994), Hoefler and Gould (2000), Murphy et al. (2001), Maree et al. (2003), Dancer and Fiebig (2004), Win and Miller (2005), Smith and Naylor (2005), Birch and Miller (2006), Birch and Miller (2007), Mills et al. (2009), Mallik and Lodewijks (2010), Bianconcini and Cagnone (2012), Chowdhury and Mallik (2012), Adelfio et al. (2013).

The statistical modelling of the academic career is challenging due to the complexity of the process. For example, the pre-enrolment test is an instrument to measure students' competencies in addition to already known characteristics, such as the high school mark, thus it is important to assess the value added by the test and to disentangle the effect of the high school mark on the academic performance into a direct effect and an indirect effect mediated by the test. To this end, the analyst has to rely on complex approaches such as path models (Murray-Harvey, 1993) or the regression chain graphs discussed in the following.

Another complication for the statistical modelling of gained credits is that the observed distribution is typically quite irregular: in fact, exams yield different number of credits and the sequence of exams varies across students; moreover, the distribution usually has peaks at zero and at the maximum. A simple approach such as OLS regression can still be used to analyze the associations, but it cannot be used to make predictions. To this end, a proper statistical model is required, even if the features of the response variable rule out conventional parametric models. Two effective methods are the quantile regression (Birch and Miller, 2006; Adelfio et al., 2013) and the mixture regression discussed in the following.

In the rest of the paper we focus on a case study about the pre-enrolment test at the University of Florence, illustrating some modelling strategies based on regression chain graphs, mixture models and hurdle models.

2 Data on freshmen at the University of Florence

In the academic year 2008/2009, the School of Economics of the University of Florence introduced a compulsory test to evaluate the background of the students wishing to enrol in a degree program. The test has 3 editions (September, November and December) and it is based on 40 multiple-choice items covering 3 areas: Logic (12 items, 30%), Reading (10 items, 25%) and Mathematics (18 items, 45%). For each item, one out of 5 alternatives is correct, with the following scoring system: 1 if

correct, 0 if blank, -0.25 if wrong. Thus the total score ranges from -10 to 40, and the threshold for passing the test is fixed at 9: candidates with a lower total score are advised against enrollment. In such a case, they can still enrol in a degree program of the School of Economics, but they can take examinations only after passing the test during one of the later editions.

We consider the participants to the first edition of the test (September 2008). The data set is obtained by merging data collected at the test with the administrative data of the School of Economics. After deleting 68 foreign students (due to missing information), the data set has 1057 observations. The available students' variables are listed in the following. *Pre-test*: Female, Far-away resident (indicator for residence in the provinces of Massa-Carrara and Grosseto or in a province out of Tuscany), Type of high school (Scientific, Humanities, Technical, Other), High school irregular career (indicator for age at high school diploma > 19), High school grade (from 60 to 100, centered at 80). *Test*: Total test score, Partial test scores (Logic, Reading, Mathematics), Test passed (indicator for total test score ≥ 9). *University career*: Delay in enrollment (indicator for being enrolled one or more years after high school diploma), Degree program (Management, Economics, Tourism, Marketing and Statistics), Credits gained during the first year (from 0 to 60), Second year enrollment at the School of Economics.

The test was passed by 853 candidates (80.7%). The test result is not mandatory for enrollment, but it influences the probability of enrollment: the enrollment rates were 65.3% overall, 67.9% for candidates who passed the test and 54.4% for candidates who did not pass the test.

The analysis is based on 690 students who took the test and then enrolled at the School of Economics. The sample distribution of gained credits after one year (December 2009) has a small percentage at the maximum (0.75% of freshmen gained 60 credits), but it has a peak at the minimum (23% of freshmen did not gain any credit). Therefore, the phenomenon is characterized by a relevant left censoring that needs to be accounted by the model. Moreover, the distribution of positive credits is quite irregular, showing peaks at 6, 15, 24, 36 and 45 credits. This pattern results from the paths followed by students, which can take exams weighting 6, 9 or 12 credits. The distribution of positive credits has a median of 30 and a mean of 29.8.

3 Modelling strategies

To disentangle direct and indirect effects of students background characteristics on the number of gained credits, the result of the admission test can be treated as an intermediate variable in a regression chain graph (Wermuth and Sadeghi, 2012). The specified chain graph model has three blocks: (i) pre-test (exogenous) variables, (ii) standardized test scores (intermediate variables), and (iii) gained credits after one year (outcome). The test result could be summarized by the total score, but this would obscure some interesting aspects of the phenomenon: first of all, the three areas (Logic, Reading and Math) have different numbers of items; moreover, we wish

to evaluate the relationships of each of the three partial scores with pre-test variables and the outcome. Therefore, we consider the Logic, Reading and Math scores as distinct variables, using standardized values to eliminate the effect of the different numbers of items. The three standardized partial scores are jointly regressed on pre-test covariates with a multivariate linear model (as compared to three separate regressions the multivariate regression yields the same point estimates but slightly different standard errors).

The model for regressing gained credits y_i on pre-test variables and test scores entails remarkable difficulties since, as noted in the previous Section, the distribution of y_i is quite irregular and has a large peak in zero. Therefore, conventional parametric models are not suitable. We tried two alternatives that we are going to outline in sequence, namely a binomial mixture model with concomitant variables (Grilli et al., 2013) and a hurdle model (Grilli et al., 2012)

3.1 Binomial mixture model with concomitant variables

The binomial mixture model is explicitly designed for a count variable with a fixed maximum, such as the number of gained credits y_i . This model assumes that the distribution $P(y_i)$ is defined by a finite mixture of conditional distributions $P(y_i | u_i)$, where u_i is a categorical latent variable taking values $k = 1, \dots, K$ with prior probabilities $\pi_k = P(u_i = k)$, where $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$.

$$P(y_i) = \sum_{k=1}^K \pi_k P(y_i | u_i = k). \quad (1)$$

where the conditional distributions $P(y_i | u_i)$ are binomial with common number of trials t and component-specific probabilities of success θ_k :

$$P(y_i | u_i = k) = \binom{t}{y_i} \theta_k^{y_i} (1 - \theta_k)^{t - y_i}. \quad (2)$$

In order to exploit the covariates, we fit a *Concomitant variable mixture model* (Dayton and Macready, 1988), where the component probabilities of the finite mixture vary across subjects according to a vector of covariates \mathbf{z}_i (usually including a constant for the intercept):

$$P(y_i | \mathbf{z}_i) = \sum_{k=1}^K \pi_{k|\mathbf{z}_i} P(y_i | u_i = k), \quad (3)$$

where $\pi_{k|\mathbf{z}_i} = P(u_i = k | \mathbf{z}_i)$, with $\pi_{k|\mathbf{z}_i} > 0$ and $\sum_{k=1}^K \pi_{k|\mathbf{z}_i} = 1$ for any subject i . Such constraints are satisfied by any model for nominal variables, like the multinomial logit model:

$$\pi_{k|z_i} = \frac{\exp(\mathbf{z}'_i \boldsymbol{\beta}_k)}{\sum_{l=1}^K \exp(\mathbf{z}'_i \boldsymbol{\beta}_l)}, \quad k = 1, \dots, K, \quad (4)$$

with $\boldsymbol{\beta}_1 = 0$ for model identifiability. Therefore, the prior probabilities of class membership depend on the covariates \mathbf{z}_i through a non-linear function.

For given K , the parameters can be estimated with Maximum Likelihood using the EM algorithm (McLachlan and Peel, 2000).

3.2 Hurdle model

A hurdle or two-part model (Cameron and Trivedi, 2005) can be used to account for the large proportion of students (23%) gaining no credits ($y_i = 0$). Such a proportion should not be regarded as a nuisance, but as a key feature of the phenomenon since those students failed to begin the university career and, indeed, most of them dropped out.

The hurdle model has two components: a logit model for the probability of gaining at least one credit $P(y_i > 0 | \mathbf{z}_i)$, and a linear model for the expected number of gained credits $E(y_i | y_i > 0, \mathbf{x}_i)$. The linear model is fitted on the subset of students who gained at least one credit. The covariates of the two sub-models, \mathbf{z}_i and \mathbf{x}_i , are distinct in principle, but they can even be the same. Since no parametric distribution appropriately describes the pattern of gained credits, we avoid a parametric specification and estimate the parameters via OLS and then compute robust standard errors.

The linear model for positive credits should be regarded as an approximation of the relationship between the mean and the covariates, without trying to model the whole distribution. Indeed, the linear model does not put restrictions on the support of y_i , so that non-integer values and out-of-range values are possible. However, in this application such issues are not critical, since non-integer values are just a problem of rounding, whereas the predicted mean is always within the range [0,60].

References

1. Adelfio, G., Boscaino, G., Capursi, V.: Quantile regression on a new indicator for higher education performance. Working Paper, CNR Solar (2013) <http://eprints.bice.rm.cnr.it/id/eprint/5181>
2. Bianconcini, S., Cagnone, S.: A General Multivariate Latent Growth Model With Applications to Student Achievement. *Journal of Educational and Behavioral Statistics* **37**, 339–364 (2012)
3. Birch, E.R., Miller, P.W.: Student Outcomes At University In Australia: A Quantile Regression Approach. *Australian Economic Papers*, Wiley Blackwell **45**, 1–17 (2006)
4. Birch, E.R., Miller, P.W.: The influence of type of high school attended on university performance. *Australian Economic Papers* **46**, 1-17 (2007)
5. Cameron, A.C., Trivedi, P.K.: *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge (2005)

6. Chowdhury, M., Mallik, G.: How Important are Introductory Subjects in Advanced Economics Studies? *Economic Papers: A journal of applied economics and policy* **31**, 255–264 (2012)
7. Dancer, D.M., Fiebig, D.G.: Modelling Students at Risk. *Australian Economic Papers* **43**, 158–173 (2004)
8. Dayton, C. M., Macready, G. B.: Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association* **83**, 173–178 (1988)
9. Grilli, L., Rampichini, C., Varriale, R.: University admission test and students' careers: an analysis through a regression chain graph with a hurdle model for the credits. 46th Scientific Meeting of the Italian Statistical Society. Rome, 20-22 June, (2012)
10. Grilli, L., Rampichini, C., Varriale, R.: Binomial mixture modelling of university credits. To appear in *Communications in Statistics - Theory and Methods* (2013)
11. Hoefer, P., Gould, J.: Assessment of Admission Criteria for Predicting Students' Academic Performance in Graduate Business Programs. *Journal of Education for Business* **75**, 225–229 (2000)
12. Mallik, G., Lodewijks, J.: Student Performance in a Large First Year Economics Subject: Which Variables are Significant? *Economic Papers: A journal of applied economics and policy* **29**, 80–86 (2010)
13. Maree, J.G., Pretorius, A., Eiselen, R.J.: Predicting success among first-year engineering students at the rand afrikaans university. *Psychological Reports* **93**, 399–409 (2003)
14. McLachlan, G., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
15. Mills, C., Heyworth, J., Rosenwax, L., Carr, S., Rosenberg, M.: Factors associated with the academic success of first year Health Science students *Advances in Health Science Education* **14**, 205–217 (2009)
16. Murphy, M., Papanicolaou, K., McDowell, R.: Entrance score and performance: A three year study of success. *Journal of Institutional Research* **10**, 32–49 (2001)
17. Murray-Harvey, R.: Identifying characteristics of successful tertiary students using path analysis. *Australian Educational Researcher* **20**, 63–81 (1993)
18. Smith, J., Naylor, R.: Schooling Effects on Subsequent University Performance: Evidence for the UK University Population'. *Economics of Education Review* **24**, 549–562 (2005)
19. Wedman, I.: The Swedish Scholastic Aptitude Test: Development, Use, and Research. *Educational Measurement: Issues and Practice* **13**, 5–11 (1994)
20. Vermunt, J. K., Magidson, J.: *LG-Syntax users guide: Manual for Latent GOLD 4.5 Syntax Module*. Statistical Innovations Inc., Belmont, MA (2008)
21. Wermuth, N., Sadeghi, K.: Sequences of regressions and their independences. *Test* **21**, 215–252 (2012)
22. Win, R., Miller, P.W.: The Effects of Individual and School Factors on University Students' Academic Performance. *Australian Economic Review* **38**, 1–18 (2005)

Robust estimation of regime switching models

Luigi Grossi, Fany Nan

Abstract When GM estimators are extended to threshold models, which are piecewise linear models, the consistency of GM estimators is guaranteed only under certain choices of the objective function. In this paper we explore, in a simulation experiment, the loss of consistency of GM-SETAR estimator under different objective functions, time series length, parameters combinations and type of contaminations. Finally the robust estimators are applied to study the dynamic of electricity prices.

Key words: GM estimator, nonlinear models, outliers

1 Introduction

Threshold Autoregressive (TAR) models are quite popular in the nonlinear time-series literature Tong (1990). In the class of non-linear models, studies addressed to robustifying this kind of models are very few, although the problem is very challenging particularly when it is not clear whether aberrant observations must be considered as outliers or as generated by a real non-linear process. Chan and Cheung (1994) extended the generalized M estimator method to SETAR models. Their simulation results show that the GM estimation is preferable to the LS estimation in presence of additive outliers. As GM estimators have proved to be consistent with a very small loss of efficiency, at least under normal assumptions, the extension to threshold models, which are piecewise linear, looks quite straightforward. Despite this observation, a cautionary note (Giordani; 2006) has been written to point out some drawbacks of GM estimator proposed by Chan and Cheung (1994). In particular, it is argued and shown, by means of a simulation study, that the GM estimator can deliver inconsistent estimates of the threshold even under regularity conditions. According to this contribution, the inconsistency of the estimates could be particularly severe when strongly descending weighting functions are used. Zhang et al. (2009) demonstrate the consistency of GM estimators of autoregressive parameters in each regime of SETAR models when the threshold is unknown. The consistency of parameters is guaranteed when the objective function is a convex non-negative function. A possible function holding these properties is the Huber ρ -function. However, the authors conclude, the problem of finding a threshold estimator with desirable finite-sample properties is still an open issue. Although, a theoretical proof has been provided, there is not a well structured Monte Carlo study to assess the extent of the distortion of the GM-SETAR estimators. In this paper we want to fill

Luigi Grossi, University of Verona, e-mail: luigi.grossi@univr.it
Fany Nan, University of Verona, e-mail: fany.nan@univr.it

this gap by presenting an extensive Monte Carlo study comparing LS and GM estimator under particular conditions. Moreover, we propose an application of robust nonlinear estimation to the series of electricity prices following the results of the simulation experiment. It is well known that among the stylized facts which empirically characterize electricity prices, the presence of sudden spikes is one of the most regularly observed and less explored feature (Janczura and Weron; 2010).

2 Robust SETAR models

Given a time series y_t , a two-regime Self-Exciting Threshold AutoRegressive model SETAR(p, d) is specified as $y_t = \mathbf{x}_t \beta_1 + \varepsilon_t$, if $y_{t-d} \leq \gamma$ and $y_t = \mathbf{x}_t \beta_2 + \varepsilon_t$, if $y_{t-d} > \gamma$ for $t = 1, \dots, N$, where y_{t-d} is the threshold variable with $d \geq 1$ and γ is the threshold value. The relation between y_{t-d} and γ states if y_t is observed in regime 1 or 2. β_j is the parameter vector for regime $j = 1, 2$ and \mathbf{x}_t is the t -th row of the $(N \times p)$ matrix \mathbf{X} comprising p lagged variables of y_t (and possibly a constant). Errors ε_t are assumed to follow an iid($0, \sigma_\varepsilon$) distribution. In general the value of the threshold γ is unknown, so that the parameter to estimate become $\theta = (\beta_1', \beta_2', \gamma, \sigma_\varepsilon)$. For a discussion on SETAR parameters estimation see Tong (1990).

In the case of robust two-regime SETAR model, for a fixed threshold γ the GM estimate of the autoregressive parameters can be obtained by applying the iterative weighted least squares: $\hat{\beta}_j^{(n+1)} = (\mathbf{X}_j' \mathbf{W}^{(n)} \mathbf{X}_j)^{-1} \mathbf{X}_j' \mathbf{W}^{(n)} \mathbf{y}_j$ where $\hat{\beta}_j^{(n+1)}$ is the GM estimate for the parameter vector in regime $j = 1, 2$ after the n -th iteration from an initial estimate $\hat{\beta}_j^{(0)}$, and $\mathbf{W}^{(n)}$ is a weight diagonal matrix, those elements depend on a weighting function $w(\hat{\beta}_j^{(n)}, \hat{\sigma}_{\varepsilon, j}^{(n)})$ bounded between 0 and 1. The threshold γ can be estimate by minimizing the objective function $\rho(r_t)$ over the set Γ of allowable threshold values. In the present paper we are going to analyze three robust estimators for SETAR models. The first method is described in Chan and Cheung (1994). For the second method, we follow Franses and van Dijk (2000). The GM weights are presented in Schweppe's form $w(\hat{\beta}_j, \hat{\sigma}_{\varepsilon, j}) = \psi(r_t)/r_t$ with standardized residuals $r_t = (y_t - \mathbf{x}_t \hat{\beta}_j) / (\hat{\sigma}_{\varepsilon, j} w_x(\mathbf{x}_t))$ and $w_x(\mathbf{x}_t) = \psi(d(\mathbf{x}_t)^\alpha) / d(\mathbf{x}_t)^\alpha$. $d(\mathbf{x}_t) = |\mathbf{x}_t - m_{y, j}| / \hat{\sigma}_{y, j}$ is the Mahalanobis distance and α is a constant usually set equal to 2 to obtain robustness of standard errors. The chosen weight function is the polynomial ψ function proposed in Lucas et al. (1996). The threshold γ is estimated by minimizing the objective function $\sum_{t=1}^N w(\hat{\beta}, \hat{\sigma}_\varepsilon)(y_t - \mathbf{x}_t \hat{\beta})^2$ over the set Γ of allowable threshold values. The third method is based on the same methodologies of the second but with ψ be the Huber weight function.

3 Simulation experiment and empirical application

To compare the performance of the three methods, we reproduced the simulation study of Chan and Cheung (1994). We generated time series from SETAR($1, d$) models for fixed sample sizes of $N = 100, 500$, with 1000 replications respectively, and

$\sigma_\varepsilon^2 = 1$. Moreover, 18 parameter combinations for $\theta = (\beta_1, \beta_2, \gamma, d)$ are considered. The series are contaminated following three schemes. For the single-outlier case, an additive outlier is located at $t = N/2$ with magnitude $\omega = 0, 3, 4, 5$. For the 3-outlier case, we fixed three outliers at $t = N/4, N/2$, and $N * 3/4$ with magnitude $-\omega, \omega, -\omega$ respectively. The multiple-outlier case is applied only for series with $N = 500$: three outliers are fixed every 100 observation with the same scheme of the 3-outlier case. For the first (called ‘‘Tukey’’) robust estimation method (Chan and Cheung; 1994), the starting values $\hat{\beta}_1^0, \hat{\beta}_2^0$ of the parameters are calculated by four iterations with Huber weights with OLS estimates as initial points. For the second (called ‘‘Polyn’’) and third (called ‘‘Huber’’) method, the starting values are calculated by least median squares.

Table 1 Ratios of the RMSE of the Robust estimates to the LS estimates. 1000 MC simulations of time series with sample size 500. Values of true parameters are in parentheses

d	ω	$\hat{\gamma}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
		$\omega = 0$	3	5	$\omega = 0$	3	5	$\omega = 0$	3	5
Tukey (0,-0.5,-1)	1	1.17	0.98	0.86	1.84	0.46	0.41	3.02	0.81	0.41
Tukey (0,-1,-0.5)	1	1.18	0.77	0.72	3.06	0.42	0.26	1.83	0.72	0.47
Tukey (0,0.5,0.8)	1	1.31	1.36	1.44	1.66	0.67	0.76	3.35	0.87	0.48
Tukey (0,-0.5,0.8)	1	2.92	2.28	1.86	3.87	1.79	1.15	3.40	1.09	0.62
Tukey (0,0.3,0.8)	2	4.68	1.58	1.23	2.59	1.33	0.86	2.73	0.77	0.44
Tukey (-0.1,0.3,-0.8)	2	18.26	10.29	5.75	3.77	2.16	1.48	2.99	0.88	0.73
Polyn (0,-0.5,-1)	1	0.99	0.76	0.69	1.12	0.38	0.17	1.12	0.36	0.16
Polyn (0,-1,-0.5)	1	0.96	0.62	0.57	1.14	0.21	0.10	1.11	0.64	0.28
Polyn (0,0.5,0.8)	1	0.92	1.02	1.04	1.14	0.41	0.22	1.18	0.42	0.19
Polyn (0,-0.5,0.8)	1	1.28	1.06	0.93	1.38	0.78	0.49	1.07	0.45	0.21
Polyn (0,0.3,0.8)	2	1.74	0.65	0.39	1.18	1.04	0.60	1.10	0.43	0.19
Polyn (-0.1,0.3,-0.8)	2	2.57	1.21	0.59	1.08	0.75	0.38	1.08	0.41	0.15
Huber (0,-0.5,-1)	1	0.98	0.72	0.66	1.07	0.49	0.21	1.01	0.37	0.17
Huber (0,-1,-0.5)	1	0.92	0.63	0.58	1.09	0.29	0.14	1.03	0.62	0.28
Huber (0,0.5,0.8)	1	1.10	1.02	1.00	1.15	0.46	0.23	1.13	0.50	0.20
Huber (0,-0.5,0.8)	1	1.31	1.15	1.11	1.28	0.78	0.61	1.06	0.48	0.24
Huber (0,0.3,0.8)	2	2.04	0.72	0.51	1.31	1.04	0.71	1.14	0.63	0.26
Huber (-0.1,0.3,-0.8)	2	1.58	1.03	0.59	1.09	0.80	0.52	1.11	0.49	0.22

In Table 1 we have summarized some results of the big Monte Carlo experiment. Each row corresponds to a combination of parameters. The values reported in the table represent the ratio between robust and LS RMSE: robust estimators are better than LS when the ratios are less than 1. For lack of space, we reported only 6 combinations out of the original 18, focusing on the sample size $T = 500$, the multiple outlier case with three outlier magnitudes. According to what it has been proved by Zhang et al. (2009) the robust estimator of the threshold parameter is very less efficient than the LS estimator in small samples. As a consequence, we found (results available upon request) that all three robust methods performed generally worse than the LS, at least for weak contamination patterns, that is in the single outlier case with small magnitude. From Table 1 is immediately clear that, while the method suggested by Chan and Cheung (1994) based on the Tukey function does not show any significant improvement with respect to LS, the other two methods look

to be competitive to LS in the estimation of the threshold. Moreover, the polynomial and Huber functions are far better than the LS estimator in estimating autoregressive coefficients with a slight prevalence of the polynomial method. These results confirm the theoretical results provided by Zhang et al. (2009).

We applied LS and the three robust methods to estimate parameters of a SETAR(2,1) model on Italian electricity price data (*PUN, prezzo unico nazionale*). We estimated the model on price returns of peak hour 18 from January 1st, 2009 to December 31st, 2011. Estimation results are shown in table 2.

Table 2 SETAR parameter estimates. PUN price returns, hour 18

Estimation method	Regime 1				Regime 2		
	$\hat{\gamma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$
LS	-0.075	-0.565	-0.367	0.039	-0.363	-0.199	0.023
Polynomial	0.001	-0.314	-0.191	0.033	-0.350	-0.192	0.023
Huber	-0.001	-0.336	-0.203	0.033	-0.327	-0.171	0.023
Tukey	-0.066	-0.76	-0.611	0.041	-0.896	-0.316	0.030

We have to emphasize that the analyzed time series is very similar to the trajectories simulated in the previous section: large sample size and high contamination level. In this case the Polynomial and Huber methods should perform better than both LS and Tukey. As a confirmation of the simulation experiment, Polynomial and Huber coefficients are very similar. The next step of the analysis will be to compare the forecasting performances of these two methods with the forecasting performance of the LS estimator.

References

- Chan, W. S. and Cheung, S. H. (1994). On robust estimation of threshold autoregressions, *Journal of Forecasting* **13**: 37–49.
- Franses, P. H. and van Dijk, D. (2000). *Non-linear time series models in empirical finance*, Cambridge University Press.
- Giordani, P. (2006). A cautionary note on outlier robust estimation of threshold models, *Journal of Forecasting* **25**(1): 37–47.
- Janczura, J. and Weron, R. (2010). An empirical comparison of alternate regime-switching models for electricity spot prices, *Energy Economics* **32**: 1059–1073.
- Lucas, A., van Dijk, R. and T, K. (1996). Outlier robust gmm estimation of leverage determinants in linear dynamic panel data models, *Discussion Paper 94-132*, Tinbergen Institute.
- Tong, H. (1990). *Non-linear Time Series: a Dynamical System Approach*, Oxford University Press, Oxford.
- Zhang, L. X., Chan, W. S., Cheung, S. H. and Hung, K. C. (2009). A note on the consistency of a robust estimator for threshold autoregressive processes, *Statistics and Probability Letters* **79**: 807–813.

Variable selection in sequential multi-block analysis

Kristian Hovde Liland

Nofima, Norway

Abstract

In sequential multi-block analysis, e.g. sequential and orthogonalised PLS (SO-PLS), data from different sources, instruments, time points, etc. are included one block at the time. This results in different analyses depending on the order of the blocks. When working with multivariate data it is sometimes beneficial to reduce the number of variables to be analysed. A vast amount of methods for variable selection strategies exist together with a range of different motivations for applying them. Introducing variable selection into the regime of sequential multi-block analysis means one has to make several assumptions, choices and limitations to avoid drowning in models and computational demand. In this work we will explore some of the possibilities applying powered PLS and selectivity ratios for multivariate variable selection on real multi-block data from Raman and NIR spectroscopy.

Robustness issues for a class of models for ordinal data

Maria Iannario

Abstract In this contribution we discuss robustness analysis of ordinal data by means of a statistical model introduced for interpreting discrete distribution derived by the psychological process of selection. Specifically, we use the influence function and analyse the contamination of an *atypical* situation on both estimators in the parametric space of the model.

Key words: CUB models, Influence function, Robustness

1 Introduction

Only few papers concern robustness analysis of ordinal data [1, 7] because of the complex structure concerning both the limited range of the data and the discreteness of the support. Generally, it has been considered the continuous latent variable surrounding the categorical choice. In this contribution we consider ordinal data as realizations of discrete distribution and motivate some robustness issues for a class of ordinal data models introduced by [6] for the analysis of ratings. He considers the scores expressed on evaluation contexts as realizations of a mixture of discrete random variables concerning the psychological process of selection.

In this case, the robustness is related to the modification induced by atypical responses on the probability distribution.

This choice could depend from many factors: low education of respondents who misunderstand the question and choose a wrong category producing a contamination of the true distribution; a *shelter effect* [2] related to a single category where the probability mass is concentrated; a *response contraction bias* which arises from respondents which refrain from using extreme values of the scale when more scale

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Naples, e-mail:
maria.iannario@unina.it

points are provided or by *scale usage heterogeneity* when respondents adopt a personal scale, shorter than the proposed one. Some possible “bumps” in the frequency distributions derived from these uncommon responses can affect the reliability of the estimators.

The focus of this work is to measure the impact on the estimator produced when the frequency of a category has been modified with respect to the expected distribution.

The paper is organized as follows: in the next Section we examine the infinitesimal contamination at a single category by means of the *influence function* with some approximation of the sensitivity curve. A discussion and some concluding remarks end the paper.

2 Influence function

When we consider categorical data model, a contamination can be produced by any discrete distribution function G with the same support $\{1, \dots, m\}$ of the assumed model. Specifically, let F be the distribution function of the starting model; the contaminated model has the distribution function

$$H = (1 - \varepsilon)F + \varepsilon, G$$

where ε is the mass of contamination. Under H a proportion ε of the data are generated by G . The probability mass function under the contaminated model is given by

$$p_r^H = (1 - \varepsilon) p_r(\boldsymbol{\theta}) + \varepsilon g_r, \quad r = 1, \dots, m,$$

where g_r is the probability of the r -th category under G .

The main interest focuses on the asymptotic value of the estimator under contamination [5] which is the solution $\boldsymbol{\theta}(H)$ of

$$E_H [S\{R, \boldsymbol{\theta}(H)\}] = \sum_{r=1}^m S\{r; m\boldsymbol{\theta}(H)\} p_r^H = 0. \quad (1)$$

We can consider different levels of contamination. The simplest one is related to the consideration of a point mass placed on the r^* -th category: there is a fraction ε of outliers located at r^* , i.e the ratings which do not belong to the model are concentrated on r^* . In this case the contaminated model is

$$H = (1 - \varepsilon)F + \varepsilon \delta_{r^*},$$

where δ_{r^*} is a probability measure which puts mass 1 on the r^* -th category. By (1) the asymptotic value

$$\theta_\varepsilon^{r^*} = \theta((1 - \varepsilon)F + \varepsilon\delta_{r^*})$$

of the estimator under the contaminated model is the solution of

$$\sum_{r=1}^m S(r; \theta_\varepsilon^{r^*}) \{(1 - \varepsilon)p_r(\theta_0) + \varepsilon I(r = r^*)\} = 0,$$

where $I(\omega)$ takes value 1 if ω holds and 0 otherwise.

For describing the effect of an infinitesimal contamination on the asymptotic value of the estimator we consider the influence function:

$$IF(r^*, F) = \lim_{\varepsilon \rightarrow 0} \frac{\theta\{(1 - \varepsilon)F + \varepsilon\delta_{r^*}\} - \theta(F)}{\varepsilon}.$$

If we consider the class of CUB models [3], characterized by the following distribution:

$$p_r(\theta) = \pi b_r(\xi) + (1 - \pi)U_r, \quad r = 1, 2, \dots, m$$

where $\theta = (\pi, \xi)'$, $b_r(\xi) = \binom{m-1}{r-1} \xi^{m-r} (1 - \xi)^{r-1}$ is a shifted Binomial distribution and U_r is a Uniform discrete probability distribution, and implement the influence function for the fitted CUB distribution \hat{F} [4] we can illustrate the effect of contamination on the parameter space

$$\Omega(\theta) = \{(\pi, \xi) : 0 < \pi \leq 1 \quad 0 \leq \xi \leq 1\}.$$

Specifically, it is possible to consider several aspects concerning the pattern of influence functions and by means of the *inspection plots* introduced in [4]. In these representations is possible to notice that although a CUB random variable can only take finite values, by varying the parameters concerning the feeling and the uncertainty, the influence function can be very large when the parameters are close to the boundary of the parameter space. Such behaviour mimics the unbounded influence functions in models for continuous random variables.

Moreover, the influence function can be interpreted as the limit of the sensitivity curve $SC(r^*)$ when the sample size diverges. This curve measures the change in the estimates produced, in a sample (r_1, r_r, r_n) of size n , by an additional observation on the r^* -th category, standardized by the amount of contamination. Formally,

$$SC(r^*) = \frac{\hat{\theta}_{n+1}(r, r_{n+1} = r^*) - \hat{\theta}_n(r)}{\frac{1}{n+1}}.$$

Consequently,

$$\{1/(n+1)\} IF(r^*, F)$$

approximates the changes in the value of the estimators due to an additional observation $r_{n+1} = r^*$.

Let $I(\theta)$ be the information matrix, whose generic element is

$$I_{ij}(\boldsymbol{\theta}) = E \left[\left\{ \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_i} \right\} \left\{ \frac{\partial \ln(p(\boldsymbol{\theta}))}{\partial \theta_j} \right\} \right], \quad (i, j) = 1, 2.$$

By applying standard results on M -estimators, whose Maximum Likelihood estimators are a special case, the influence function corresponding to a contamination at the r^* -th category is obtained as

$$IF(r^*, F) = I(\boldsymbol{\theta})^{-1} S(r^*, \boldsymbol{\theta}).$$

3 Discussion and concluding remarks

In this contribution we introduce some aspects concerning robustness issues relevant for a specific class of ordinal data. For this model the contaminations in the data can produce substantial changes of the final model within the parameter space, with a consequent significant impact on the interpretation of the results.

Specifically, the study of the influence function for the estimators of the parameters in CUB mixture enhances several aspects of robustness related to anomalous scores: atypical responses or outliers in sample surveys, for instance.

On the basis of these preliminary results, we are planning to consider the construction of robust estimators for CUB models parameters and the extension of robustness issues generated by the introduction of covariates concerning respondents.

Acknowledgements This work has been realized with the partial support of FIRB2012 project on “Mixture and latent variable models for causal inference and analysis of socio-economic data”, at University of Perugia.

References

1. Croux, C., Haesbroeck, G., Ruwet, C.: Robust Estimation for Ordinal Regression. Katholieke Universiteit Leuven Business & Economics Working Paper, Department of Decision sciences and Information management, Faculty of Business and Economics, KBI 1110 (2011)
2. Iannario, M.: Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications* **21**, 1–22 (2012)
3. Iannario, M., Piccolo, D.: CUB Models: Statistical Methods and Empirical Evidence. In Kenett, R. S. and Salini, S. (eds). *Modern Analysis of Customer Surveys*. J.Wiley & Sons, New York, 231–254 (2012)
4. Iannario, M., Monti, A.C., Piccolo, D.: Robustness issues for CUB models. Preliminary report (2013)
5. Maronna R. A., Martin R. D. & Yohai V. J. (2006). *Robust Statistics: Theory and Methods*. J. Wiley & Sons, NY.
6. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* **5**, 85–104 (2003)
7. Victoria-Feser, M.-P., Ronchetti E.: Robust Estimation for Grouped Data. *Journal of the American Statistical Association*. **92**, 333–340 (1997)

A class of ordinal data models in R

Maria Iannario, Domenico Piccolo

Abstract This paper is devoted to present a statistical program for a class of distributions, called CUB models, introduced for the purpose of interpreting and fitting ordinal responses. More specifically, we present the procedure in R for the estimation and validation of CUB models also when some covariates are included. Special emphasis is given to the graphical tools generated by such models since they allow an immediate visualization of the effects of covariates with respect to space, time and circumstances.

Key words: CUB models, R program, ordinal data, covariates effect

1 Introduction

This paper documents the essential steps for performing statistical inference by using an R program designed for the estimation and testing of CUB models: we refer to [8] for discussion about the main features and the motivations of this approach. The current version of the program significantly improves initial value estimation, fitting measures, plotting facilities for several items and also deals with generalizations and extensions of the basic framework. Moreover, the paper shows how to display several CUB models related to several items or to the same item evaluated in different circumstances.

Maria Iannario

Department of Political Sciences, University of Naples Federico II, Naples, e-mail: maria.iannario@unina.it

Domenico Piccolo

Department of Political Sciences, University of Naples Federico II, Naples, e-mail: domenico.piccolo@unina.it

The paper is organized as follows. In Section 2 we report the main structure of the program, Section 3 comments a graphical system to synthesize more CUB models over the same parameter space. Some conclusion ends the paper.

2 Implementing a CUB model

A recent alternative to standard generalized linear models has been proposed by means of CUB (Combination of Uniform and Shifted Binomial random variables) models [10, 4]. The main structure of the program was firstly implemented in the GAUSS language and then translated in the R environment. Here, we present the version 4.0 (released in summer 2013) freely available from the Authors upon request. Several aspects on the methodological issue of the program are listed in [8] with an Appendix containing the main structure of the R program.

We assume that a vector of ordinal data or a data matrix of several ordinal data related to a set of items is available (without missing values) in a text (ASCII) file or a file readable by R. In case of missing value it is possible to use the framework of CUB models within a procedure for efficient imputation of data (see [1]).

In the following, we denote by *ordinal* a single vector and *matord* a matrix of ordinal data.

```
> matord=read.table(filename.txt,header=T)
> ordinal=dati[,j] # if ordinal data are in the jth column
```

The program runs a function CUB(.) which in turn is related to several other modular functions used for more specific analysis in order to estimate and testing a CUB model. All the inference rests on the asymptotic maximum likelihood theory [11].

The main commands are:

```
> source(CUB.R)
> m=number_of_categories
> CUB(ordinal,Y=paicov, W=csicov, shelter=c)
```

Explicit indication of m is fundamental for this version of program. Here we denote by Y, W the matrices of data whose columns contain subjects' covariates for explaining uncertainty and feeling, respectively. Since default values are set to $Y = 0, W = 0, shelter = 0$, it is quite simple to run specific CUB models. In addition, the presence of a *shelter* effect [5] allows to include the possibility to analyse higher concentration of responses in specific modalities within the same framework.

3 Graphical display for several estimated CUB models

For synthesizing the facilities offered by the current program, we report the code for representing several CUB models.

```

multicub<-function(matord,m,etich=as.character
  (1:ncol(matord)), titolo="CUB models"){
  k=ncol(matord);  vettpai=vettcsi=c()
  for(j in 1:k){
  CUB(matord[,j])
  vettpai[j]=pai; vettcsi[j]=csi
  }
  plot(1-vettpai,1-vettcsi,main=titolo,cex=1.2,cex.main=1,
    font.lab=4,cex.lab=1, pch=19, xlim=c(0,1),ylim=c(0,1),
    xlab=expression(paste("Uncertainty ", (1-pi))),
    ylab=expression(paste("Feeling ", (1-xi))))
  text(1-vettpai,vettcsi,labels=etich,pos=2,offset=0.3,
    cex=0.7)
  }

```

Figure 1 is obtained by implementing the previous code on a sample of 311 respondents on a survey concerning the satisfaction of transport system in Naples (data collected in 2012). It summarizes several CUB models (fitted to 16 items) concerning a different scale ($m = 5, 7, 10$, respectively). In this way, in a single graphical representation, we show how and if the number of modalities affects both feeling

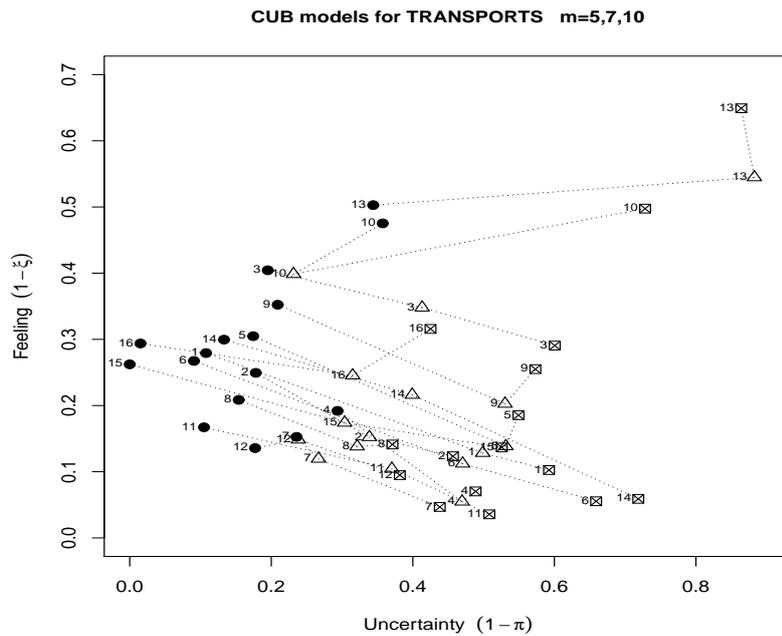


Fig. 1 CUB models on several item concerning customer satisfaction of Naples mobility system on different scale $m = 5, 7, 10$.

and uncertainty. This graphic underline as the changing of the scale produces higher uncertainty in the responses (dots, triangles and squares represent CUB models on 5, 7 and 10 categories, respectively). Notice that some items (13, for instance) show a behaviour completely different from the others: they clearly represent some questions concerning different aspects of satisfaction, generally related to staff or services.

4 Conclusion

The current version should be generalized to take into account also some recent developments as hierarchical models [6], Latent CUB models [3], and a varying uncertainty [2]. The generalization for CUBE models, an extension for taking into account the possible overdispersion sometimes presents in ordinal data [7], and for GECUB models [9], which include covariates also in the case of *shelter* effect, imply more specific programs. Then, the submission of the standard program as a package in the CRAN website should be concluded within next months.

Acknowledgements This work has been realized with the partial support of FIRB2012 project on “Mixture and latent variable models for causal inference and analysis of socio-economic data”, at University of Perugia.

References

1. Cugnata, F., Salini, S.: Comparison of alternative imputation methods for ordinal data. Preliminary Report (2013)
2. Gottard, A., Iannario, M., Piccolo, D.: Varying Uncertainty in CUB Models. Submitted (2013)
3. Grilli, L., Iannario, M., Piccolo, D., Rampichini C. (2012) Latent Class CUB Models. In Proceedings of the 46th Scientific Meeting of the Italian Statistical Society (Available from <http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/view/2337>).
4. Iannario, M. (2010) On the identifiability of a mixture model for ordinal data. *Metron* **LXVIII**, 87–94.
5. Iannario, M.: Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications* **21**, 1–22 (2012)
6. Iannario, M.: Hierarchical CUB Models for ordinal variables. *Communications in Statistics. Theory and Methods* **41**, 3110–3125 (2012)
7. Iannario, M.: Modelling Uncertainty and Overdispersion in Ordinal Data. Submitted (2013)
8. Iannario, M., Piccolo, D.: CUB Models: Statistical Methods and Empirical Evidence. In Kenett, R. S. and Salini, S. (eds). *Modern Analysis of Customer Surveys*. J.Wiley & Sons, New York, 231–254 (2012)
9. Iannario, M., Piccolo, D.: A framework for modelling ordinal data in survey of ratings and evaluations. Proceedings of JSM. San Diego, CA (2012)
10. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica* **5**, 85–104 (2003)
11. Piccolo, D.: Observed information matrix for MUB models. *Quaderni di Statistica* **8**, 33–78 (2006)

Parsimony in Mixtures with Random Covariates

Salvatore Ingrassia and Antonio Punzo

Abstract In the class of mixtures with random covariates, the generalized linear Gaussian cluster-weighted model (GLGCWM) has been recently proposed; in each mixture component, it models the response variable within the exponential family of distributions and the vector of real-valued covariates according to the multivariate Gaussian distribution. Due to the number of free parameters of each covariance matrix of the component Gaussian distributions, a family of fourteen parsimonious GLGCWMs is here introduced by applying some constraints on the eigen decomposition of these matrices. This novel family of models is also applied to a real data set where it gives good classification performance, especially when compared with more established mixture-based approaches.

Key words: cluster-weighted models, model-based clustering, generalized linear models, eigen decomposition, parsimonious mixtures.

1 Introduction

Let (Y, \mathbf{X}') be a random vector where Y is the response variable and \mathbf{X} is the p -variate random vector of real-valued covariates. Moreover, let $p(y, \mathbf{x})$ be the joint density of (Y, \mathbf{X}') . A flexible frame for density estimation and clustering of data $\{(y_i, \mathbf{x}_i)'\}_{i=1}^n$ from (Y, \mathbf{X}') is represented by the family of mixture models with random covariates that, by considering modeling also for \mathbf{X} , are to be preferred to mixtures with fixed covariates for most of the applications (see, e.g., Hennig, 2000).

An eminent member of the family of mixtures with random covariates is the cluster-weighted model (CWM; Gershenfeld 1997). The CWM principle consists in factorizing $p(y, \mathbf{x})$, in each mixture component, into the product between the conditional density of $Y|\mathbf{X} = \mathbf{x}$ and the marginal density of \mathbf{X} by assuming a parametric functional dependence of Y on \mathbf{x} . Thus, CWMs embrace the potential of mixtures of regression models and of mixtures of distributions; the idea of the former approach

Salvatore Ingrassia and Antonio Punzo
Department of Economics and Business, University of Catania (Italy)
e-mail: [s.ingrassia,antonio.punzo]@unict.it

is adopted to model the conditional distribution of $Y|\mathbf{x}$, while the principle of the latter is used to model both $p(y, \mathbf{x})$ and the marginal density of \mathbf{X} . For recent developments in CWMs see Ingrassia *et al.* (2012b), Ingrassia *et al.* (2013), Ingrassia *et al.* (2012a), Ingrassia and Punzo (2013), Punzo (2012), and Subedi *et al.* (2013). In particular, Ingrassia *et al.* (2012a) propose the generalized linear Gaussian CWM (GLGCWM) which adopts a generalized linear model for the relationship of Y on \mathbf{x} in each mixture component. This implies the possibility to model, for example, a count response via a Poisson distribution and a dichotomous response by a Bernoulli distribution.

However, when p increases, the number of parameters to be estimated in the GLGCWM increases too, especially due to the contribute of the covariance matrices of the component Gaussian distributions. To make the approach parsimonious, in line with Celeux and Govaert (1995), a family of fourteen GLGCWMs is introduced in Section 2 by applying some constraints on the eigen decomposition of the component covariance matrices. The EM algorithm is adopted for maximum likelihood parameters estimation and the BIC is considered for model selection in the novel family. An application to real data is presented in Section 3.

2 Parsimonious generalized linear Gaussian CWMs

The generalized linear Gaussian cluster-weighted model (GLGCWM; Ingrassia *et al.*, 2012a) is a finite mixture model, with k components, of equation

$$p(y, \mathbf{x}; \Psi) = \sum_{j=1}^k \pi_j p(y|\mathbf{x}; \theta_j, \zeta_j) \phi(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \quad (1)$$

where π_j is the weight of the j th component, with $\pi_j > 0$ and $\sum_{j=1}^k \pi_j = 1$, $\phi(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is the density of a p -variate Gaussian random vector with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$,

$$p(y|\mathbf{x}; \theta_j, \zeta_j) = \exp \left\{ \frac{y\theta_j - b(\theta_j)}{a(\zeta_j)} + c(y; \zeta_j) \right\}, \quad (2)$$

for specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$, is an exponential family distribution, and Ψ contains all of the parameters of the mixture. It is well known that the exponential family model in (2) is strictly related with the generalized linear models. In (1), like in the classical generalized linear models, a monotone and differentiable link function $g(\cdot)$ is introduced which relates the expected value μ_j , of the response $Y|j$, to the covariates \mathbf{X} through the relation $g(\mu_j) = \eta_j = \beta_{0j} + \boldsymbol{\beta}'_{1j}\mathbf{x}$.

Because there are $p(p+1)/2$ free parameters for each $\boldsymbol{\Sigma}_j$, it is usually necessary to introduce parsimony into the general model (1) for real applications. With this end, following Celeux and Govaert (1995), we consider the eigen decomposition $\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{\Gamma}_j \boldsymbol{\Delta}_j \boldsymbol{\Gamma}_j'$, where $\lambda_j = |\boldsymbol{\Sigma}_j|^{1/p}$, $\boldsymbol{\Delta}_j$ is the scaled ($|\boldsymbol{\Delta}_j| = 1$) diagonal

matrix of the eigenvalues of Σ_j , and Γ_j is a $p \times p$ orthogonal matrix whose columns are the normalized eigenvectors of Σ_j . Each eigen-component has a different geometric interpretation: λ_j determines the volume of the cluster, Δ_j its shape, and Γ_j its orientation. The constraints we pose on them generate the family of fourteen parsimonious GLGCWMs models summarized in Table 1.

Table 1 Nomenclature, covariance structure, type of maximum likelihood (ML) solution in the M-step of the EM algorithm (CF=“closed form” and IP=“iterative procedure”), and number of free covariance parameters for each parsimonious GLGCWM.

Model	Volume	Shape	Orientation	Σ_j	ML	Free covariance parameters
EII	Equal	Spherical	-	$\lambda \mathbf{I}$	CF	1
VII	Variable	Spherical	-	$\lambda_j \mathbf{I}$	CF	k
EII	Equal	Equal	Axis-Aligned	$\lambda \Delta$	CF	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_j \Delta$	IP	$k + p - 1$
EVI	Equal	Variable	Axis-Aligned	$\lambda \Delta_j$	CF	$1 + k(p - 1)$
VVI	Variable	Variable	Axis-Aligned	$\lambda_j \Delta_j$	CF	kp
EEE	Equal	Equal	Equal	$\lambda \Delta \Gamma \Delta'$	CF	$p(p + 1)/2$
VEE	Variable	Equal	Equal	$\lambda_j \Delta \Gamma \Delta'$	IP	$k + p - 1 + p(p - 1)/2$
EVE	Equal	Variable	Equal	$\lambda \Delta_j \Gamma \Delta_j'$	IP	$1 + k(p - 1) + p(p - 1)/2$
EEV	Equal	Equal	Variable	$\lambda \Delta \Gamma_j \Delta_j'$	CF	$p + kp(p - 1)/2$
VVE	Variable	Variable	Equal	$\lambda_j \Delta_j \Gamma \Delta_j'$	IP	$kp + p(p - 1)/2$
VEV	Variable	Equal	Variable	$\lambda_j \Delta \Gamma_j \Delta_j'$	IP	$k + p - 1 + kp(p - 1)/2$
EVV	Equal	Variable	Variable	$\lambda \Delta_j \Gamma_j \Delta_j'$	CF	$1 + k(p - 1) + kp(p - 1)/2$
VVV	Variable	Variable	Variable	$\lambda_j \Delta_j \Gamma_j \Delta_j'$	CF	$kp(p + 1)/2$

3 Real data analysis: the `f.voles` data set

The `f.voles` data set, detailed in Flury (1997, Table 5.3.7) and available in the **Flury** package for R, consists of measurements of female voles from two species, *M. californicus* and *M. ochrogaster*. The data consist of 86 observations for which we have a binary variable Species denoting the species (45 *Microtus ochrogaster* and 41 *M. californicus*), a response variable Age measured in days, and $p = 6$ real-valued covariates related to skull measurements. For our purpose, we assume that data are unlabelled with respect to Species and that our interest is in evaluating clustering using the GLGCWMs as well as comparing the algorithm with some well-established mixture model-based techniques. Therefore, Age can be considered the natural Y variable and the $p = 6$ skull measurements can be considered as the vector of covariates \mathbf{X} .

By considering a Gaussian distribution for Y , all fourteen GLGCWMs were fitted, assuming no known group membership, for $k \in \{1, \dots, 5\}$. The model with the largest BIC (-3863.451) was VEE with $k = 2$. Table 2(a) displays the clustering results from this model (group memberships are individuated by *maximum a posteriori* probabilities). Table 2 also shows the clustering results of mixtures of linear Gaussian regression models (MLGRMs) and of mixtures of linear Gaussian regression with concomitant models (MLGRCMs). They were fitted via the `stepFlexmix()` function of the R-package **flexmix** (Grün and Leisch, 2008). The best value for $k \in \{1, \dots, 5\}$ was selected by BIC. The best results are clearly ob-

Table 2 Clustering of the `f.voles` data using three different approaches.

(a) VEE-GLGCWM				(b) MLGRM						(c) MLGRCM			
True \ Est.	1	2		True \ Est.	1	2	3	4	5	True \ Est.	1	2	
<i>Ochrogaster</i>	43	2		<i>Ochrogaster</i>	14	5	6	15	5	<i>Ochrogaster</i>	15	30	
<i>Californicus</i>	0	41		<i>Californicus</i>	10	3	8	9	11	<i>Californicus</i>	3	38	

tained for the VEE-GLGCWM where the number of groups is correctly selected and only two *M. ochrogaster* observations are misclassified as *M. californicus*.

4 Conclusions

We extended the generalized linear Gaussian cluster-weighted model (GLGCWM) by introducing a family of fourteen parsimonious GLGCWMs. We illustrated our approach on real data where our method gave impressive superior clustering results when compared to other more famous mixture approaches. Although we have demonstrated the usefulness of our models for model-based clustering, they can also be applied for model-based classification or discriminant analysis.

References

- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.
- Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer, New York.
- Gershensfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.
- Grün, B. and Leisch, F. (2008). **FlexMix** version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**(4), 1–35.
- Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, **17**(2), 273–296.
- Ingrassia, S. and Punzo, A. (2013). Fitting bivariate mixed-type data via the generalized linear exponential cluster-weighted model. Technical report, available at: <http://arxiv.org/abs/1304.0150>.
- Ingrassia, S., Minotti, S. C., Punzo, A., and Vittadini, G. (2012a). Generalized linear Gaussian cluster-weighted modeling. Technical report, available at: <http://arxiv.org/abs/1211.1171>.
- Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012b). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.
- Ingrassia, S., Minotti, S. C., and Punzo, A. (2013). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*. DOI: 10.1016/j.csda.2013.02.012.
- Punzo, A. (2012). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. Technical report, available at: <http://arxiv.org/abs/1207.0939>.
- Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.

International Relations Based on the Voting Behavior in General Assembly

Hiroshi Inoue¹

Abstract The aim of this paper is to analyze the structure of similarity of voting behaviours in the General Assembly of United Nations in every five years from 1946 to 1985. It is an attempt to know the structure of world system as a whole. Methodologically, the method of blockmodel analysis which is a type of network analysis is applied. As a result, the structure in the Assembly reflecting the Cold War is found in all periods, especially in the earlier periods, while, in the later stages, the structure reflecting the North-South problem is more salient, though nations of the Third World have been led to more heterogeneous behaviours.

1 What are Problems?

In this paper, two themes being related each other will be argued. One is substantive so that the international relations after the World War II are focused from the view of historical sociology in a broad sense. Another one is methodological so that the network analysis is focused.

International relations have been changing very dynamically. However, on the other hand, some kind of structure has been observed in each stage. The international relations after the World War II have been depicted in some ways. Probably, three moments which have different direction may be well-known. One is a crisis moment called the Cold War. In fact, the Cold War is said to have continued from 1945 to 1989. The second moment is the extension of global economy. The third one is a cooperative

¹ Hiroshi Inoue ; email: inoue-h@bbplus.net

moment symbolized by the United Nations. Each of three moments is related with different paradigm---one of hegemony, one of interdependence, and one of cooperation.

Various types of studies from these perspectives have appeared. Among them, there are some researches with the methodological breakthrough in a sense that they attempted to treat the international relation as a whole and to introduce formal approach. For example, the study of Snyder and Kick (1979) explored multiple international relations---trade, military relations, diplomats, and treaties---by applying methods of network analysis, especially the blockmodel analysis. They found different structure for each relation, though they insisted the core-semiperiphery-periphery structure as a basic one. The star structure appears in the sphere of trade in which the developed countries are core nodes, while there is a relatively independent leagues of nations belonging to the Third World in other relations.

This paper is going to analyze the data of voting behaviours of nations in the Assembly by using the method of network analysis so that we will find the proper structure and its change.

2 Data and Method

The raw data is the collection of voting behaviors of nations in the General Assembly in each year from 1946 to 1985. The data source is Urano (1987). The raw data set contains the year of resolution, the attitude of each nation, and so on. The attitude is shown by one choice from four alternatives---approval for, opposition against, abstention, absent. In this paper, the data is simplified in two points. Areas of agenda---disarmament, peacekeeping, human rights etc --- are not distinguished. The data was aggregated in every five years.

At first, I make a coincidence matrix in which each of the row and column has three categories---approval, objection, abstention--- between every pair of nations. Next, I calculated a coincidence coefficient being defined as the ratio of frequency on the main diagonal. We are able to produce a $n \times n$ square matrix by collecting those coefficients in each period. (The n means the number of nations.) Eventually, we reach eight matrices that are the data to be analyzed.

This coincidence matrix is regarded as a kind of social network in which each nation is connected with other nations. The degree of coincidence in each cell is interpreted as showing the similarity of two nations or the strength of a coalition tie between them. Methods of formal network analysis are available to explore some hidden structure of this network. We adopt the blockmodel analysis developed by White, Boorman, and Breiger (1976) so that we are able to pull out a partition of classes as a structure. The blockmodel analysis is to find a reduced image graph from an original graph. A simple formal expression may be, using a matrix notation of a graph, as

$$B(c, d) \stackrel{\phi^{-1}}{\mapsto} A(i, j)$$

where A is an original data matrix and B is a blockmodel. What should be questioned is a specification of ϕ . First, ϕ generates a partition of the original graph A in terms of the structural equivalence which is defined on the graph A as

$$SE(u, v) \Leftrightarrow uAx \text{ iff } vAx \text{ and } xAu \text{ iff } xAv \text{ for } \forall x$$

where u, v, x are nodes of the graph A , and $SE(u, v)$ is a pair of structurally equivalent nodes. Second, ϕ means giving either 1 or 0 to each subgraph made of partitioned classes by evaluating the density of the subgraph. Now, we get a reduced graph called blockmodel.

The blockmodel is a theoretical concept based on the idea of structural equivalence. The concept has been eased operationally. Here, I will apply the CONCOR procedure which is developed by White, Boorman and Breiger (1976)...

3 Findings and Discussion

I cannot go into detail of all periods because of limited space, I will argue two periods as illustrations. (The notation of table number is given according to the full-paper.)

In the third period, from 1956 to 1960, the four block model is possible. Members of each class are listed below (the name of nation is abbreviated). The table 3a and 3b show the structure of blockmodel. The block-1 (The USSR block) and the block-4 (USA block) are not connected. It means the structure of the Cold War. At the same time, we realize that four blocks form a line on which the relatively independent blocks exist. They intervene between the USSR block and USA block that are terminated in the far ends of the path.

Block-1 (27 nodes) : Afghn Alban Bulga Burma CzecS Eygpt Ghana Guine Hunga
India Indon Iraq Mali Moroc Nepal Nigeria Polan Roman Seneg SriLa Sudan
Syria USSR Yemen Yugo etc

Block-2 (22 nodes) : CentA Chad Congo Cypr Cambo Ethio Finld Gabon Coted
Jorda Leban Libya Niger SaudA Somal Togo Tunij Camer UppVo Zaire etc.

Block-3 (26 nodes): Argen Austri Boliv Costa Cuba Denmk Ecuad Guate Haiti Icela
Iran Irela Israe Laos Liber Malay Mexic Norw Pakis Panam Parag Philp Swed
Thail Urugu Venez

Block-4 (25 nodes): Austra Belgi Brazi Canad Chilie China Colom DomRe Franc
Grec Hondu Italy Japan Lux Nethe NewZe Nicar Peru Portu SouthA Spain
Turkey UK USA etc.

Table 3a. Density matrix of the blockmodel for the period from 1956 to 1960.

	1	2	3	4
1	0.77	0.66	0.56	0.39
2	0.66	0.81	0.68	0.46
3	0.56	0.68	0.81	0.72
4	0.39	0.46	0.72	0.76

Table 3b. The image graph matrix
The cut-off value is the average density (0.630)

	1	2	3	4
1	1	1	0	0
2	1	1	1	0
3	0	1	1	1
4	0	0	1	1

In the fifth period, from 1966 to 1970, the structure seems to be changing. Let's see the list below and the table 5a and 5b. They show the four blocks model. The block-4 with the size of 20 is mainly composed of developed capitalistic countries and disconnected with other blocks including 107 countries. On the other hand, USSR belongs to the block-1 which is larger than the block-4 of USA and also is linked with the block-2, though it is difficult to evaluate whether USSR is more influential than USA or is absorbed into a peripheral block. We may say that the Cold War still exists.

But, rather, it is clear that there appears the confrontation between the developed countries and developing ones, that is, the South-North problem.

Block-1 (49 nodes): Afghn Alban Alger Bulga Burma Congo Cuba Cypr CzecS
 Cambo Eygpt Ghana Hunga India Indon Iraq Jorda Kenya Kuwai Leban Libya
 Mongo Moroc Nepal Nigeria Pakis Polan Roman SaudA Somal SriLa Sudan Syria
 Tunij Ugand USSR Camer Yemen Yugo Zamb etc.

Block-2 (44 nodes): Barba Benin CentA Chad Chile China Colom DomRe Ecuad
 Ethio Gabon Gambi Greec Guate Guyan Haiti Hondu Iran Coted Jamai Laos Liber
 Madag Malay Mexic Niger Panam Peru Philp Rwand Seneg Singap Spain Thail
 Togo Turky UppVo Urugu Venez Zaire etc.

Block-3 (14 nodes): Argen Boliv Botsw Brazi Costa ELSal Fiji Irela Israe Japan Lesot
 Nicar Parag Swadi

Block-4 (20 nodes): Austra Austri Belgi Canad Denmk Finld Franc Icela Italy Lux
 Malw Malta Nethe NewZe Norw Portu SouthA Swed UK USA

Table 5a. The density matrix of
 blockmodel for the 66-70 period

	1	2	3	4
1	0.812	0.688	0.545	0.332
2	0.688	0.768	0.703	0.462
3	0.545	0.703	0.713	0.550
4	0.332	0.462	0.55	0.681

Table 5b. The image graph matrix
 The cut-off value is
 the average density (0.596)

	1	2	3	4
1	1	1	0	0
2	1	1	1	0
3	0	1	1	0
4	0	0	0	1

4. Concluding Remarks

The UN is an international institution to solve problems of disparity of wealth, conflict, and so on in this world. It is a cooperative regime, but, at the same time, an arena of confrontations and coalitions. According to the network analysis of voting behaviors in the Assembly of UN, the structure of the East-West confrontation, that is, the Cold War has been found in the earlier stages after the World War II, and has been maintained until the 1980's. However, the North-South problem has appeared in early stages and has become more salient in the late 1960's. The Third World has gotten larger, but been partitioned heterogeneously.

References

1. Snyder, R., Kick, E.: Structural Position in the World System and Economic Growth 1955-1970: A Multiple-Network Analysis of Transitional Interactions. *Am. J. of Sociol.* 84(5), pp.1096-1126 (1979)
2. Urano, Yuki: *Kokusai Shakai no Henyo to Kokuren Tohyo Kodo ---1945~1985---II: Kiso deta oyobi Kakkoku Tohyo Kodo Bunseki.* Kokusai Chiiki Shiryo Center (1987). (Lang. in Japanese)
3. White, H. C., Boorman, S. A., and Breiger, R. L: Social Structure from Multiple Networks I. *Am. J. of Socil.* 81, pp.730-779. (1976)

Visual model representation and selection for classification and regression trees

Carmela Iorio, Massimo Aria, Antonio D'Ambrosio

Abstract Within the framework of recursive partitioning algorithms by tree-based methods, this paper provides a contribution on both the visual representation of the final data partition in a geometrical space and the selection of the decision tree. The results in terms of error rate are really similar to the ones returned by the Classification And Regression Trees procedure, showing how this novel way to select the best tree is a valid alternative to the well know cost-complexity pruning.

Key words: Classification trees, Regression trees, Pruning, Visual Data Analysis.

1 A short introduction to recursive tree-based partitioning methods

Recursive partitioning tree procedures have been the subject of extensive research in the past. Specially tree-based methods have been proposed for both prediction and exploratory purposes. Hierarchical segmentation obtained by decision trees can be seen as stepwise procedures, performed according to an optimization criterion, that provides a progressive sequence of partitions of a set initial of objects, described by some explanatory variables (either numerical or/and categorical) and a response variable (Hastie et al., 2009), via a top down criteria.

The oldest tree based method was Automatic Interaction Detector (AID) proposed by Morgan and Sonquist (1963). AID is an algorithm for growing a binary regression tree in which the outcome variable is quantitative and the splitting rule taken into account the reduction in unexplained sum of squares. Messenger and Mandell (1972) and Morgan and Messenger (1973) extended AID for categorical outcome using a so called *theta criterion* (THAID, THeta Automatic Interaction Detector).

Dept of Economics and Statistics, Dept of Economics and Statistics, Dept of Industrial Engineering, University of Naples Federico II, e-mail: (carmela.iorio, massimo.aria, antdambr)@unina.it

The aim of these earlier methods was the segmentation of the data into groups being as much different as possible in terms of the distributions of the outcome variable.

A descendant of AID and THAID is CHAID (CHi-square Automatic Interaction Detector), introduced by Kass (1980). CHAID uses a Chi-square splitting criterion to classify a categorical response variable. To correct for multiple testing a Bonferoni correction is applied.

One of the most popular tree-based techniques is Classification And Regression Trees (CART) developed by Breiman et al. (1984). Induction of decision trees is typically performed in two steps.

First, a training set is used to grow the tree and the splitting criterion is expressed in terms of decrease in impurity. In the second step, called pruning, the tree is reduced to prevent "overfitting". The CART pruning procedure considers both the aspect of accuracy (evaluated by some error measure which may not necessarily conclude with the one used in the growing phase) and the complexity (given by the number of terminal nodes) of the tree, introducing the so called *cost-complexity measure* (Breiman et al., 1984; Cappelli et al., 2002). The algorithm uses a separate set of samples for pruning, distinct from the learning set and known as test set, that produces a best sequence of pruned nested sub-trees of the full tree (cross-validation).

2 The key idea

The most common approach to build a classification or regression tree is to grow a full tree and prune it back. Each node in the tree is the starting point for a sub-tree which will end with several leaves or terminal nodes. The data at the leaf are evaluated via the misclassification rate or the expected value according to the nature of the response. This involves that the global badness of fit measure which can be either the misclassification rate or the mean squared error.

The classical approach uses the so-called cost-complexity pruning procedure to select the optimal decision tree. The latter is based on the definition of a trade-off measure between the accuracy (*cost*) and the size (*complexity*) of the tree. Through a recursive procedure, the weakest links are cut by minimizing the cost-complexity measure. The result is a sequence of optimal sub-trees candidate to be the final decision tree. The best one is chosen either via test-set or cross-validation.

Our approach is based on the definition of a new way to represent the tree structure. Indeed the length of a path is proportional to the error measure decrease in the sense that lower is the error in the descendant nodes, the longest is the length of the path. In this way, the graph shows immediately the best splits and the purest nodes of the tree.

This consideration suggests us to define an alternative pruning sequence based on a visual approach. Each internal node is an equally likely candidate to be the point which identifies a cut to the depth level of the structure. In this way, given a node t candidate as cutting point, for each path which span the ideal cutting line, the nodes departing from it will become terminal. At each potential cutting line is linked a er-

ror measure of the tree. The distribution of errors follows a typical descending trend over the training sample and a convex trend over the test or CV sample.

Figure 1 shows the visualization of the tree structure using our approach on the breast cancer Wisconsin dataset (left side of figure) compared with the classical one (right side of figure). Immediately the figure points out the relative importance of splits and the best cut level to obtain optimal decision tree.

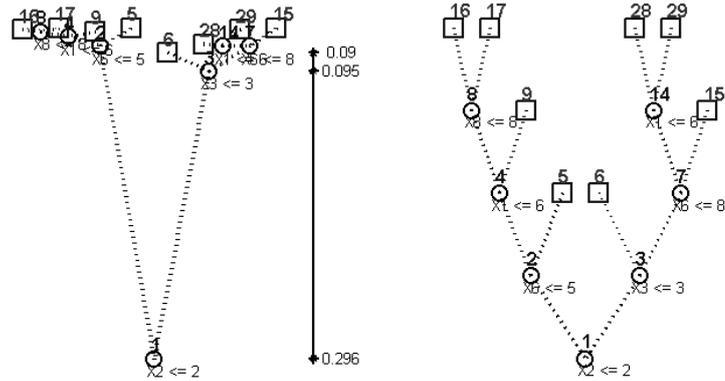


Fig. 1 Comparison between visual (left graph) and classical (right graph) tree structures (Breast cancer Wisconsin dataset)

Figure 2 shows a comparison of pruning sequences of our approach versus cost-complexity procedure. Empirical evidence suggests how both procedures return similar results in terms of prediction error. Moreover in visual approach the identification of the best tree and the weakest links is immediately evaluable by the graphical analysis of the tree structure without consider the pruning sequence.

3 Conclusion

The proposed visual model representation and selection seems to be a valid alternative to the cost-complexity strategy to select decision trees. Moreover we propose a novel tree structure visualization which allows to identify more discriminant splits, weakest links and help the user to catch the optimal sub-structure as decision tree.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont, California (1984)

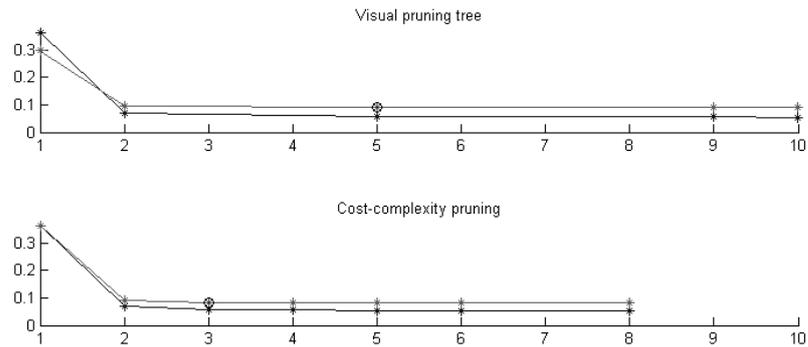


Fig. 2 Comparison between visual and classical pruning sequences (Breast cancer Wisconsin dataset)

2. Cappelli, C., Mola, F., Siciliano, R.: A statistical approach to growing a reliable honest tree. *Computational Statistics & Data Analysis* **38**, 285–299 (2002)
3. Hastie T., Tibshirani R., Friedman J.: *The Elements of Statistical Learning*, Springer (2009)
4. Kass, G. V.: Significance testing in automatic interaction detection (A.I.D.). *Applied Statistics* **24** **2**, 178–189 (1975)
5. Kass, G. V.: An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29** **2**, 119–127 (1980)
6. Messenger, R., Mandell, L.: A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association* **67** **340**, 768–772 (1972)
7. Morgan, J. N., Messenger, R. C.: THAID a sequential analysis program for analysis of nominal scale dependent variables. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor (1973)
8. Morgan, J. N., Sonquist, J. A.: Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* **58**, 415–434 (1963)
9. Quinlan, J.R.: Induction of decision trees. *Machine Learning* **1**, 81–106 (1986)
10. Quinlan, J. R.: C.4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California (1993)

Model risk measurement in currency option pricing

Kuziak Katarzyna

Abstract Empirical research conducted in currency option pricing models pointed out that results are very sensitive to assumptions lying upon them and necessity of managing this type of risk. One of the steps in risk management process is the risk measurement. For the options good risk measure is vega coefficient (first derivative of the option value with respect to volatility of underlying). Two sources of model risk will be considered: estimation error (bad specification of stochastic process for the exchange rate) and use of wrong assumptions. For the exchange rate Geometric Brownian Motion and GARCH family processes will be assumed. Different foreign exchange rates will be analyzed. The goal is to check which of GARCH family processes cause higher model risk comparing to Geometric Brownian Motion.

1 Model risk

Model risk is impossible to totally eliminate. Only when we do not use models, we are free of model risk. Peter Temple compared risk modeling to search for the Holy Grail (Temple 2001).

In pricing models, model risk is defined as the risk arising from the use of a model which cannot accurately evaluate market prices, or which is not a mainstream model in the market.^f Sources of model risk in pricing models include (Kato, Yoshida, 2000):

- 1) use of wrong assumptions,
- 2) errors in estimations of parameters,
- 3) errors resulting from discretization, and
- 4) errors in market data.

In this paper estimation errors (bad specification of stochastic process for the exchange rate) and use of wrong assumptions will be analyzed

¹Katarzyna Kuziak, Department of Financial Investments and Risk Management, Wrocław University of Economics; Poland, email: katarzyna.kuziak@ue.wroc.pl.

2 Currency option pricing models

As is the case with equity derivatives, the implied volatility surface corresponding to vanilla European currency option prices is neither flat nor constant. It is proved that the Garman Kohlhagen model (where Geometric Brownian Motion is assumed) is a poor model for foreign exchange markets. As is the case with equity derivatives, however, vanilla currency option prices are quoted and their Greeks are calculated using the Black-Scholes-Merton framework (Haugh, 2010). It is therefore necessary to compare option values of the Black-Scholes-Merton framework (Black, Scholes 1973, Merton 1973) and pricing models using GARCH approach.

The theory for pricing options assuming the GARCH process was first developed by Duan (1995). By introducing the locally risk-neutral valuation relationship (LRNVR), Duan (1995) derived an option pricing model which depends upon, among other factors, a risk premium parameter. Researches of Amin and Ng (1993) and Chaudhury and Wei (1996) were very promising. They have found a significantly better performance by the GARCH option pricing model in comparison to the BSM framework.

In the paper following processes for the foreign exchange rates r_t are assumed:

- 1) Geometric Brownian Motion (in Garman-Kohlhagen model):

$$r_t = \mu \Delta t + \sigma \sqrt{\Delta t} \varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$

where: μ ; expected rate of return, σ ; volatility (standard deviation of rate of returns).

In this process constant volatility is assumed.

- 2) GARCH(1,1)

$$r_t = \sqrt{h_t} \times \varepsilon_t$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \quad t \in \mathbb{Z}$$

where: α, β, ω ; process parameters of rate of returns, ε_t ; i.i.d. standard normal random variables with 0 expected value and variance of 1

- 3) GARCH(1,1) in mean (in Duan model)

$$r_t = \mu + \lambda \sqrt{h_t} + \sqrt{h_t} \varepsilon_t$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} \quad t \in \mathbb{Z}$$

where: λ ; risk premium

- 4) process AR(1)-GJR-GARCH(1,1)¹:

$$r_t = \mu + \varphi r_{t-1} + \varepsilon_t$$

¹ Later, GJR-GARCH abbreviation will be used.

$$h_t = \omega + (\alpha + \alpha^{-1} I_{(\varepsilon_{t-1} < 0)}) \varepsilon_{t-1}^2 + \beta h_{t-1}$$

$$V = \frac{\omega}{1 - \alpha - \beta} \times \frac{1}{1 - \varphi^2}$$

where: μ , φ , α ; process parameters of rate of returns, h_t ; conditional variance, V ; unconditional variance if the process.

This process allows modeling of fat tails of the distribution, volatility clustering and autocorrelation.

Currency pricing models:

1) Garman-Kohlhagen model (Garman, Kohlhagen, 1973)

Risk-neutralized exchange rate process under Black-Scholes-Merton is following:

$$d \ln(S_t) = (r - r_f - \frac{\sigma^2}{2}) dt + \sigma dW_t^*$$

Then, European call option pricing formula is:

$$c = S e^{-r_f T} N(d) - X e^{-r T} N(d - \sigma \sqrt{T})$$

$$\text{where: } d = \frac{\ln\left(\frac{S}{X}\right) + \left(r - r_f - \frac{\sigma^2}{2}\right) T}{\sigma \sqrt{T}}$$

S ; spot price of exchange rate, X ; exercise price, r ; domestic risk free rate, r_f ; foreign risk free rate, σ ; volatility of exchange rate, T ; time to maturity.

2) GARCH option pricing model (Duan, 1995)

Locally risk-neutralized exchange rate process under GARCH:

$$\ln \frac{S_{t+1}}{S_t} = r - r_f - \frac{\sigma_{t+1}^2}{2} + \sigma_{t+1} \varepsilon_{t+1}^*$$

$$\sigma_{t+1}^2 = \beta_0 + \beta_1 \sigma_t^2 + \beta_2 \sigma_t^2 (\varepsilon_t^* - \theta - \lambda)^2,$$

$$\varepsilon_{t+1}^* | F_t \sim N(0, 1).$$

Then, foreign exchange option price under GARCH process is following:

$$C(S, \sigma_t; X, T, r, r_f, \beta_0, \beta_1, \beta_2, \theta + \lambda) = e^{-r T} E_0^Q \{ \text{Max}(S_T - X, 0) \}.$$

For valuing option price standard Monte Carlo simulation is used.

3 Vega coefficient

Vega coefficient is the first derivative of the option value with respect to volatility of underlying i exchange rate. General formula for it is as follows:

$$vega = \frac{\partial c(\sigma_i)}{\partial \sigma_i} = \lim_{\kappa \rightarrow 0} \frac{c_i(\sigma_i + \kappa) - c_i(\sigma_i - \kappa)}{2\kappa}.$$

4 Analysis of model risk measurement

Simulation study based on chosen foreign exchange rate and hypothetical option will be given.

5 Summary

Conclusions arising from simulation study will be presented.

References

1. Amin, K., Ng, V.: ARCH Processes and Option Valuation, unpublished manuscript, University of Michigan (1993)
2. Black, F., Scholes, M.: The Pricing of Options and Corporate Liabilities. *J. of Polit. Econ.* 81, pp. 637-659, (1973)
3. Bollerslev, T.: Generalized autoregressive conditional heteroscedasticity. *J. of Econom.* 31, pp. 307-327, (1986)
4. Chaudhury, M. M., Wei, J. Z.: A Comparative Study of the GARCH(1, 1) and Black-Scholes Option Prices, Working Paper, University of Saskatchewan (1996)
5. Cont, R.: Model uncertainty and its impact on the pricing of derivative instruments. *Math. Finance*, (2005)
6. Derman, E.: Model risk. *Risk* 9, pp. 34-37, (1996)
7. Duan J.: The GARCH Option Pricing Model. *Math. Finance*, no 5, pp. 13-32, (1995)
8. Duan J., Wei, J.Z.: Pricing Foreign Currency and Cross-Currency Options Under GARCH, *J. of Deriv.* 7, pp. 51-63, (1999)
9. Garman, M.B., Kohlhagen, S.W.: Foreign Currency Option Values. *Int. Money and Finance* 2, pp. 231-237, (1983)
10. Glosten, L. , Jagannathan, R., Runkle, D.: On the relation between the expected value and the volatility of the nominal excess return on stocks, *J. of Finance* 48, pp. 1179-1801, (1993)
11. Kato, T., Yoshida, T.: Model risk and its control. *Monet. and Econ. Studies*, Dec. pp. 129-156, (2000)
12. Merton, R.: Theory of Rational Option Pricing, *Bell J. Econ. and Manag. Science* 4, pp. 141-183, (1973)
13. Temple, P.: *Hedge Funds: Courtesans of capitalism*, Wiley and Sons, New York (2001)

Log-Mean Linear Parameterizations for Smooth Independence Models

Monia Lupparelli, Luca La Rocca and Alberto Roverato

Key words: marginal independence, smooth model, sparse table

1 Introduction

In categorical data analysis the choice of suitable parameterizations is a relevant aspect for several reasons: (i) the parameter space is often involved, (ii) its dimension rapidly increases with the number of variables, (iii) tables are sparse for high dimensional data, (iv) models specified by non-linear constraints on joint probabilities can result in non-smooth models. There is, in particular, an interest in parameterizations defining smooth and interpretable models by means of linear constraints on the parameter space. These considerations motivate the increasing attention for novel parameterizations; see [5], [1] and [9].

We focus on the log-mean linear (LML) parameterization recently introduced by [9] for the binary case and then generalized by [8], which is suitable for models of marginal independence, also known as bi-directed graph or covariance graph models; see [3] and [4]. These models investigate the marginal independence structures of the variables and are very useful in high dimensional data analysis, where working in low-dimension sub-spaces is highly desirable.

Monia Lupparelli

University of Bologna, Department of Statistical Sciences, Via Belle Arti 41, 40126 Bologna, Italy,
e-mail: monia.lupparelli@unibo.it

Luca La Rocca

University of Modena and Reggio Emilia, Department of Computer, Mathematical and Physical Sciences, Via Campi 213/b, 41125 Modena, Italy, e-mail: luca.larocca@unimore.it

Alberto Roverato

University of Bologna, Department of Statistical Sciences, Via Belle Arti 41 40126 Bologna, Italy,
e-mail: alberto.roverato@unibo.it

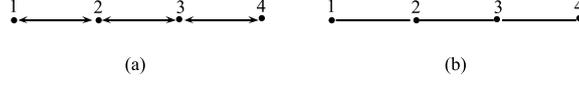


Fig. 1 Independence models for Copen data: (a) a bi-directed graph giving $Y_1 \perp\!\!\!\perp \{Y_3, Y_4\}$ and $\{Y_1, Y_2\} \perp\!\!\!\perp Y_4$, under the connected set Markov property, with $\chi_{(5)}^2 = 8.6$ (p -value = 0.13, BIC = -20.85); (b) an undirected graph giving $Y_1 \perp\!\!\!\perp \{Y_3, Y_4\} | Y_2$ and $\{Y_1, Y_2\} \perp\!\!\!\perp Y_4 | Y_3$, under the global Markov property, with $\chi_{(8)}^2 = 13.9$ (p -value = 0.09, BIC = -33.26).

We show, through an example, how linear constraints on the LML parameterization allow us to specify, at the same time, marginal independencies and partial conditional independencies, thus obtaining a class of smooth parsimonious bi-directed models defined in a lower dimensional space where the constraints have a clear interpretation; see [2] on a similar topic. In addition, through simulations, we show that a convenient choice of variable coding focusses LML models on partial tables with relatively large counts, which results in increased efficiency.

2 Smooth Parsimonious LML Models for Bi-Directed Graphs

We consider a vector $Y_V = (Y_v)_{v \in V}$ of discrete random variables taking values $i_v \in \mathcal{S}_V$, where $\mathcal{S}_V = \times_{v \in V} \{0, 1, \dots, d_v\}$, according to a Multinomial distribution with strictly positive probability parameter $\pi_V = (\pi^{i_v})_{i_v \in \mathcal{S}_V}$, where $\pi^{i_v} = P(Y_V = i_v)$ and $\sum_{i_v \in \mathcal{S}_V} \pi^{i_v} = 1$; π_V belongs to the $|\mathcal{S}_V| - 1$ dimensional simplex Π_V . For every $D \subseteq V$, Y_D takes values $i_D \in \mathcal{I}_D$ with \mathcal{I}_D defined accordingly. The mean parameter is the vector $\mu_V = (\mu^{j_D}, j_D \in \mathcal{I}_D)_{D \subseteq V}$, $\mu_V \in \mu(\Pi_V)$, where $\mu^{j_D} = P(Y_D = j_D)$, $\mu^{j^0} = 1$, and $\mathcal{I}_D = \times_{v \in D} \{1, \dots, d_v\}$. The LML parameter proposed by [9] and [8] is the vector $\gamma_V = (\gamma^{j_D}, j_D \in \mathcal{I}_D)_{D \subseteq V}$ defined by the smooth mapping $\Pi_V \rightarrow \gamma(\Pi_V)$

$$\gamma^{j_D} = \sum_{E \subseteq D} (-1)^{|D \setminus E|} \log(\mu^{j_E}); \quad (1)$$

for every $D \subseteq V$ we define $\gamma_D = (\gamma^{j_D})_{j_D \in \mathcal{I}_D}$, which is a subvector of γ_V .

Let $\mathcal{B} = (V, E)$ be a bi-directed graph defined by a finite set V of nodes and a symmetric set of edges $E \subseteq V \times V$ drawn as bi-directed. Under the *pairwise Markov property*, for the vector Y_V , a missing edge between a pair of nodes $(u, v) \notin E$ corresponds to the marginal independence $Y_u \perp\!\!\!\perp Y_v$. The set of all independencies encoded by \mathcal{B} can be derived using the *connected set Markov property*: given any disconnected set $D \subseteq V$ of nodes in \mathcal{B} , the vectors associated to its connected components Y_{C_1}, \dots, Y_{C_r} are mutually independent; see Fig. 1(a) for an illustration and [4] for technical details. Given a graph \mathcal{B} , the probability distribution of Y_V satisfies the connected set Markov property iff the vector $\gamma_D = 0$ for every disconnected set D of \mathcal{B} ; see [8, Thr. 4.1]. Parameterizations for these models have also been studied by [4] and [6] using respectively the mean and multivariate logistic (MLT) parameter.

We consider the Copen data set including four binary variables concerning symptoms of 362 psychiatric patients: $Y_1 \equiv$ stability (0 = extroverted, 1 = introverted); $Y_2 \equiv$ validity (0 = energetic, 1 = psychasthenic); $Y_3 \equiv$ acute depression (0 = yes, 1 = no); $Y_4 \equiv$ solidity (0 = hysteric, 1 = rigid). These data were analysed by [10], finding the conditional independence model in Fig. 1(b). More recently, [6] and [9] obtained the model in Fig. 1(a) using marginal independence models. Both models achieve a good fit, but they encode different independencies that can be combined only adding further independence relationships; however this operation requires some care because it may lead to non-smooth models; see [1, Ex. 7].

LML models represent a tool which allows us to partially combine the independencies under the two graph models into a single smooth model. In details, we can define an LML model under two sets of linear constraints:

$$\mathcal{Y}_{\{1,3\}} = \mathcal{Y}_{\{1,4\}} = \mathcal{Y}_{\{2,4\}} = \mathcal{Y}_{\{1,3,4\}} = \mathcal{Y}_{\{1,2,4\}} = 0; \quad (2)$$

$$\mathcal{Y}_{\{1,3\}} + \mathcal{Y}_{\{1,2,3\}} = 0, \quad \mathcal{Y}_{\{2,4\}} + \mathcal{Y}_{\{2,3,4\}} = 0, \quad \mathcal{Y}_{\{1,3,4\}} + \mathcal{Y}_{\{1,2,3,4\}} = 0. \quad (3)$$

Constraints in (2) define the bi-directed graph model, while constraints in (3) define the independencies $Y_1 \perp\!\!\!\perp \{Y_3, Y_4\} | \{Y_2 = 1\}$ and $Y_4 \perp\!\!\!\perp \{Y_1, Y_2\} | \{Y_3 = 1\}$ both implied by the undirected graph; for proofs and details about partial conditional independencies see [7, Thr. 6, Cor. 8]. In this way, we achieve a smooth parsimonious bi-directed graph model with $\chi^2_{(8)} = 11.45$ (p -value = 0.18, BIC = -35.68) where all constraints have a clear interpretation in term of independencies.

Partial conditional independencies which can be tested using LML models are of the form $\{Y_C = 1_C\}$ and thus depend on the coding of the variables. We propose to adopt the criterion of *maximal count coding*, so that hypotheses are tested in partial tables with many observations: given a set of binary variables, we will code them so that the cell with the largest count corresponds to all variables taking level 1. We deem this approach should improve the efficiency of inference, especially for large and sparse tables. This feature is illustrated by the following simulation study, which compares the performance of the LML and MLT parameterizations in achieving parsimonious models; the latter parameterization is denoted by η and attains parsimoniousness by setting to zero higher order interactions.

3 A Simulation Study

Consider four binary variables indexed by $V = \{A, B, C, D\}$. We compare the performance in testing the hypothesis $\eta_V = 0$ using the MLT parameter with the performance in testing the hypothesis $Y_C \perp\!\!\!\perp Y_D | \{Y_A = 1, Y_B = 1\}$ using the LML parameter with maximal count coding. Both hypotheses are implied by the independence $Y_C \perp\!\!\!\perp Y_D | \{Y_A, Y_B\}$. We generated a sequence of probability vectors π_k , $k = 1, \dots, 40$, satisfying the constraint $Y_C \perp\!\!\!\perp Y_D | \{Y_A, Y_B\}$. For each π_k , we sampled 5000 multinomial vectors n_w , $w = 1, \dots, 5000$, of size N . For each random sample n_w , we tested the two above hypotheses at $\alpha = 0.05$ nominal significance level, using the $\chi^2_{(1)}$ dis-

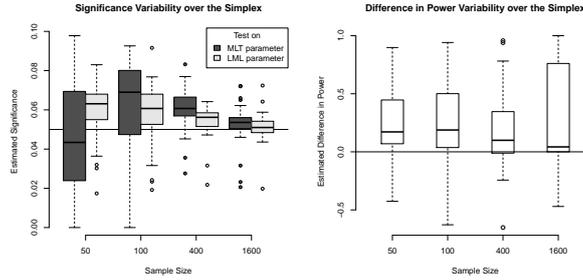


Fig. 2 (a) Box-plots of the estimated significance levels. (b) Box-plot of the difference in power.

tribution. For each π_k , we estimated the finite sample significance level $\hat{\alpha}_k^\eta$ and $\hat{\alpha}_k^\gamma$ of the two tests through the proportion of rejected models in the 5000 random samples, thus obtaining two distributions of estimates. The procedure was repeated for $N = 50, 100, 400, 1600$. Fig. 2(a) compares for every N the two box-plots of the estimated significance levels. The plot shows a lower variability in the estimates and a faster convergence to the nominal value (0.05) for the test on the LML parameter.

We also compared the two tests in terms of power. We replicated our simulations using a sequence of 40 unconstrained probability vectors π_k . For every k , we estimated the type II error of the two tests, $\hat{\beta}_k^\eta$ and $\hat{\beta}_k^\gamma$, through the proportion of accepted models in the 5000 random samples. Fig. 2(b) reports, for every N , the box-plot of the differences in power $\hat{\delta}_k = \hat{\beta}_k^\eta - \hat{\beta}_k^\gamma$, $k = 1, \dots, 40$. The plot shows a clear gain in power for the test based on the LML parameter.

References

1. Bergsma, W. P. and Rudas, T.: Marginal log-linear models for categorical data. *Annals of Statistics* **1**, 140–159 (2002)
2. Colombi, R. and Forcina, A.: A class of smooth models satisfying marginal and context specific conditional independencies. *arXiv:1210.8050v1* (2012)
3. Cox, D. R. and Wermuth, N.: Linear dependencies represented by chain graphs. *Statistical Science* **8**, 204–218 (1993)
4. Drton, M. and Richardson, T.S.: Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B* **70**, 287–309 (2008)
5. Ekholm A., McDonald J. W. and Smith P. W. F.: Association models for a multivariate binary response. *Biometrics* **56**, 712–718 (2000)
6. Lupparelli M., Marchetti G. M. and Bergsma W. P.: Parameterizations and fitting of bi-directed graph models to categorical data. *arXiv:0801.1440v1* (2008)
7. Roverato, A., Lupparelli, M. and La Rocca, L.: Log-mean linear models for binary data. *arXiv:1109.6239v2* (2012)
8. Roverato, A.: Dichotomization invariant log-mean linear parameterization for discrete graphical models of marginal independence. *arXiv:1302.4641* (2013)
9. Roverato, A., Lupparelli, M. and La Rocca, L.: Log-mean linear models for binary data. *Biometrika* **100**, 485–494 (2013)
10. Wermuth, N.: Model search among multiplicative models. *Biometrics* **32**, 253–263 (1976)

Nonlinear CUB models

Marica Manisera and Paola Zuccolotto

Abstract In this contribution, rating data are modelled using the new class of Nonlinear CUB models, recently introduced to generalize the standard CUB. Both CUB and Nonlinear CUB can be viewed as special cases of a general model explaining the cognitive mechanism driving the individuals' responses on a Likert-type scale. A very interesting feature of the Nonlinear CUB is given by the so-called transition probabilities, useful to understand and describe the individual's mental stance towards the response scale. This and all the other nice features of Nonlinear CUB models are presented by an application to real data coming from a survey aimed at evaluating the parents' satisfaction with the service delivered by public kindergartens in Brescia, Italy.

Key words: rating data, Likert-type scales, latent variables, transition probabilities, satisfaction

1 Introduction

The aim of this contribution is to present a case study where rating data are modelled using Nonlinear CUB models, a new class proposed by [5] as a generalization of the well-known CUB models of Piccolo ([2]). The paper is organized as follows: in Section 2 we briefly recall the basic features of CUB models and we present the new class called Nonlinear CUB, in Section 3 we show the results of an application to real data from a survey dealing with satisfaction measurement and draw the main conclusions.

Marica Manisera
University of Brescia, c.da S. Chiara, 50, 25122 Brescia, Italy, e-mail: manisera@eco.unibs.it

Paola Zuccolotto
University of Brescia, c.da S. Chiara, 50, 25122 Brescia, Italy, e-mail: zuk@eco.unibs.it

2 CUB and Nonlinear CUB models

CUB models have been introduced in the literature ([6], [2], [7] [8], [4]) to analyse ordinal data and fit in the latent variable framework. With CUB models rating or ranking data are modelled as a mixture of a Uniform and a Shifted Binomial random variables: the observed rating r ($r = 1, \dots, m$) is the realization of a discrete random variable R whose probability distribution is given by

$$Pr\{R = r|\theta\} = \pi Pr\{V(m, \xi) = r\} + (1 - \pi)P\{U(m) = r\} \quad r = 1, 2, \dots, m$$

with $\theta = (\pi, \xi)'$, $\pi \in (0, 1]$, $\xi \in [0, 1]$. For a given m , $V(m, \xi)$ is a Shifted Binomial random variable with parameter m and success probability $1 - \xi$, modelling the *feeling* component, and $U(m)$ is a discrete Uniform random variable defined over the support $\{1, \dots, m\}$, aimed to model the *uncertainty* component. CUB models are identifiable for $m > 3$ ([3]).

Nonlinear CUB models (NLCUB) are a generalization of CUB, introduced by [5]. In detail, with NLCUB the discrete random variable R generating the observed rating r has a probability distribution depending on a new parameter \bar{m} , $\bar{m} \geq m$, given by

$$Pr\{R = r|\theta\} = \pi \sum_{w \in l^{-1}(r)} Pr\{V(\bar{m}, \xi) = w\} + (1 - \pi) \sum_{w \in l^{-1}(r)} P\{U(\bar{m}) = w\}$$

where l is a step function mapping from $(1, \dots, \bar{m})$ into $(1, \dots, m)$. When $\bar{m} = m$ and $l(w) = w$ for all $w = 1, \dots, \bar{m}$, then the proposed model collapses to classical CUB. In [5] this formulation is derived as a special case of a more general framework aimed to model the cognitive process underlying the individual decision about the rating to express about a certain latent trait. This general model proposed by the authors assumes the presence of two different phases in the decision process, called *feeling* and *uncertainty* phases, respectively. The *feeling* phase of the decision process proceeds through T consecutive steps, so that the final rating of this phase is the result of many elementary judgments that are, firstly, summarized and, secondly, transformed into a Likert-scaled rating. The *uncertainty* phase consists of a final random judgment that can replace the final rating of the *feeling* phase with some given probability. The main feature of the proposed model is the possibility to express the so-called transition probabilities $\phi_t(s)$, i.e. the probability of moving to rating $s + 1$ at step $t + 1$ of the *feeling* phase, given that the rating at step t is s , $s = 1, \dots, m - 1$. Transition probabilities describe the state of mind of the respondents about the Likert scale used to express judgments and the decision process has been defined by the authors to be linear or nonlinear according to whether the transition probabilities are constant or non-constant for different t and s . In [5], the authors derive, under some general assumptions, a sufficient condition for linearity of the decision process and show both that CUB is a particular case of the general framework and that it meets the sufficient condition for linearity. NLCUB models, instead, are a nonlinear variant of the general model and this is the reason for their

name (a graphical explanation is in [5]). When different NLCUB models are considered, while the corresponding parameters π can be fairly compared, this is not allowed for parameters ξ , due to the presence of the function l which can be different from one model to another. We overcome this drawback by introducing the parameter μ , the expected number of one-rating-point increments during the feeling phase ([5]). The estimation of NLCUB models is a difficult task and presents identifiability problems. Nevertheless, thanks to some constraints and a simplifying assumption, it is possible to resort to the algorithm used to estimate classical CUB models, and the results of simulations studies are encouraging ([5]).

3 Case study and conclusions

The presented case study deals with the dataset ([1]) collected during a survey aimed at evaluating the quality of the service delivered by public kindergartens in Brescia, Italy. The parents of the 1916 children attending public kindergartens in 2003/2004 were asked to evaluate, using a 5-point Likert scale, their degree of satisfaction with 22 aspects of the service. The questionnaire, when necessary, was translated into foreign languages. The final dataset was composed of the responses of $n = 1338$ families, with a low number of sparse missing values, which we treat with listwise deletion for each analysed aspect. Due to space constraints, we present here only the results of NLCUB models fitted to data of four selected aspects (Table 1).

Table 1 Aspects under study

A1	Are you satisfied with your child's daily activities at school?
A2	Do you think your child feel affection for his/her teacher?
A3	Are you satisfied with auxiliary operators of the school?
A4	Are you satisfied with the daily organization of children's check-in/check-out?

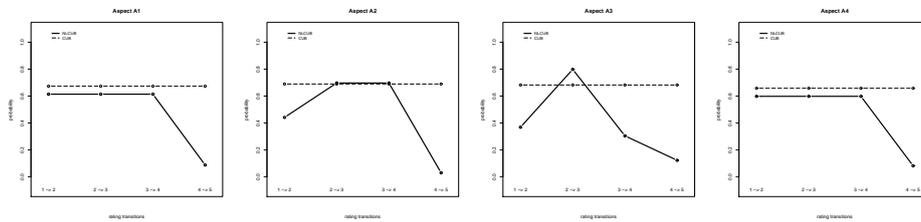
We carried out the analysis using CUB and NLCUB (Table 2). With respect to CUB, NLCUB models provide very similar estimates of μ and sometimes slightly smaller estimates of π . However, the NLCUB estimates of \bar{m} and $\{l^{-1}(r)\}_{r=1,\dots,5}$ suggest the presence of a decision process with a nonlinear structure, that cannot be identified by CUB.

Figure 1 shows the patterns of average transition probabilities implied by the estimated CUB and NLCUB models. For the four analysed aspects, we detect nonlinear decision processes with a decreasing probability of moving to the response denoting the highest degree of satisfaction.

Concluding, this case study shows how NLCUB models allow us to model rating data resulting from cognitive mechanisms with non-constant transition probabilities, thus extending the possibilities of application of the well-known framework of CUB models.

Table 2 Parameter estimates

Aspect	Parameter	CUB	NLCUB
A1	ξ	0.3266	0.3861
	$\bar{m}; \{l^{-1}(r)\}_{r=1, \dots, 5}$	-	6; {(1); (2); (3); (4, 5); (6)}
	μ	2.6935	2.7081
	π	1	1
A2	ξ	0.3104	0.3028
	$\bar{m}; \{l^{-1}(r)\}_{r=1, \dots, 5}$	-	9; {(1, 2, 3); (4); (5); (6, 7, 8); (9)}
	μ	2.7585	2.7846
	π	1	0.9895
A3	ξ	0.3178	0.2014
	$\bar{m}; \{l^{-1}(r)\}_{r=1, \dots, 5}$	-	11; {(1, 2, 3, 4, 5); (6); (7, 8); (9, 10); (11)}
	μ	2.7289	2.7387
	π	1	0.9769
A4	ξ	0.3420	0.4019
	$\bar{m}; \{l^{-1}(r)\}_{r=1, \dots, 5}$	-	6; {(1); (2); (3); (4, 5); (6)}
	μ	2.6319	2.6567
	π	1	0.9999

**Fig. 1** Average transition probabilities, CUB and NLCUB models

References

- Brentari, E., Carpita, M., Zuccolotto, P.: Qualità e Customer Satisfaction nei Servizi - Un'indagine statistica nelle Scuole dell'Infanzia del Comune di Brescia. FrancoAngeli, Milano (2006)
- D'Elia, A., Piccolo, D.: A mixture model for preference data analysis. *Comput. Stat. Data An.* **49**, 917–934 (2005)
- Iannario, M.: On the identifiability of a mixture model for ordinal data. *Metron* **LXVIII**, 87–94 (2010)
- Iannario, M., Piccolo, D.: CUB Models: Statistical Methods and Empirical Evidence. In: Kenett, R. S., Salini, S. (eds.) *Modern Analysis of Customer Surveys*, pp. 231–258. Wiley, NY (2012)
- Manisera, M., Zuccolotto, P.: Modelling rating data by Nonlinear CUB models. Submitted for publication (2013)
- Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. *Quad. Stat.* **5**, 85–104 (2003)
- Piccolo, D.: Observed information matrix for MUB models. *Quad. Stat.* **8**, 33–78 (2006)
- Piccolo, D., D'Elia, A.: A new approach for modelling consumers' preferences. *Food Qual. Prefer.* **19**, 247–259 (2008)

Finding number of groups using a penalized internal cluster quality index

Marica Manisera and Marika Vezzoli

Abstract In cluster analysis, the identification of number of groups is a non trivial question. Many papers investigated this issue and several criteria have been introduced. The objective of this study is to propose a new method that automatically identifies the optimal number of groups in a hierarchical cluster algorithm. Starting from the idea of pruning, introduced in the context of classification and regression trees, we propose to use a penalized internal cluster quality index in order to identify the best cut in the dendrogram able to provide a partition easily interpretable. In this paper, we show the results obtained by applying our procedure on simulated data with known structure.

Key words: hierarchical cluster analysis, internal cluster quality index, optimal number of clusters, penalized score function, pruning

1 Introduction

Identifying the optimal number of groups is of central importance in cluster analysis ([3]). Many authors handled this issue by exploring several criteria (among others, see [2], [4]), based on the trade-off between a low inter-cluster similarity and a high intra-cluster similarity. It is reasonable to expect that the optimal partition using this trade-off criterion is obtained when the number of groups equals the number of subjects analyzed. However, this type of partition is useless.

The objective of this study is to propose a new method that automatically identifies the optimal number of groups k^* in a hierarchical cluster analysis. In detail, we

Marica Manisera
University of Brescia, c.da S. Chiara, 50, 25122 Brescia, Italy, e-mail: manisera@eco.unibs.it

Marika Vezzoli
University of Brescia, viale Europa, 11, 25123 Brescia, Italy, e-mail: marika.vezzoli@med.unibs.it

want to overcome the subjective cutting of the dendrogram, which appears to be a common choice in practice.

We identify k^* by optimizing an internal cluster quality index, penalized by the number of clusters in order to take account of the interpretability of the resulting groups. This idea was inspired by the pruning ([1]), used in classification and regression trees as a method to avoid the overfitting problem that, in the extreme case, arises when each leaf of the tree contains only one subject. Indeed our idea is conceived along the same line and has the same aim: in a hierarchical cluster analysis, we grow the dendrogram by imposing a penalty depending on the k number of clusters so as to stop the procedure up to identify a reduced, and therefore interpretable, number of groups. In this study, we penalize the intrinsic index proposed in [2], suitable for quantitative data. However, a penalization can be proposed on alternative cluster quality indices, whenever they show a behaviour leading to choose k^* equal to the number of subjects in the analysis (for example, the error measure W_k in [4]).

The paper is organized as follows. Section 2 describes the proposed penalized internal quality index. Section 3 shows an illustrative example based on simulated data with known structure discussing results and future research.

2 Methodology

Starting from the $n \times p$ data matrix \mathbf{X} with n subjects and p quantitative variables, cluster analysis aims at partitioning subjects into k clusters. Many criteria identify the optimal number k^* of groups on the basis of the trade-off between a low inter-cluster similarity and a high intra-cluster similarity, where similarity is usually defined starting from a chosen distance function. In this study, we focus on the Calinski and Harabasz (CH) index ([2]), suitable for quantitative data, which measures the internal cluster quality for a given k as

$$\text{CH}(k) = \frac{\text{BGSS}/(k-1)}{\text{WGSS}/(n-k)}.$$

WGSS (Within-Group Sum of Squares) summarizes the intra-cluster similarity and is given by $\text{trace}(\mathbf{W})$, where \mathbf{W} is a $k \times k$ matrix whose generic element $\{w_{ht}\}_{h,t=1,\dots,k}$ is the distance of the subjects belonging to group h from the centroid \mathbf{c}_t of group t (\mathbf{c}_t is a p -dimensional vector containing the means $m_j^{(t)}$, $j = 1, \dots, p$, computed on subjects of group t). BGSS (Between-Group Sum of Squares) summarizes the inter-cluster similarity and is given by $(\text{trace}(n\Sigma) - \text{WGSS})$ where Σ is the variance-covariance matrix of \mathbf{X} . When Euclidean distance is used, the concepts of inter-cluster and intra-cluster similarities are related to the dispersion between and within the clusters in the analysis of variance. The best k is given by

$$k^* = \underset{k=2,\dots,n-1}{\text{argmax}} \text{CH}(k).$$

Whenever CH increases as k increases, the optimal partition is expected for $k^* = n$. However, this result is useless and does not comply with the aim of a cluster analysis. In order to identify an interpretable partition, k should be reasonably small and this is commonly achieved by subjective choices. In hierarchical clustering, that is the focus of this study, this corresponds to a subjective cutting of the dendrogram. In order to avoid such arbitrariness, we propose to identify k^* as:

$$k^* = \underset{k=2, \dots, n-1}{\operatorname{argmax}} Q(k|\lambda). \quad (1)$$

where $Q(k|\lambda) = \text{CH}(k) - \lambda \times k$ is obtained by introducing the penalty $\lambda \in \mathbb{R}_+$ on the number k of groups, in order to keep k^* reasonably small and find it automatically (equation (1) holds for internal cluster quality indices, alternative to CH, that have to be maximized and increase with k). If $\{0\}$ is included in the domain of λ , for $\lambda = 0$ we have $Q(k|\lambda) = \text{CH}(k)$ and no penalization is imposed. The larger the values of λ , the stronger the penalty (and *viceversa*). The effect of a fixed λ on k^* depends on the magnitude of the chosen cluster quality index.

3 An illustrative example

We applied the proposed procedure on an artificially generated data described in [5] and referred to 5 interval-type variables on 75 subjects clustered into 5 groups. We performed a hierarchical cluster analysis, using the `hclust` function in \mathbb{R} , with complete linkage. The cluster dendrogram, shown in Figure 1 (left), does not provide strong evidence for the simulated 5-class structure, which is instead confirmed by CH in [5], but only because the authors considered small k 's ($k = 2, \dots, 6$, see p. 12). We replicated their procedure computing CH for $k = 2, \dots, 50$, as shown in Figure 1 (right). CH increases as k increases and is maximized for $k = 50$ (see the black circle in the right part of Figure 1). However, if we want a partition with an interpretable number of groups, CH suggests $k^* = 8$. This choice requires the inspection of the dendrogram and a subjective reasoning by the researcher.

Instead, the use of the penalized CH index makes the choice automatic. Figure 2 (left) shows the penalized CH index computed for fixed $\lambda = 0.3$ and $k = 2, \dots, 50$. The penalization introduced in the cluster quality index automatically leads to the choice of the 8-class structure as the best solution (see the black circle in the left part of Figure 2). The maximization of the penalized CH index identifies 8 as the best number of groups for a wide range of fixed λ ($[0.154, 0.335]$), while outside that range the best number of groups is always 2 or 50, as shown in Figure 2 (right). Therefore, in this data set we can fairly choose $k^* = 8$.

Concluding, results obtained in this illustrative example show that the proposed procedure is able to reach the objective of automatically identifying the best number of clusters in a data set by taking account of the interpretability of the resulting groups. Current research is being devoted to refine the optimization algorithm, especially with reference to the choice of λ . Simulation studies and the analysis of

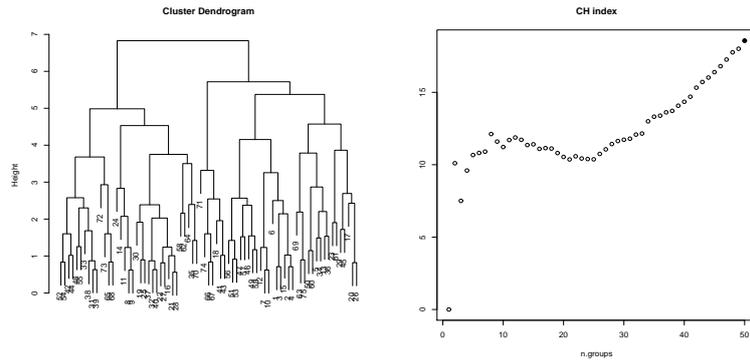


Fig. 1 Cluster dendrogram (left) and CH index for $k = 2, \dots, 50$ (right) - artificial data

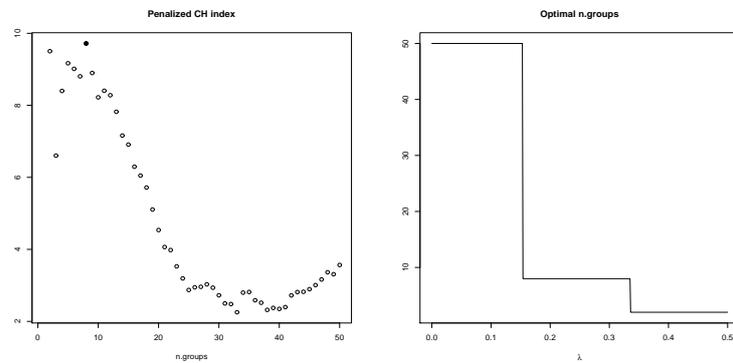


Fig. 2 Penalized CH index for $k = 2, \dots, 50$ and $\lambda = 0.3$ (left) and optimal number k^* of groups based on the penalized CH index for $\lambda = 0, \dots, 0.5$

real data sets, involving several internal cluster quality indices suitable for different data types, could confirm the validity of our proposal.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group, Belmont(1984)
2. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27 (1974)
3. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, Chichester (2011)
4. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of data clusters via the Gap statistic. *J. Roy. Stat. Soc. B* **63**, 411-423 (2001)
5. Walesiak, M., Dudek, A.: clusterSim: Searching for optimal clustering procedure for a data set. R package version 0.41-8. <http://CRAN.R-project.org/package=clusterSim> (2012)

Object-Oriented Bayesian Network to deal with measurement error in household surveys

Daniela Marella and Paola Vicard

Abstract In this paper we propose to use the Object-Oriented Bayesian Networks architecture to model measurement errors in the Italian Survey on Household Income and Wealth (SHIW) 2008 when the variable of interest is categorical. The network implemented is used to stochastically impute micro data for households. Potentialities and possible extensions of this approach are discussed.

Key words: categorical variable, mixed measurement model, underreporting.

1 Introduction

Measurement error is the difference between the value of a feature provided by the respondent and the corresponding true but unknown value. Together with nonresponse, measurement error is one of the main nonsampling error source. In fact, the presence of measurement errors may severely affect the quality of survey results leading to erroneous conclusions.

Object-oriented Bayesian networks (OOBNs) have been recently proposed as a new tool by which modelling and correcting measurement errors. In particular, the measurement error in a categorical variable is described by a mixed measurement model implemented in an Bayesian Network (BN); for details see [2]. The aim of the paper is to apply this model to 2008 Survey on Household Income and Wealth (SHIW). SHIW is conducted by Banca d'Italia every two years. Its main objective is

Daniela Marella
Dipartimento di Scienze della Formazione, Via del Castro Pretorio 20, 00185 Roma e-mail:
daniela.marella@uniroma3.it

Paola Vicard
Dipartimento di Economia, Via Silvio D'Amico 77, 00145 Roma e-mail:
paola.vicard@uniroma3.it

to study the economic behaviors of Italian households. Financial assets in SHIW are affected by misreporting of financial amounts with a prevalence of underreporting.

A two step strategy has been followed: first of all the measurement model parameters have been estimated using a validation sample; secondly the estimated model implemented in a BN has been used to impute micro data for units in SHIW 2008. More specifically, for each respondent the evidence (*i.e.* observed value in SHIW 2008) is inserted and propagated throughout the network to estimate the probability distribution of the true value given the observed one. The individual true value can then be predicted by a random draw from such a distribution.

The paper is organized as follows. In section 2 the measurement error model is described and the performance of the imputation procedure is evaluated. In Section 3, potentialities and possible extensions of our approach for dealing with measurement error in sample surveys for continuous variables are discussed.

2 An application to SHIW 2008

Let X be an ordered categorical variable with K response categories whose frequencies p_k , $k = 1, \dots, K$, are assumed known. When a measurement error takes place the observed category is different from the true category. Let $q_{i \rightarrow j}$ be the intercategory transition probability from the true category i to the observed category j , where $\sum_{j=1}^K q_{i \rightarrow j} = 1$. In order to estimate the $K(K-1)$ probabilities $q_{i \rightarrow j}$, we could carry out an interview-reinterview study. An alternative way to proceed is to express the transition probabilities $q_{i \rightarrow j}$ by means of models characterized by a smaller number of parameters to be estimated. A realistic and plausible representation of the measurement error generating process is the mixed measurement model,

$$s_{i \rightarrow j}^{mix} = (1 - h)s_{i \rightarrow j}^{prop} + hs_{i \rightarrow j}^{MMT} \quad (1)$$

given by a mixture of the proportional model $s_{i \rightarrow j}^{prop}$ and the one-T step model $s_{i \rightarrow j}^{MMT}$. The model $s_{i \rightarrow j}^{prop}$ reflects the assumption that, whenever a measurement error occurs, the observed value j is generated at random from the population frequency distribution, *i.e.* regardless of the true value i . The model $s_{i \rightarrow j}^{MMT}$ implies that the observed category j can only be a neighbouring category or a category up to T steps away from the true category i .

The model (1) is completely general, and can be applied to various contexts thanks to the measurement model parameters (h, μ, α_t) , where h is the mixture parameter, μ is the misreport probability and α_t is the probability that the difference between the observed and the true category is t , for $|t| = 1, \dots, T$. This model has been implemented using the OOBN architecture. For details and graphical model representation, see [2].

Here we apply the mixed measurement model (1) to SHIW 2008. A validation sample has been used to estimate the measurement model parameters. Data have been collected through an independent experiment survey carried out by Banca

d'Italia and a major Italian bank group on a sample of customers of the latter. The survey was carried out in 2003 on a sample of 1.681 households where at least one member was a customer of the bank group. Survey data had then been matched with the bank customers database containing the amount of the assets actually held by the individuals selected in the sample. The resulting dataset will be referred to as our *validation sample*, see [3] for details.

In our analysis we focus on bonds amount (specifically government and private bonds). Since model (1) can be applied to categorical variables, the true and the observed amount of bonds in the validation sample have been discretized. First of all, the measurement model parameters μ and α_t (for $t = -7, \dots, 3$) have been estimated. The estimates are reported in Table 1.

Table 1 *Estimates of measurement model parameters*

μ	α_{-7}	α_{-6}	α_{-5}	α_{-4}	α_{-3}	α_{-2}	α_{-1}	α_1	α_2	α_3
0.54	0.07	0.07	0.08	0.15	0.12	0.13	0.16	0.11	0.06	0.05

The value of the mixture parameter $h = 0.9$ has been evaluated through a sensitivity analysis. Under the assumption that the underreporting behavior observed in the validation sample remained unchanged in 2008, the estimated model suitably implemented in a BN is used to impute the (discretized) amount of bonds in SHIW 2008.

In order to evaluate the performance of the imputation procedure the following evaluation criteria have been used:

1. the Kullback-Leibler distance between the true distribution and the observed and the imputed distribution denoted by KL^{TI} and KL^{TO} , respectively;
2. the percentage of correct imputations given by

$$\psi = \frac{1}{n^*} \sum_{i \in S^*} I_{x_i}(x_i^*) * 100 \quad (2)$$

where S^* is the subsample composed of units affected by measurement errors, I_{x_i} is the indicator function assuming the values 1 if the true value for unit i denoted by x_i is equal to the corresponding imputed value x_i^* and 0 otherwise.

This imputation procedure reveals a good performance when applied to categorical or previously discretized variables. The distance between the true and the imputed distribution $KL^{TI} = 0.13$ is less than the distance between the true and the observed distribution $KL^{TO} = 0.47$. Furthermore, the proposed imputation method is able to correctly reconstruct 11% of data affected by measurement error in the validation sample. It is known that results could be improved defining household profiles as unique combinations of covariate values when adequate sample sizes for each profile are available. Unfortunately, this is not the case in our application because of the small size of the validation sample.

The good performance of imputation procedure is not necessarily guaranteed if we want to retrieve the continuous value from the imputed category using the intervals defined by the discretization process, since the continuous imputed value does not necessarily improve the observed value. This is mainly due to the trade-off between the number of intervals defined by the discretization on one hand, and the accuracy of parameters estimates and the imputation classes sizes on the other hand. Then alternative procedures must be investigated for dealing with continuous variables.

3 Potentialities and limitations of BNs when dealing with measurement errors in sample surveys

We believe that BNs are an important and promising tool to deal with measurement error in sample surveys. However, at the moment, they do have some limitations that may complicate their application. The two main problems regard: (i) the use of hybrid BNs where continuous and discrete variables are considered; (ii) the necessity to take into account the complexity of sampling design when BNs are applied to sample surveys.

Regarding the first point, it is well known that mixtures of Gaussian distributions can approximate any probability distribution. Then it should be possible to solve any hybrid BN by first approximating it by a mixture of Gaussian BNs and then using the Lauritzen algorithm [1] to solve this mixture, see [4].

As far as the second point is concerned, design complexity must be taken into account in survey analysis by an appropriate use of sampling weights in order to obtain unbiased estimates. Furthermore, when auxiliary variables are used to improve either parameter estimates or imputation procedure, BNs should be learnt suitably accounting for the sampling design features.

In spite of these challenging problems we believe that, in a world where large amount of data are increasingly available, BNs are particularly promising and appealing since they define a common language to computer scientists and statisticians, *i.e.* those managing large amount of data and those making inference on them.

References

1. Lauritzen, S. L.: Propagation of probabilities, means and variances in mixed graphical association models. *Journal of American Statistical Association*, **87**, 1098–1108 (1992)
2. Marella, D., Vicard, P.: Object-Oriented bayesian network for modelling the respondent measurement error. *Communications in Statistics: Theory and Methods*. To appear.
3. Neri, A., Ranalli, M.G.: To misreport or not to report? The case of the Italian Survey on Household Income and Wealth. *Statistics in Transition*, **12**, 281–300 (2011).
4. Shenoy, P.P.: Inference in Hybrid Bayesian Networks using mixtures of Gaussians. UAI (2006).

Beyond tandem analysis: joint dimension reduction and clustering in R

Angelos Markos, Alfonso Iodice D'Enza and Michel Van de Velden

Abstract We describe a class of methods that use a combination of dimension reduction and clustering. In particular, two methods for quantitative data and three for qualitative data are reviewed that will be included in an R package. A unified framework is used to underline differences among the methods in question and facilitate their comparison.

Key words: dimension reduction, clustering

1 Introduction

There exist several methods for clustering high-dimensional data. One popular approach is to use a combination of dimension reduction and clustering methods. A sequential application of dimension reduction and clustering is referred to as *tandem analysis* proposed by [1]. This approach is implemented in R packages such as `FactoMineR` [6] and `FactoClass` [7]. However, more sophisticated methods combining dimension reduction and clustering have been proposed in the literature, for both quantitative and qualitative data cases. The methods in question are, for quantitative data, the Factor-Kmeans and the Reduced K-means, proposed by [9] and [2], respectively; for qualitative data methods have been proposed by [4], [8] and [5]. In this paper we describe a re-formulated version of extant joint dimension reduction and clustering methods implemented in a R package.

The paper is structured as follows: Section 2 introduces the input and output quantities and the data structures; Section 3 introduces the functions imple-

Angelos Markos

Department of Primary Education, Democritus University of Thrace, Greece, e-mail: amarkos@eled.duth.gr

Alfonso Iodice D'Enza

Department of Economics and Law Program, Università di Cassino e del Lazio Meridionale, Italy e-mail: iodicede@unicas.it

Michel Van de Velden

Econometrics Department, Erasmus University of Rotterdam, Netherlands, e-mail: vandevelde@ese.eur.nl

menting the proposed quantitative methods and the corresponding optimization criteria; similarly, in Section 4 the functions for qualitative methods are introduced.

2 Notation and I/O quantities

Consider n observations and q variables. In case of quantitative data, the $n \times q$ reference matrix is \mathbf{X} : the q variables are supposed to be centered or standardized. In case of qualitative data, let \mathbf{Z}_j denote an $n \times p_j$ indicator matrix, with p_j the number of categories for the j th variable with $j = 1, \dots, q$ categorical variables. The general matrix $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_j, \dots, \mathbf{Z}_q]$ is an $n \times Q$ block matrix, where $Q = \sum_{j=1}^q p_j$. Let \mathbf{C} be an $n \times K$ indicator matrix that assigns each statistical unit to one of the K groups. We refer to d as the dimensionality of the solution, to \mathbf{Y} as the $n \times d$ matrix of unit coordinates in the low-dimensional space and to \mathbf{B} as the variable coordinates: the size of \mathbf{B} is $q \times d$ for the quantitative case, and $Q \times d$ for the qualitative case. Of course, the computation of \mathbf{Y} and \mathbf{B} varies with the nature of the data being analysed as well as with the considered method. Finally, \mathbf{G} is a $K \times d$ centroid matrix that contains the scores of the cluster centroids on a reduced number of components. Further method-specific structures will be defined along with the presentation of the methods.

Table 1 contains a list of input and output quantities of the functions described in Sections 3 and 4. The input quantities are common to all functions, with the obvious exception of `data` that can be quantitative or qualitative. The cluster membership vector aside, the output quantities depend on the chosen function, and will be defined in the following sections.

input	description	output	description
<code>data</code>	dataset	<code>cluID</code>	cluster membership
<code>nclus</code>	number of clusters	<code>obscoord, attcoord</code>	low-dimensional coordinates
<code>ndim</code>	dimensionality of the solution	<code>criterion</code>	optimal value
<code>nstart</code>	number of random starts		

Table 1 Input and output quantities, common to all the implemented functions

3 Functions for quantitative data

For quantitative data, two functions are available, FKM and RKM. In particular the function FKM implements Factorial K-means [9], that simultaneously classifies observations and finds a subset of factors that best describe the

classification. The function `RKM` implements Reduced K-means [2], that can be seen as a constrained k-means where the centroids are forced to lie in a subspace of reduced dimension.

The output quantities of `FKM` are defined as follows:

- **criterion:** the factorial K-means is defined by the minimization of the following loss function [9]:

$$\min_{\mathbf{B}, \mathbf{C}, \mathbf{G}} \phi(\mathbf{B}, \mathbf{C}, \mathbf{G}) = \|\mathbf{XBB}^\top - \mathbf{CGB}^\top\|^2 = \|\mathbf{XB} - \mathbf{CG}\|^2 = \|\mathbf{Y} - \mathbf{CG}\|^2 \quad (1)$$

$$\text{s.t. } \mathbf{B}^\top \mathbf{B} = \mathbf{I}.$$

- **obscoord:** \mathbf{Y} ; **attcoord:** \mathbf{B}

The output quantities of `RKM` are defined as follows:

- **criterion:** the reduced K-means minimizes the following loss function [2]:

$$\min_{\mathbf{B}, \mathbf{C}, \mathbf{G}} \phi(\mathbf{B}, \mathbf{C}, \mathbf{G}) = \|\mathbf{X} - \mathbf{CGB}^\top\|^2 \text{ s.t. } \mathbf{B}^\top \mathbf{B} = \mathbf{I}. \quad (2)$$

- **obscoord:** $\mathbf{Y} = \mathbf{XB}$; **attcoord:** \mathbf{B}

4 Functions for qualitative data

For qualitative data, three functions are available `MCAk`, `iFCB` and `GROUPALS`. In particular, the function `MCAk` implements the joint MCA and K-means clustering approach, proposed by Hwang *et al.* [4]; the function `iFCB` implements the i-FCB method [5], that combines dimension reduction and clustering for binary data, however it can also be used for categorical data with more than two modalities per variable. The function `GROUPALS` implements the `GROUPALS` method [8], that consists of a combination of optimal scaling (homogeneity analysis) and clustering techniques. The output quantities of the `MCAk` function are:

- **criterion:** the target function being optimized is

$$\min_{\mathbf{Y}, \mathbf{B}, \mathbf{C}, \mathbf{G}} \phi(\mathbf{Y}, \mathbf{B}, \mathbf{C}, \mathbf{G}) = \alpha_1 \sum_{j=1}^q \|\mathbf{Y} - \mathbf{Z}_j \mathbf{B}_j\|^2 + \alpha_2 \|\mathbf{Y} - \mathbf{CG}\|$$

$$\text{s.t. } \mathbf{Y}^\top \mathbf{Y} = \mathbf{I}_d$$

which is a weighted sum of the MCA's homogeneity criterion ([3]) and of the K-means objective.

- **obscoord**: \mathbf{Y} ; **attcoord**: $[\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_q]$, where $\mathbf{B}_j = (\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j^\top \mathbf{Y}$, $1, \dots, q$, contains the category quantifications of the j^{th} qualitative variable.

The output quantities of the iFCB function are:

- **criterion**: the solution is achieved by optimizing the following quantity

$$\max_{\mathbf{C}} \phi(\mathbf{C}) = \text{tr} \left[\frac{1}{n} \mathbf{C}^\top \mathbf{Z} (\mathbf{D}_z)^{-1} \mathbf{Z}^\top \mathbf{C} - \frac{p}{n^2} (\mathbf{C}^\top \mathbf{1} \mathbf{1}^\top \mathbf{C}) \right] \quad (3)$$

where $\mathbf{D}_z = \text{diag}(\mathbf{Z}^\top \mathbf{Z})$ and $\mathbf{1}$ is an n -dimensional vector of ones.

- **obscoord**: $\mathbf{Y} = \mathbf{D}_w \mathbf{Z} \mathbf{D}_z \mathbf{B} \boldsymbol{\Sigma}^{-1}$, with $\mathbf{D}_w = \text{diag}(\mathbf{C} \mathbf{w})$ and \mathbf{w} is the group size vector; \mathbf{U} , \mathbf{V} and $\boldsymbol{\Sigma}$ being the left and right singular vectors and values of the standardized residual version of matrix $\mathbf{P} = \mathbf{F}(nq)^{-1}$; **attcoord**: $\mathbf{B} = \mathbf{D}_Q^{-1/2} \mathbf{V} \boldsymbol{\Sigma}$.

The output quantities of the GROUPALS function are:

- **criterion**: the optimized loss function is

$$\min_{\mathbf{B}, \mathbf{C}, \mathbf{G}} \phi(\mathbf{B}, \mathbf{C}, \mathbf{G}) = \frac{1}{q} \sum_{j=1}^q \|\mathbf{C} \mathbf{G} - \mathbf{Z}_j \mathbf{B}_j\|^2 \quad (4)$$

- **obscoord**: $\mathbf{Y} = \mathbf{C} \mathbf{G}$; **attcoord**: \mathbf{B} .

The package also contains functions for ggplot2-based visualization and for measuring the quality of the solutions.

References

1. Arabie, P. & Hubert, L. (1994). 'Cluster analysis in marketing research'. *IEEE T Automat Contr*, 19: 716–723
2. De Soete, G. & Carroll, J. D.: K-means clustering in a low-dimensional Euclidean space. In Diday E. et al. (Eds.). *New Approaches in Classification and Data Analysis* (pp. 212–219). Heidelberg: Springer (1994)
3. Gifi, A.: *Nonlinear multivariate analysis*. John Wiley & Sons, NY (1990)
4. Hwang, H., Dillon, W. R. & Takane, Y.: An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. *Psychometrika* 71: 161–171 (2006)
5. Iodice D'Enza, A. & Palumbo, F.: Iterative factor clustering of binary data. *Computation Stat* 28(2), 789–807 (2013)
6. Lê, S., Josse, J. & Husson, F.: FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, 25 (1), 1–18 (2008)
7. Pardo, C. E., & Del Campo, P. C. Combination of Factorial Methods and Cluster Analysis in R: The Package FactoClass. *Revista Colombiana de Estadística*, 30 (2), 231–245, (2007)
8. Van Buuren, S. & Heiser, W. J.: Clustering n objects in k groups under optimal scaling of variables. *Psychometrika* 54, 699–706 (1989)
9. Vichi, M. & Kiers, H.: Factorial k-means analysis for two-way data. *Comput Stat Data An* 37(1), 49–64 (2001)

A biclustering approach for discrete outcomes

F. Martella and M. Alfò

Abstract We discuss an extension of mixtures of factor analyzers (MFA) to allow for simultaneous clustering of subjects and variables where discrete manifest variables are available. To estimate model parameters, we propose a modified EM algorithm in a ML framework.

Key words: Model-based biclustering, discrete data

1 Introduction

Biclustering - known under a broad range of names, including double clustering, block clustering, bidimensional clustering, co-clustering, simultaneous clustering and blockmodeling - may be thought as a two-dimensional extension of standard clustering approaches. In a typical clustering approach, a set of objects is considered and each object should be assigned to a cluster optimizing a given criterion. Objects in the same cluster are considered “homogeneous” with respect to some measured features. If the interest is in clustering two sets of objects from a given data matrix (usually units and variables), a clustering method can be applied to both sets of objects successively and/or independently (see [1]). However, results depend on which set is classified first. To overcome this problem, and to take into account association between the two partitions, [2] proposed to partition units and variables simultaneously rather than successively. The most important advantage of biclustering is that it allows to highlight the interaction between the two sets of objects and

F. Martella

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le Aldo Moro 5, 00185 Rome (Italy), e-mail: francesca.martella@uniroma1.it

M. Alfò

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le Aldo Moro 5, 00185 Rome (Italy), e-mail: marco.alfò@uniroma1.it

help in their characterization by using an “overall” objective function that cannot be reduced to a simple combination of the two objective functions. During the past three decades, starting with the pioneering work of [3] and with decision-theoretic work by [4], this class of methods has been widely developed by various authors in different fields such as marketing, customer satisfaction, social network, psychology, text mining, election nutritional analyses and microarray studies. The interested reader can refer to [5] and [6] for a comprehensive review of biclustering.

Mixture-based (sometimes referred to as model-based) biclustering approaches have been developed by [7], who proposed a block mixture model for binary data (further extended in [8]), and [9], who proposed a two-way Poisson mixture model for text analysis. Recently, [10] and [11] have extended the Mixture of Factor Analyzers model (MFA, see [12]) to allow for biclustering of genes and tissue samples.

In the present paper, we address the issues of applying biclustering in the context of multivariate binary and count responses highlighting potential applications to mixed type responses.

2 Model-based biclustering for continuous data

Recently, [10] have introduced a biclustering model, where the data density is approximate by a mixture of Gaussian distributions with a particular component-specific covariance structure. More precisely, they propose to use a binary and row stochastic matrix representing a column partition, i.e. a partition of variables, whereas the traditional mixture approach is used to define a partition of subjects. In details, let \mathbf{y}_i be a J -dimensional vector representing J quantitative variables measured on the i -th subject ($i = 1, \dots, n$) and assume that the observed sample is drawn from a population \mathcal{P} partitioned in K clusters $\mathcal{P}_1, \dots, \mathcal{P}_K$, with $\mathcal{P} = \bigcup_{k=1}^K \mathcal{P}_k$, $\mathcal{P}_k \cap \mathcal{P}_{k'} = \emptyset$, $k \neq k' = 1, \dots, K$. The k -th cluster has prior probability $\pi_k = Pr(\mathbf{y}_i \in \mathcal{P}_k)$, conditional on belonging to the k -th cluster \mathbf{y}_i is specified by a factor analysis model as follows:

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{u}_{ik} + \mathbf{e}_{ik} \quad (1)$$

where $\boldsymbol{\mu}_k$ is the J -dimensional component-specific mean vector, $\mathbf{V}_k = \{v_{jl}\}$ ($j = 1, \dots, J$, $l = 1, \dots, Q_k$, $k = 1, \dots, K$) is a binary row stochastic matrix representing variable cluster membership. Here $v_{jl} = 1$ if and only if the j -th variable belongs to the l -th column cluster, and 0 otherwise, \mathbf{u}_{ik} is a Q_k -dimensional ($Q_k < J$) vector of component-specific latent variables (factors), which are assumed to be i.i.d. draws from $N(\mathbf{0}, \mathbf{I}_{Q_k})$, and \mathbf{I}_{Q_k} denotes the $Q_k \times Q_k$ identity matrix. Last, \mathbf{e}_{ik} are i.i.d. Gaussian component-specific random variables with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}_k = \text{diag}(\sigma_{1k}^2, \dots, \sigma_{jk}^2)$, that are assumed to be independent of \mathbf{u}_{ik} . The model parameters are estimated through a maximum likelihood approach by using an Alternating Expectation Conditional Maximization (AECM) algorithm (see e.g. [13]).

3 Extension to discrete data

In line with the previous issues, we propose a model-based biclustering approach developed by [10] for discrete data. Specifically, we turn our attention to (i) binary data, i.e. when the response represent success and failure, or more generally the presence or absence of an attribute of interest; (ii) count data, i.e. the number of events recorded in a given time interval, with known or unknown index (maximum count).

3.1 Binary data

Let Y_{ij} be a binary variable representing the response of the i -th subject to the j -th test item. In this case, $Y_{ij} = 1$ where the i -th subject provides a correct answer to the j -th item, 0 otherwise. Responses y_{ij} 's are assumed to be (conditionally) independent Binomial random variables within the k -th cluster (local independence) and, thus, it follows that the probability of the response pattern \mathbf{y}_i within the k -th cluster, $f(\mathbf{y}_i|k)$, is expressed by:

$$f(\mathbf{y}_i|k) = \prod_{j=1}^J \theta_{j|k}^{y_{ij}} (1 - \theta_{j|k})^{1-y_{ij}}. \quad (2)$$

where $\theta_{j|k}$ indicates the conditional probability of success for a subject in the k -th cluster, i.e. the probability that subjects from the k -th cluster gives a correct answer to the j -th item, that is:

$$\theta_{j|k} = Pr(Y_j = 1 | \forall i \in P_k).$$

In line with the idea of [10], to introduce an item-specific partition, conditional on the k -th cluster with probability π_k , we assume that the conditional probabilities of success $\text{logit}(\theta_{j|k})$ may be reparametrized as follows:

$$\text{logit}(\theta_{j|k}) = (\phi_k + \mathbf{a}'_{kj}\boldsymbol{\beta}), \quad j = 1, \dots, J, \quad k = 1, \dots, K. \quad (3)$$

where ϕ_k is the measure of the component-specific latent trait level, e.g. it may be interpreted as the ability measured by the items, $\pi_k = Pr(i \in P_k) = Pr(\phi = \phi_k)$, $k = 1, \dots, K$. In this parametrization, \mathbf{a}_{kj} is a Q -dimensional component-specific row stochastic vector ($Q < J$) representing item cluster membership, i.e. $a_{kj} = 1$ if and only if the j -th item belongs to the q -th item cluster, and 0 otherwise while $\boldsymbol{\beta} = (\beta_1, \dots, \beta_Q)$ is a Q -dimensional fixed random vector, which is constant across components and, may be interpreted, in an educational setting context, as a common measure of simplicity shared by a subset of the J items. Notice that, in the Rasch model context, it's common to use the negative sign in equation (3) and $\boldsymbol{\beta}$ is interpret as difficulty/complexity parameters. Thus, the marginal density of \mathbf{y}_i is

$$f(\mathbf{y}_i) = \sum_{k=1}^K \pi_k \prod_{j=1}^J \theta_{j|k}^{y_{ij}} (1 - \theta_{j|k})^{1-y_{ij}}. \quad (4)$$

Notice that, differently from the work proposed by [7], the variable partition is conditioned by the unit partition and that we implicitly assume that (i) the population under study is made up by a finite number of clusters, with subjects in the same cluster sharing the same ability level; (ii) the ability has a discrete distribution (more flexible than a continuous one) with support $\{\phi_1, \dots, \phi_K\}$; (iii) the assumption of unidimensionality, according to which all the items measure the same component-specific ability, holds; (iv) all items discriminate equally between subjects. On the basis of these assumptions, we simultaneously perform clustering of subjects and items, so that items in the same item-specific cluster share the same difficulty level. The proposed model may be related to several proposal in the literature, mainly focused on unidimensionality and clustering of items in IRT models (see [14], [15], [16]).

3.2 Count data

Suppose that Y_{ij} represent the number of events of the j -th type for the i -th subject, which takes nonnegative values. In this context, we may assume that responses y_{ij} 's are (conditionally) independent Poisson random variables within the k -th cluster and, thus, the conditional probability of \mathbf{y}_i becomes:

$$f(\mathbf{y}_i|k) = \prod_{j=1}^J \frac{e^{-\theta_{j|k}} \theta_{j|k}^{y_{ij}}}{y_{ij}!}. \quad (5)$$

where $\theta_{j|k}$ indicates the component-specific Poisson parameter. Here, to introduce variable clustering, conditional on the k -th cluster with probability π_k , we assume that $\log(\theta_{j|k})$ may be reparametrized as follows:

$$\log(\theta_{j|k}) = (\phi_j + \mathbf{a}'_{kj}\boldsymbol{\beta}), \quad j = 1, \dots, J, \quad k = 1, \dots, K, \quad i = 1, \dots, n. \quad (6)$$

where ϕ_j is a variable-specific intercept which determines the general scale for the j -th variable, it follows that $\pi_k = Pr(i \in P_k) = Pr(\phi = \phi_k)$, $k = 1, \dots, K$, the departures from this value are determined by the column (count) partitions through the influence of the Q -dimensional component-specific row stochastic vector ($Q < J$) representing variable cluster membership \mathbf{a}_{kj} defined as above through $\boldsymbol{\beta} = (\beta_1, \dots, \beta_Q)$. In this case, expression (4) becomes:

$$f(\mathbf{y}_i) = \sum_{k=1}^K \pi_k \prod_{j=1}^J \frac{e^{-\theta_{j|k}} \theta_{j|k}^{y_{ij}}}{y_{ij}!}. \quad (7)$$

3.3 ML parameter estimation

Regardless of the type of discrete data, the log-likelihood function based on n independent observations takes the following form:

$$l(\Xi) = \sum_{i=1}^n \log f(\mathbf{y}_i) = \sum_{i=1}^n \log \sum_{k=1}^K \pi_k \prod_{j=1}^J f(y_{ij}|k) \quad (8)$$

where Ξ is the vector containing all model parameters and $f(\mathbf{y}_i)$ is defined as in equation (4) with reparametrization given by expression (3) for binary data, and in equation (7) with the reparametrization in expression (6) for count data. Parameter estimation is performed via a modified EM type algorithm.

References

1. Tryon, R.C.: Cluster Analysis, Edwards Brothers (1939).
2. Fisher, W.: Clustering and aggregation in economics. Baltimore: Johns Hopkins, (1969).
3. Hartigan, J.A.: Direct clustering of a data matrix. *Journal of the American Statistical Association* **67**, 123–129 (1972).
4. Bock, H.H.: Automatische Klassifikation, Vandenhoeck and Ruprecht, Göttingen, (1974).
5. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**, 24–45 (2004).
6. Van Mechelen, I., Schepers, J.: A unifying model for biclustering. *Compstat 2006 - Proceedings in Computational Statistics*, 81–88 (2006).
7. Govaert, G., Nadif, M.: Clustering with block mixture models. *Pattern Recognition*, **36**(2), 463–473 (2003).
8. Govaert, G., Nadif, M.: Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, **52**, 3233–3245, (2008).
9. Li, J., Zha, H.: Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics and Data Analysis*, **50**(1), 163–180 (2006).
10. Martella, F., Alfò, M., Vichi, M.: Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics*, **4**(1), 3 (2008).
11. Martella, F., Alfò, M., Vichi, M.: Hierarchical mixture models for biclustering in microarray data. *Statistical Modelling*, **11**(6): 489-505 (2011).
12. Ghahramani, Z., Hinton, G.E.: The EM algorithm for mixture of factor analyzers. Technical Report, CRG-TR-96-1, 8, University of Toronto, (1997).
13. Meng, X.L., Van Dyk, D.A.: The EM algorithm -an old folk song sung to a fast new tune. Reading paper in *Journal of the Royal Statistical Society: Serie B (Statistical Methodology)*, **59**, 511–567 (1997).
14. Bartolucci, F.: A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, **72**, 2, 141–157 (2007).
15. Bartolucci, F., Montanari, G.E., Pandolfi, S.: Dimensionality of the latent structure and item selection via latent class multidimensional IRT models. *Psychometrika*, **77**, 4, 782–802 (2012).
16. Debelak, R., Arendasy, M.: An Algorithm for Testing Unidimensionality and Clustering Items in Rasch Measurement. *Educational and Psychological Measurement*, **72**, 375–387 (2012).

A Multidimensional IRT approach to analyze learning achievement of Italian students

Mariagiulia Matteucci, Stefania Mignani, Roberto Ricci

Abstract In the field of educational measurement, it often happens that tests consist of different subscales involving explicitly several dimensions. With the aim of estimating the test item psychometric properties and the candidate ability, item response theory (IRT) models are often used with multidimensional structure. In this work, a comparison among multidimensional models with different ability structures is conducted in a case study. In particular, data from Italian student assessments conducted at the end of the lower secondary school by INVALSI are taken into account.

1 Introduction

In educational settings, the use of standardized tests to assess students' learning nationally and across countries is widespread, as demonstrated by large-scale international assessments such as PISA, TIMSS, and PIRLS. Large-scale tests are under strictly standardized conditions to provide reliable and comparable scores for individuals. In Italy, the use of standardized assessment has assumed an increasing importance only recently, thanks to the annual surveys conducted by the National Evaluation Institute for the School System (INVALSI) at different school grades. The INVALSI develops tests to assess pupils' reading comprehension, grammar knowledge and mathematics competency, and administers them to the whole population of primary school students (second and fifth grade), lower secondary school students (sixth and eighth grade), and upper secondary school students (tenth grade).

Mariagiulia Matteucci

Department of Statistical Sciences, University of Bologna m.matteucci@unibo.it

Stefania Mignani

Department of Statistical Sciences, University of Bologna, stefania.mignani@unibo.it

Roberto Ricci

Evaluation Institute for the School System, roberto.ricci@INVALSI.it

The assessment of a competence is a process involving several steps. Firstly, it is fundamental to identify learning objectives, secondly the design for assessment is arranged to detect how a student might demonstrate that achieved a particular learning objective. To this purpose it is fundamental the development of a proper measurement instrument, typically a test containing a set of items related to specific content domains. With the aim of estimating the test item psychometric properties and the candidate ability, item response theory (IRT) models are often used. The focus of IRT is on the specification of the relationship between item psychometric properties (such as difficulty and discrimination) and the latent, non-observable, trait (ability) measured by the candidate's responses. IRT models express the probability of an item response as a function of the latent variable and the item properties.

In most of the applications IRT models are applied under the assumption of unidimensionality, i.e. the presence of a one predominant latent ability however when a test consists of different subscales, it may be more appropriate to apply a multidimensional model with a more complex structure underlying the response process (Reckase, 2009; Sheng and Wikle, 2007, 2008, 2009).

In this work, a comparison among multidimensional models with different ability structures is conducted in a case study. Data from Italian student assessments conducted at the end of the lower secondary school by INVALSI are analysed. The aim of the test is to evaluate competencies in language and mathematics, giving a final score to each student. This final score is calculated as the arithmetic mean of the scores obtained in the two subtests, by assuming unidimensionality within each subtest. However, the Italian language test is further divided into two subscales (reading comprehension and grammar) while the mathematics test contains rather heterogeneous items belonging to different domains. For these reasons, the study of dimensionality is very important to interpret correctly the test structure and to estimate test scores reflecting the presence of different ability dimensions. Our work represents the first attempt to model this complex structure with a multidimensional approach and to determine an overall score for each student based on the whole test.

2 Multidimensional IRT models

Within the multidimensional models, different approaches are possible: explorative models where all latent traits are allowed to load on all item response variables or confirmative models. Another distinction is among non compensatory and compensatory models, where a lack in one ability naturally compensates for the other (Reckase, 2009). Finally, models can be distinguished on the basis of the number of item parameters, on the number of response categories (binary or polytomous) and on the presence or not of a general trait, besides the specific ones.

By adopting a confirmatory approach, it is also possible to assume the concurrent presence of general and specific latent traits underlying the response process (Sheng and Wikle, 2008). In this work we consider two-parameter normal ogive (2PNO) models for binary data in a confirmatory approach and assume that a test consisting of k items divided into m subtests each containing k_v items, where $v=1,..m$. We estimate the hierarchical and the additive models. The additive model assumes that the general ability directly affects the candidate's responses, and that this effect is summed to the

effect of specific factors in order to determine the probability of success to a given test item, as follows

$$P(Y_{vij} = 1 | \theta_{0i}, \theta_{vi}, \alpha_{0vj}, \alpha_{vj}, \delta_{vj}) = \Phi(\alpha_{0vj}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_{vj}) = \int_{-\infty}^{\alpha_{0vj}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_{vj}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \quad (1)$$

where $\theta_i \sim N_{m+1}(\mathbf{0}, \mathbf{P})$ and \mathbf{P} is the ability correlation matrix, with $i=1, \dots, n$ subjects. The model involves the estimation of a general and a specific discrimination parameter α_{0vj} and α_{vj} , respectively, and a threshold (or difficulty) parameter δ_{vj} for each item j . Moreover, for each subject i , an overall ability θ_{0i} and m specific abilities θ_{vi} are estimated.

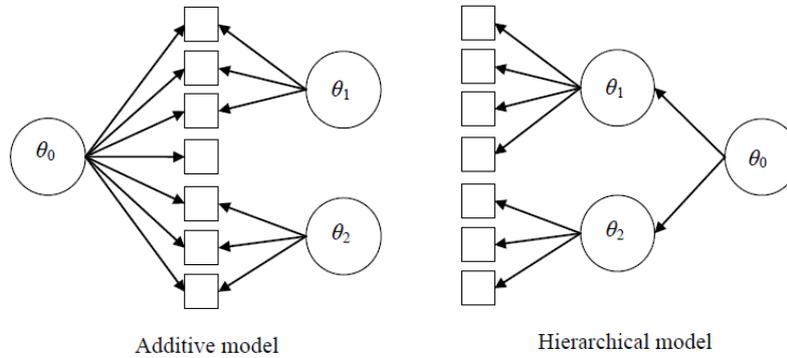
The hierarchical model assumes that each specific ability is a linear function of the general ability, as follows

$$P(Y_{vij} = 1 | \theta_{vi}, \alpha_{vj}, \delta_{vj}) = \Phi(\alpha_{vj}\theta_{vi} - \delta_{vj}) = \int_{-\infty}^{\alpha_{vj}\theta_{vi} - \delta_{vj}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \quad (2)$$

where $\theta_{vi} \sim N(\beta_v \theta_{0i}, 1)$.

A simple example is represented in Figure 1, where abilities are indicated by circles and items by squares. In the additive model, all abilities are assumed to be correlated.

Figure 1: Graphical representation of the additive and the hierarchical models



For these models, the joint estimation of item parameters and individual abilities is rather complex and marginal maximum likelihood may be computationally heavy. In this paper we resort to simulation techniques as MCMC methods and particularly to the Gibbs sampler, adopting a fully Bayesian approach. The advantages are the possibility of estimating item parameters and individual abilities jointly, the capability of including uncertainties about item parameters and abilities in the prior distributions, and the use of Bayesian model comparison techniques (Matteucci, 2013).

3 Results

We consider the INVALSI test administered in the scholastic year 2008-2009 at the end of the lower secondary school (eighth grade) consisting of 30 reading, 10 grammar and 21 mathematics items. We show only the main estimation results for both models with three specific abilities. In general the item parameters have been estimated accurately, because both standard deviations and Monte Carlo standard error are very low. The discrimination parameters are all largely positive, indicating a strong relationship among the item responses and the abilities, while the threshold parameters (difficulty parameters) are spread along the interval $[-2; 2]$, meaning that the test contains items with different levels of difficulty.

As regards the additive model the correlations between the general ability and reading comprehension is 0.60, mathematics 0.35 and grammar 0.24. As the hierarchical model the parameters for the linear relation between the specific and general abilities are quite similar (around 1.80) for reading and grammar and 1.24 for mathematics.

The results for both models seem to indicate that the general ability should be interpreted as a “cognitive factor” conditioning the performance in all the domains. Moreover by using the additive model it is possible to compare the results of two students, for example in mathematics, considering the estimates of the specific ability obtained by conditioning to the effect of the general ability. In other words, the difference in the mathematics ability of two students is a mere result of their learning achievement which is not influenced by the “cognitive factor”. Two students could have the same level of general factor but different level of mathematics ability and this difference is due only to the actual learning achievement.

Considering these issues we determine for the two models the distribution of the general and specific abilities as shown in Figure 2. Then, considering only the “Cognitive factor”, we compare the results to the score assigned directly by INVALSI.

Figure 2: Ability estimate for the two models

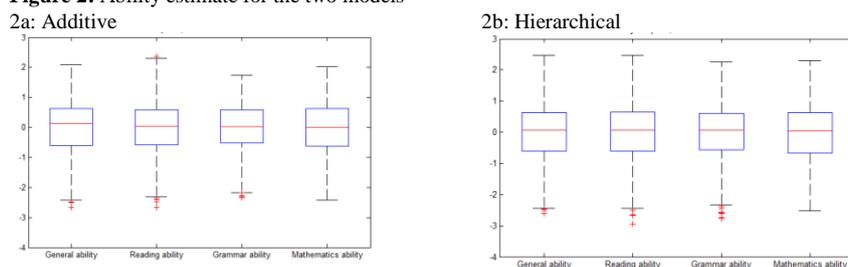


Table 1: Comparison between abilities and score

	<i>INVALSI score < 60</i>		<i>INVALSI score ≥ 60</i>	
	θ_0 Ad	θ_0 Hi	θ_0 Ad	θ_0 Hi
Mean	-1.44	-1.35	0.21	0.18
Median	-1.44	-1.40	0.20	0.24
S.D	0.47	0.55	0.73	0.73

The experts of INVALSI suggest to set 60 as cut score for fail/pass decision. As we can see the results for the models are coherent with the INVALSI score, in particular low abilities are associate to students that fail the test and vice versa. The two estimated θ_0 are quite similar though the additive model seems to explain slightly better the differences in the tails of distribution.

The use of estimated ability to assign a score seems to be recommended as to obtain a more accurate procedure to graduate students in order to define appropriate pass/fail criterion. The results confirm the need of taking into account the complex structure of the test to obtain appropriate guidelines for interpreting the learning performances.

References

1. Matteucci M.: An investigation of parameter recovery in MCMC estimation for the additive IRT model. *Communication in Statistics - Theory and Methods* (2013)
2. Reckase, M.: *Multidimensional Item Response Theory*. Springer-Verlag (2009)
3. Sheng, Y., Wikle, C.: Comparing multiunidimensional and unidimensional item response theory models. *Educ. and Psychol. Measurement*, 899-919 (2007)
4. Sheng, Y., Wikle, C.: Bayesian multidimensional IRT models with an hierarchical structure. *Educ. and Psychol. Measurement*, 413-430 (2008)
5. Sheng, Y., Wikle, C.: Bayesian IRT models incorporating general and specific abilities. *Behaviormetrika*, 27-48 (2009)

Extending the Forward Search to the Combination of Multiple Classifiers: A Proposal

Sabina Mazza

Abstract This contribution presents the extension of the Forward Search approach to the particular ensemble learning scheme StackingC - a model that combines predictions deriving from different supervised classification methods - with the aim of creating a robust procedure that enables us to evaluate the effects exerted by the single units on the model and on the decision rule of each method in order to be able to verify dynamically the stability and robustness and to identify any anomalous values.

Keywords: Combination of supervised classification methods, Forward Search, Stacking scheme

1 Introduction

In order to satisfy the need for models that are more stable and more precise in their predictions, various methods have been proposed by the literature, based on the combination of models from the same class, among which: Bagging [2], Boosting [3] and on others based on the combination of predictions deriving from a set of different supervised classification algorithms (*base-level classifiers*) by means of a *meta-level classifier* in order to improve performances.

This approach is also known as an *ensemble of classifiers* in the supervised classification task. The trend of studies in this direction that started with *Stacked Generalization* [7] is particularly interesting, and is consolidated by the proposals offered by *Stacking* [6] and *StackingC* [5], which tackle and overcome crucial problems previously unsolved in continuity with the original theory.

The presence of outliers in the dataset could also alter the structure of the model,

and thus cause the generation of predictions that might not be reliable.

For this reason, the proposed contribution presents the extension of the Forward Search approach [1] to the particular ensemble learning scheme StackingC, in order to build a robust procedure and to monitor the effects that each observation, outlier or not, can exert on the model and to evaluate the performances of the various base classifiers and the final classifier, as well as the behaviour of the decision rule that presides over the functioning of each algorithm and of Stacking. The “philosophy” at the heart of the Forward Search approach is the creation of a dynamic data analysis process.

Starting from the construction, using robust methods, of a subset $S(m)$ free from anomalous values and which represents the heart of the distribution, a dynamic implementation will be achieved, thus increasing the size of the robust sample selected with the introduction of one observation at a time. The choice of the new cardinality subset $m+1$ is found by using the Mahalanobis distances calculated to the step m . More precisely, the observations are chosen with the $m+1$ with the smallest distances to form the new subset $S(m+1)$. The process is repeated at every step of the search and continues until $m = n$.

2 Extending the Forward Search to Stacking

We created a specific routine which performs Forward Search in multivariate analysis in the context of the combination of supervised classification methods, which is then inserted into the field of the FSDA toolbox [4], created for multivariate data analysis.

Thus the typical Forward Search procedures are extended to Stacking, procedures which refer to the choice of the best robust subset, the criteria for search progress and the creation of specific plots that support the monitoring of the quantity of interest graphically.

Since we are dealing with classification problems, the strength of the search in this case lies in the graphic representation and the plots generated by the procedure which will allow us to monitor several interesting aspects, especially in the presence of outliers. In particular, we can observe the effects on the performances of the classifiers in terms of prediction error and on the rule of decision.

3 Experimental Results

For the analysis we used both datasets generated by the experimental plan and some real datasets. The same Stacking scheme with 7 base classifiers¹ was used in each application, with multi-response linear regression and ridge regression as meta-learner

¹ Linear Discrim. Analysis (LDA), Quadratic Discrim. analysis (QDA), Classification Tree (TRE), Bagged Classification Tree (BAG), AdaBoost (ADA), Naïve Bayes (NBA), Logistic Regression (GLM)

Only the results of one experiment on simulated contaminated data will be displayed in the following plots:

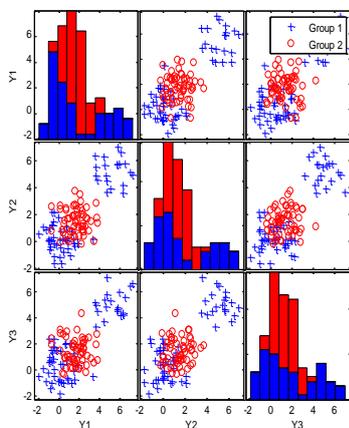


Figure 1. Simulated contaminated data. Scatterplot matrix with bivariate scatters of the three variables and histograms on the main diagonal. The units in Group 1 are represented by blue crosses.

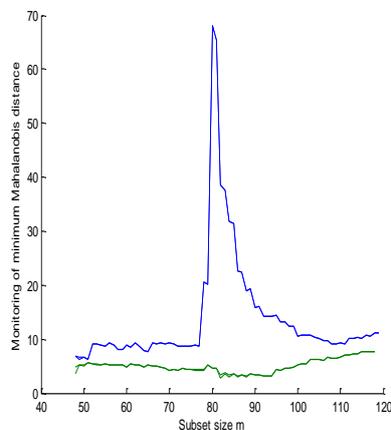


Figure 2. Simulated contaminated data. Plot of minimum Mahalanobis distances for units not belonging to the subset. Blue line, Group 1. Balanced search

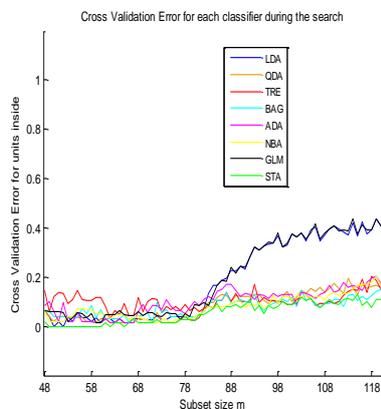


Figure 3. Simulated contaminated data. Cross-validation error for units belonging to the subset of the seven base classifiers and Stacking. Balanced search.

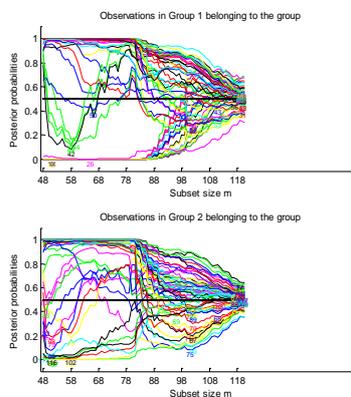


Figure 4. Simulated contaminated data. Linear Discriminant Analysis: posterior probabilities of correct classification of the units of Group 1 (upper panel) and of Group 2 (bottom panel). Balanced search

We started with a subset size of 48 out of the total 120 units, and a balanced search (the units will be inserted into the subset, taking into account the ordering of the Mahalanobis distances carried out separately for each group) and a common covariance matrix for the two groups.

On examining the scatterplot matrix in Figure 1 we can see that the situation appears to be very interesting, since the contamination of the group of observations is quite high and the outliers of the first group are positioned to the right of the second group for all the pair wise scatter plots of the variables. An analysis of the plot of the minimum Mahalanobis distances among the units not belonging to the subset (Figure 2) clearly reveals the entrance of anomalous observations into the subset, which is shown by a peak in the curve (in blue) relative to group 1 in the step prior to the inclusion of the first anomalous unit. In the cross-validation error graph, in Figure 3, at the point where we have established that anomalous observations make their entrance, by means of the Mahalanobis distance plot, there is a general increase in the curves relative to classifiers but there is clearly an increase for those relative to Linear Discriminant Analysis and Logistic Regression, while Stacking (STA) remained the best for all the steps of the search. The graph of the posterior probabilities is very interesting, since it shows that even if all the classifiers are sensitive to the entrance of outliers, all of them respond in different ways and there are changes of direction. By analysing the plot of the posterior probabilities of Linear Discriminant Analysis (Figure 4), we can observe that there are no units that are perfectly classified until the end of the search. When the outliers enter, the trajectories undergo a consistent change of direction, and they are concentrated on the last steps around 0.5.

In the passage from the traditional approach to Forward Search, we acquire a much wider knowledge of the structure of the data, the trajectories of the observations, and the importance exerted by each of these factors, whether outlier or not, on the model or on the aspects that are of interest for evaluating the performances of the Stacking scheme.

References

1. Atkinson, A.C., Riani, M., Cerioli, A.: Exploring Multivariate Data With the Forward Search Springer Verlag, New York (2004)
2. Breiman, L.: Bagging Predictors. *Machine Learning*, **24**, 123-140 (1996)
3. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm, *Proceedings of the International Conference on Machine Learning*, 148-156, Morgan Kaufmann, San Francisco (1996)
4. Riani, M., Perrotta, D., Torti, F.: FSDA: A MATLAB toolbox for robust analysis and interactive data exploration, *Chemometrics and Intelligent Laboratory Systems*, **116**, 17-32 (2012)
5. Seewald, A.K. : How to make Stacking better and faster while also taking care of an unknown weakness. In *Proceedings of the 19th International Conference on Machine Learning, ICML-2002*. Morgan Kaufmann Publisher, San Francisco (2002).
6. Ting, K. M., Witten, I. H. : Issues in stacked generalization. *Journal of Artificial Intelligence Research*, **10**, 271-289 (1999)
7. Wolpert, D.H. : Stacked Generalization. *Neural Networks*, **5**, 241-259, Pergamon Press (1992)

Plug-in Bootstrap for Sample Survey Data

Fulvia Mecatti¹ and M. Giovanna Ranalli

Abstract

Bootstrap algorithms are simple and appealing solutions for variance estimation, confidence intervals and p-value. However for sample survey data they must account for the complex nature of the sampling design. In this paper, a new perspective to finite population bootstrap is given, according to fundamental bootstrap principles such as the mimicking and the plug-in. The method illustrated allows for significant computational advantages and qualifies as a first step towards a unified approach to bootstrapping sample survey data.

1 Introduction and Motivation

Bootstrap algorithms are simple and general computational tools for assessing estimators' accuracy - for instance via variance estimation - and for producing confidence intervals and p-values. Bootstrap applies to finite sample sizes and provides numerical solutions for non standard situations so that it is particularly appealing when dealing with sample survey data and complex sampling designs. We refer to the expression *complex sampling* in a broad sense, intending any departure from the classical system of independent and identically distributed (*iid*) sample variables, including without replacement selection (WOR) and unequal probability selection. Since the original Efron's Bootstrap applies to *iid* samples, adaptations are required for dealing with the non-*iid* nature of data from a complex sample. Literature about bootstrapping

¹ Fulvia Mecatti, Department of Sociology & Social Research, University of Milan-Bicocca; email: fulvia.mecatti@unimib.it

² M. Giovanna Ranalli, Department of Economics, Finance & Statistics, University of Perugia; email: giovanna.ranalli@stat.unipg.it

sample survey data appears to have developed according to two major approaches:

(1) a *functional* approach is based on *iid* re-sampling - as for the original bootstrap - and it requires the *re-scaling* of sample data in such a way that if the parameter to be estimated were the population mean or total (linear case, for short) then the final bootstrap estimate for the estimator's variance would perfectly match the (usually known) analytic variance estimate. This approach includes for instance the *rescaling* bootstrap (Rao and Wu, 1988);

(2) a more *fundamental* approach is based on bootstrap principles such as the *mimicking* principle and the *plug-in* principle (Hall, 1992). It implies re-sampling from a *bootstrap population* by mimicking the very design that provides sample data (Chao and Lo, 1985; Booth et al., 1994).

Recent proposals trying to integrate the two approaches are the *direct* bootstrap (Antal and Tillé, 2011) and the *generalized* bootstrap (Beaumont and Patak, 2012). Notice that approach (2) besides being consistent with bootstrap foundations, appears preferable for handling complex survey data for which more than one variance estimator for the linear case is available, such as for instance probability-proportional-to size samples (πPS). In fact this would require problem-dependent arbitrary choices. Moreover, non-*iid* bootstrap algorithms based on the bootstrap population (approach 2) are proved to ensure good inferential properties as for the Efron's *iid* bootstrap (Hall, 1992).

In this paper we propose a *plug-in* bootstrap based on re-sampling from a *working* distribution which is equivalent to re-sampling from the bootstrap population according to approach (2), thus avoiding its physical construction with significant computational advantages. We will use *iid* sampling, i.e. with replacement and equal probability, to show how to implement the *plug-in* bootstrap in more complex non-*iid* designs. We also discuss how the *plug-in* bootstrap can encompass as special cases several of the bootstrap methods under both approaches (1) and (2) above, thus qualifying as a first step towards a unified approach to bootstrapping sample survey data.

2 Bootstrapping under the mimicking principle and implementing the *plug-in* bootstrap

Let $U = \{1 \cdots k \cdots N\}$ be the target population from which a sample s is selected under a (probability) sampling design $p(s|U)$ and let $\hat{g} = \hat{g}(s)$ denote the estimator of the target parameter $g = g(U)$. A Bootstrap Population is an empirical population re-constructed from sample data as $U^* = \{k \in s \text{ each replicated } d_k^* \text{ times}\}$. A bootstrap sample s^* is produced by re-sampling in U^* and the replication $\hat{g}^* = \hat{g}(s^*)$ is computed. The process is

iterated a large number B of times producing the bootstrap distribution $\{\hat{\vartheta}_b^*, b=1 \dots B\}$ which is a Monte Carlo estimate of the distribution of the estimator $\hat{\vartheta}$, no matter how complex it might be. The bootstrap distribution is used for variance estimation and for estimating percentiles and p-values of $\hat{\vartheta}$. For instance, the variance of the B replications $\hat{\vartheta}_b^*$ defines the bootstrap estimate $v^*(\hat{\vartheta})$ of the estimator's variance $V(\hat{\vartheta})$. As a consequence, the potential of the bootstrap distribution and bootstrap variance as accurate estimates of, respectively, the distribution and variance of estimator $\hat{\vartheta}$ lies significantly upon the capability of the entire re-sampling process of *mimicking* the original sampling process generating $\hat{\vartheta}$ as an estimate of ϑ . Hence, according to the mimicking principle (a) re-sampling should be performed into U^* where the frequency d_k^* are chosen to mirror the known features of the original population U at the largest extent; and (b) the re-sampling design $p(s^* | U^*)$ should be copying the original one. However, it is often objected that the reconstruction of U^* where to physically perform the re-sampling can be heavily resource-consuming, especially under complex though popular sampling designs such as πPS . The *plug-in* bootstrap allows for a full bootstrap population approach by avoiding the actual reconstruction of U^* .

For illustrating the implementation of the plug-in bootstrap we start focusing on the *iid* case, i.e. the basic simple random sample with replacement. We largely refer upon the notation used in Tillé (2006). Let $\mathbf{S} = \{S_k, k=1 \dots N\}$ be a random vector where \mathbf{S} indicates the number of times each population unit k is selected in the sample so that $\sum_k S_k = n$ is the sample size. Then, the sampling design can be expressed as the (multivariate) probability distribution of \mathbf{S} : $p(s | U) = p(\mathbf{s}) = \Pr(\mathbf{S} = \mathbf{s})$. For the basic *iid* case we have $S_k \sim \text{Binomial}(n, 1/N)$ so that the (original) sample $s | U$ is produced by generating a value \mathbf{s} of the N -dimensional random variable $\mathbf{S} \sim \text{Multinomial}(n, 1/N \dots 1/N)$. Consequently for producing a bootstrap sample $s^* | U^*$ by both re-sampling from U^* and mimicking the original sampling design it would require (a) to choose $n^* = n$ and $d_k^* = N/n$; and (b) to generate from $\mathbf{S}^* \equiv \mathbf{S} \sim \text{Multinomial}(n, 1/N \dots 1/N)$. Notice that the re-sampling procedure above is equivalent to randomly select from an urn (U^*) containing N balls of the n different colours contained in the original sample s each with frequency n/N . Thus a bootstrap sample $s^* | U^*$ is equivalently produced by directly selecting from the original sample s with probabilities $(1/N)(N/n) = 1/n$, i.e. by generating a value \mathbf{s} of the n -dimensional random

variable $\mathbf{S}^* | s \sim \text{Multinomial}(n, 1/n \cdots 1/n)$. Thus a drastic computational reduction is allowed by re-sampling from the *working* distribution $\mathbf{S}^* | s$ while guaranteeing both the equivalence with the re-sampling from U^* and the same properties over the final bootstrap distribution. This defines the *plug-in* bootstrap. Finally notice that in the illustrative *iid* case above the *plug-in* bootstrap perfectly matches the original Efron's bootstrap.

More complex sampling design can be treated accordingly. For instance it is proved that (Ranalli and Mecatti, 2012): a) for simple random sampling WOR and equal probability, the natural choice $d_k^* = N/n$ gives the working re-sampling $\mathbf{S}^* | s \sim \text{Multi Hypergeom.}(n, N/n \cdots N/n)$. This is equivalent to the WOR bootstrap (Chao and Lo, 1985); b) for random size list-sequential Poisson sampling the natural choice $d_k^* = \pi_k^{-1}$ gives $\mathbf{S}^* | s \sim \text{Binomial}(d_k^*, \pi_k)$. This is equivalent to the *generalized* bootstrap (Beaumont and Patak, 2012); and c) for simple random sampling with replacement and unequal probability the natural choice $d_k^* = \pi_k^{-1}$ gives $\mathbf{S}^* | s \sim \text{Multinomial}(n, 1/n \cdots 1/n)$ as for the equal probability case. The *direct* bootstrap provides a similar result (Antal and Tillé, 2011).

For π PS WOR sampling which is extensively used in large scale surveys, the potential computational advantages of the *plug-in* bootstrap appears even greater. However the natural choice $d_k^* = \pi_k^{-1}$ gives more complex working distributions depending on the particular design implemented. For instance Conditional Poisson design associates with a non- central multivariate Hypergeometric working distribution of the Fisher type. For this case more research is currently undergoing.

References

1. Antal, E., Tillé, Y: A direct bootstrap method for complex sampling design from a finite population, Journal of the American Statistical Association, 106, 534-543 (2011)
2. Beaumont, J-F., Patak, Z: On the generalized bootstrap for sample surveys with special attention to Poisson sampling, International Statistical Review, 80, 127-148 (2012)
3. Booth J.G., Butler R.W., Hall P.: Bootstrap methods for finite populations, Journal of the American Statistical Association, 89, 1282-1289 (1994)
4. Chao M.T., Lo A.Y.: A bootstrap method for finite population, Sankhya, 47, 399-405 (1985)
5. Hall, P.: The bootstrap and the Edgeworth expansion, Springer-Verlag (1992)
6. Ranalli, M.G., Mecatti, F: Comparing recent approach for bootstrapping sample survey data: a first step towards a unified approach, Proceedings of the Section on Survey Research Methods, American Statistical Association, 4088-4099 (2012)
7. Rao, J.N.K., Wu, C.F.J.: Resampling inference with complex survey data, Journal of the American Statistical Association, 83, 231-241 (1988)
8. Tillé, Y.: Sampling Algorithms, Springer (2006)

A BLU Predictor for Spatially Dependent Functional Data of a Hilbert Space

Alessandra Menafoglio, Matilde Dalla Rosa and Piercesare Secchi

1 Introduction

In most geophysical and environmental applications, the observed natural phenomena are very complex and they rarely show a uniform behavior over the spatial domain. Collected data inevitably reflect these features, being often intrinsically highly dimensional, strongly spatially dependent and possibly non-stationary. This work addresses the problem of performing optimal linear spatial prediction when the available data are functional and georeferenced.

The problem of spatial prediction is a classic topic in geostatistics [3] and it has recently received a great deal of attention in the literature on infinite-dimensional data belonging to the space L^2 [5, 7, 4]. The aim of this communication is to provide a summary about a novel extension of the kriging methodology to non-stationary functional random fields belonging to any separable Hilbert Space, that has been introduced and thoroughly explored in [6]. Theoretical and practical issues are faced, finally applying the proposed methodology to a real case study, dealing with daily mean temperatures curves recorded in the Canada's Maritimes Provinces.

2 Model and Prediction Problem

Let H be a separable Hilbert space, with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$, whose points are functions $x : \mathcal{S} \subset \mathbb{R} \rightarrow \mathbb{R}$. Consider a random field $\{\chi_s, s \in D \subset \mathbb{R}^d\}$ such that, for each $s \in D$, χ_s is a random element of H . Let $m_s = \mathbb{E}[\chi_s]$, $s \in D$, be the mean function, or drift, and $\{\delta_s = \chi_s - m_s, s \in D\}$ be the residual process. Assume for the process a non-stationary model describing the drift m_s through a linear model and the residual δ_s through a zero-mean, globally second-order stationary and isotropic random field with covariance function $C(\cdot)$, i.e. [6]:

Alessandra Menafoglio and Piercesare Secchi

MOX-Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy, e-mail: alessandra.menafoglio@polimi.it; piercesare.secchi@polimi.it

Matilde Dalla Rosa

Eni S.p.A., Exploration & Production Division, Via Emilia 1, San Donato Milanese (MI), Italy, e-mail: matilde.dalla.rosa@eni.com

$$m_s(t) = \sum_{l=0}^L a_l(t) f_l(s), \quad f_0(s) = 1 \quad \forall s \in D, \quad t \in \mathcal{T}; \quad a_l \in H, \quad \forall l = 0, \dots, L;$$

$$C(h) = \text{Cov}(\delta_{s_i}, \delta_{s_j}) = \mathbb{E}[\langle \delta_{s_i}, \delta_{s_j} \rangle], \quad \forall s_i, s_j \in D, \quad h = \text{dist}(s_i, s_j).$$

Let $\chi_{s_1}, \dots, \chi_{s_n}$ be a sample –observed in the sites $s_1, \dots, s_n \in D$ – of a realization of the process, that can be figured as a surface whose points are functions. The aim of this work is the prediction of the unobserved point $\chi_{s_0}^*$ of the same realized surface, through a Universal Kriging (UK) predictor, which is the best linear unbiased predictor (BLUP) $\chi_{s_0}^* = \sum_{i=1}^n \lambda_i^* \chi_{s_i}$, whose weights $\lambda_1^*, \dots, \lambda_n^* \in \mathbb{R}$ solve:

$$\min_{\lambda_1, \dots, \lambda_n \in \mathbb{R}} \mathbb{E}[\|\chi_{s_0}^* - \chi_{s_0}\|^2] \quad \text{subject to} \quad \mathbb{E}[\chi_{s_0}^* - \chi_{s_0}] = 0. \quad (1)$$

Assuming the covariance function $C(\cdot)$ and the family of regressors $\{f_l\}_{l \geq 0}$ to be known and $\{a_l\}_{l \geq 0}$ independent from the spatial location, problem (1) can be explicitly solved, as it reduces to the following linear system (in block-matrix form):

$$\begin{pmatrix} \gamma(h_{i,j}) & f_l(s_i) \\ f_l(s_j) & 0 \end{pmatrix} \begin{pmatrix} \lambda_j \\ \mu_l \end{pmatrix} = \begin{pmatrix} \gamma(h_{0,i}) \\ f_l(s_0) \end{pmatrix}, \quad (2)$$

where $\{\mu_l\}_{l \geq 0}$ are Lagrange multipliers, while $\gamma(\cdot)$ denotes the trace-semivariogram function of the residual process:

$$\gamma(h_{i,j}) = \text{Var}[\delta_{s_i} - \delta_{s_j}] = \mathbb{E}[\|\delta_{s_i} - \delta_{s_j}\|^2], \quad s_i, s_j \in D, \quad h_{i,j} = \text{dist}(s_i, s_j). \quad (3)$$

3 Assessing the Drift and the Covariance Structure

Two main issues are to be faced to derive the desired prediction: 1) the covariance structure is rarely ‘a priori’ known and needs to be estimated from the residuals; 2) the residuals need to be estimated themselves since they are in general unobserved. As in classical geostatistics, the first issue can be addressed in two steps: first the computation of an empirical estimate from (an estimate of) the vector of residuals $\delta_s = (\delta_{s_i})$, then the fitting of a variogram valid model. The second point can be appropriately handled with a generalized least squares (GLS) estimation of the drift, since a residuals estimation can be derived by difference from it. Indeed, unlike classical ordinary least squares (OLS) estimators, this estimator takes properly into account the spatial dependence among observations by minimizing the functional Mahalanobis distance between the observations vector $\chi_s = (\chi_{s_i})$ and the drift estimates $\hat{m}_s = (\hat{m}_{s_i})$: $d_{\Sigma^{-1}}(\chi_s, \hat{m}_s) = \sum_{i=1}^n \|\Sigma^{-1/2}(\chi_s - \hat{m}_s)\|_i^2$, $\Sigma = \text{Cov}(\chi_s)$.

This estimation problem admits a unique solution \hat{m}_s^{GLS} explicitly derived in [6] as: $\hat{m}_s^{GLS} = \mathbb{F}_s(\mathbb{F}_s^T \Sigma^{-1} \mathbb{F}_s)^{-1} \mathbb{F}_s \Sigma^{-1} \chi_s$, $(\mathbb{F}_s)_{il} = f_l(s_i)$. In [6] it is also shown that \hat{m}_s^{GLS} coincides with the best linear unbiased estimator (BLUE) for the mean vector m_s , that allows to derive the unbiased minimum variance prediction of $\chi_{s_0}^*$ when the drift is unknown. However in practice, to compute the GLS drift estimation, an iterative algorithm is necessary, since the drift estimator \hat{m}_s^{GLS} depends substantially on the residual covariance structure, that can be assessed only once an estimation of

the residual process —obtained by difference from the estimate \widehat{m}_s^{GLS} — is available. Therefore we propose to initialize the procedure to the OLS estimate, computing at each step the residual estimate and the related trace-variogram structure, as well as the update of the drift estimate on the basis of the structure of spatial dependence currently available. Hence, the final estimate of the variogram model can be used to solve the Universal Kriging system (2), deriving the desired prediction.

4 Case Study

The proposed methodology is applied to the Canada’s Maritime Provinces Temperatures dataset, that collects daily mean temperatures data, observed during 1980 in 27 meteorological stations located in Canada’s Maritimes Provinces [1] (Fig. 1).

Geostatistical analysis is performed by considering a non-Euclidean metric on the spatial domain and modeling the non-stationary mean behavior of the process through a drift term. In the absence of ‘a priori’ information on the deterministic variability of the process, a preliminary drift model selection is performed by cross-validation among polynomials of order lower than 2, singling out as optimal drift model a quadratic form ($\{f_l\}_{l=0,\dots,3} = \{1, y, x^2, y^2, xy\}$). Five iterations of the drift estimation algorithm proves sufficient for its convergence, leading finally to GLS drift estimation maps and Universal Kriging maps (Fig. 2, top and bottom panels respectively).

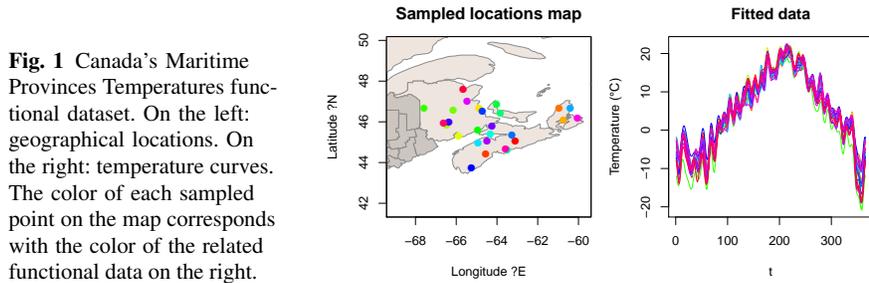


Fig. 1 Canada’s Maritime Provinces Temperatures functional dataset. On the left: geographical locations. On the right: temperature curves. The color of each sampled point on the map corresponds with the color of the related functional data on the right.

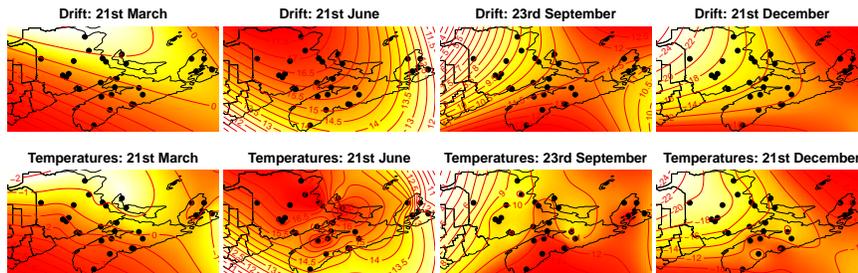


Fig. 2 GLS drift estimation maps (top panels) and UK prediction maps (bottom panels).

These results reflect the climatical features of the Maritimes region: the alternation of Atlantic warm-humid currents from S-E with N-W freezing streams coming from the internal Canadian regions drives the temperatures with a consequent rotation in the drift contour lines (Fig. 2, top panels). Furthermore, these results are consistent with respect to the seasonal reference maps [1]. Notice that the regularizing effect of kriging is partially mitigated by the presence of the drift term, that turns the prediction from a strongly data driven behavior to a model driven one. The resulting prediction is thus precise also in peripheral regions, catching the local structures as well (Fig. 2, June panel). The increased flexibility achieved by modeling a drift term significantly improves the predictive power of the method –assessed by cross-validation– with respect to stationary results [6].

5 Discussion and Further Research

In this work a novel kriging methodology for functional data belonging to any separable Hilbert space has been illustrated, with application to a real case study where a spatial varying mean is modelled.

Developing statistical models and inferential procedures for general Hilbert spaces, instead of working out *ad hoc* techniques for L^2 , opens broad perspectives of research. Indeed, it may allow the integration of the kriging methodology with the physical model underlying the observed phenomenon. For instance, in our setting, both point-wise and differential properties can be taken explicitly into account by embedding the analysis in a proper Sobolev space. Constrained data can be dealt with as well by choosing an appropriate geometry (e.g., Aitchison geometry [2]).

Finally, within our framework, a cluster-varying drift might be modelled, with a KED (Kriging with External Drift) approach. However, in such a case, the problem of the spatial prediction of a function χ_{s_0} would turn into two subproblems, since the prediction of χ_{s_0} would require a first (non-trivial) prediction of the cluster label in s_0 –needed to solve system 2–. To deal with highly heterogeneous data, more complex linear models would be worth investigating, possibly including more complex regressors, influencing or driving the physical system.

References

1. Natural Resources of Canada website: <http://atlas.nrcan.gc.ca/> (2012).
2. Aitchison, J.: The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B* **44**(2), 139–177 (1982).
3. Cressie, N.: *Statistics for Spatial data*. John Wiley & Sons, New York (1993).
4. Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data. *Environmetrics* **21**, 224–239 (2010).
5. Goulard, M., Voltz, M.: Geostatistical Interpolation of Curves: A Case Study in Soil Science. In: *Geostatistics Tróia '92*, ed. A. Soares, Dordrecht: Kluwer Academic, pp. 805-816 (1993).
6. Menafoglio, A., Dalla Rosa, M., Secchi, P.: A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. MOX-report 34/2012. Politecnico di Milano, Milano. <http://mox.polimi.it> (2012).
7. Nerini, D., Monestiez, P., Manté, C.: Cokriging for spatial functional data. *Journal of Multivariate Analysis*, **101**(2), 409–418 (2010).

Comparing fuzzy and multidimensional methods to evaluate well-being at regional level

Maria Adele Milioli, Lara Berzieri, Sergio Zani¹

Abstract. Well-being composite indicators aim to measure the global life conditions of people and are based on a set of observable variables. We compare the fuzzy set approach and the classical multidimensional methods (principal component analysis, cluster analysis) in order to obtain well-being measures in European regions. The analysis of the results points out the advantages and the shortcomings of the different criteria.

1. Introduction

The problems involved in using GDP as a measure of economic welfare have been well recognized and alternative approaches have been suggested (e.g. Stiglitz, Sen and Fitoussi, 2009; Bleys, 2012, CNEL-ISTAT, 2013). The concept of well-being may be considered as a latent and multidimensional phenomenon, encompassing a wide range of domains described by sets of manifest variables. Well-being composite indicators are sensitive to the variables that are selected, to the methods and weights used in the aggregation: different choices may entail quite different results (Paruolo *et al.*, 2013).

In a previous paper (Berzieri *et al.* 2013) we suggested a fuzzy set approach in order to obtain a composite indicator of well-being in the European regions at level NUTS2 (Nomenclature of Territorial Units for Statistics). The values of this fuzzy indicator, in the range $[0, 1]$, show the gradual transition from the poorest regions to the areas with the highest well-being.

In this communication we compare such results with the ones obtained by applying classical multidimensional methods (principal component analysis, k -means cluster analysis) to the same data set. The NUTS 2 classification subdivides the 27 European States into 266 regions (after deletion of 5 units not belonging to European Union actually). Starting from the list of all available variables for European regions in Eurostat database (reference year 2010), a subset of 16 suitable variables has been selected, with respect to six domains: wealth and free time; labour market, education, demography, health, environment.

Related recent works on the measurement of well-being in European regions are: Pittau *et al.* (2010), Annoni *et al.* (2012), Okulicz-Kozaryn (2012).

¹Maria Adele Milioli, Dep. of Economics, University of Parma, mariaadele.milioli@unipr.it;
Lara Berzieri, Comune di Parma, l.berzieri@comune.parma.it ;
Sergio Zani, Department of Economics, University of Parma, sergio.zani@unipr.it

2. The suggested fuzzy indicator

Fuzzy set theory provides an approach to deal with vague concepts as well-being or quality of life (Balioune-Lutz, 2006; Lazim and Osman, 2009; Facchinetti *et al.*, 2012). The measure of well-being can be expressed as membership degree to the subset A of the best areas.

Consider a set of n regions r_i ($i = 1, 2, \dots, n$) and p manifest variables X_s ($s = 1, 2, \dots, p$) reflecting the different aspects of well-being. Without loss of generality, let us assume that each variable is positively related with well-being. If a quantitative variable X_s shows negative correlation, we substitute it with a simple decreasing transformation, e. g. $f(x_{si}) = \max(x_{si}) - x_{si}$.

Each crisp variable X (for simplicity of notation we omit index s) is transformed into a fuzzy variable, defining the membership function (m.f.) $\mu_A(x_i)$ as follows:

$$\begin{aligned} \mu_A(x_i) &= 0 & x_i &\leq l \\ \mu_A(x_i) &= (x_i - l) / (u - l) & l < x_i < u \\ \mu_A(x_i) &= 1 & x_i &\geq u \end{aligned} \quad (1)$$

We have chosen: *lower* threshold l = median of the variable; *upper* threshold u = 90th percentile. With this choice the regions with value of the variable under the median don't belong to the subset A of the best regions, with reference to this aspect, and the regions with the 10% highest values totally belong to the subset of richest areas.

In this paper, for lack of space, we consider only the fuzzy composite indicator defined as unweighted arithmetic mean of the m.f. values of the 16 variables. This criterion is justified by the previous careful selection of the variables, ensuring a balance of the different aspects of well-being.

The values of this fuzzy composite indicator have an interesting interpretation: a value equal to 0 corresponds to a region under the median for all the variables, a value equal to 1 identifies a region over the 90th percentile for all the variables; a value in the open range (0, 1) may be assumed as membership degree of the region to the subset A of the areas with the highest well-being. See Fig. 1 for the map of the regions according to 5 classes based on percentiles of the values of the indicator.

3. Multidimensional methods

We have applied principal component analysis (PCA) to the same dataset of 16 variables. The first PC accounts for 37.5% of the total variance and the second PC for 20.2%. The percentage explained by the two PC equal to 57.7% is superior to the threshold $0.95^{16} = 0.44$ (Cronbach's alpha = 0.844). The first PC is highly related to the variables measuring income and wealth, education, labor market and life expectancy; the second PC describes demographic domain.

Six regions may be considered as outliers for too low values: Guyane, Réunion, Martinique, Guadelupe (FR); Melilla, Ceuta (ES) and three regions for too high values: Bruxelles, Inner and Outer London. They have been omitted in the following comparisons.

The linear correlation between the previous fuzzy indicator and the first PC is sufficiently high ($r = 0.932$) and also the rankings of the regions obtained by the two criteria are similar, but not equal (Spearman rho = 0.950). The correlation of these two composite indicators with p.c. GDP is moderate ($r = 0.712$ and $r = 0.759$ respectively).

The visualization of the differences between the two criteria is presented in Fig. 1. The regions arranged in an inferior or superior class by the first PC score with respect to the fuzzy value are highlighted by different type of grid overlapping the color. The figure shows that only a few regions present different percentile classification.

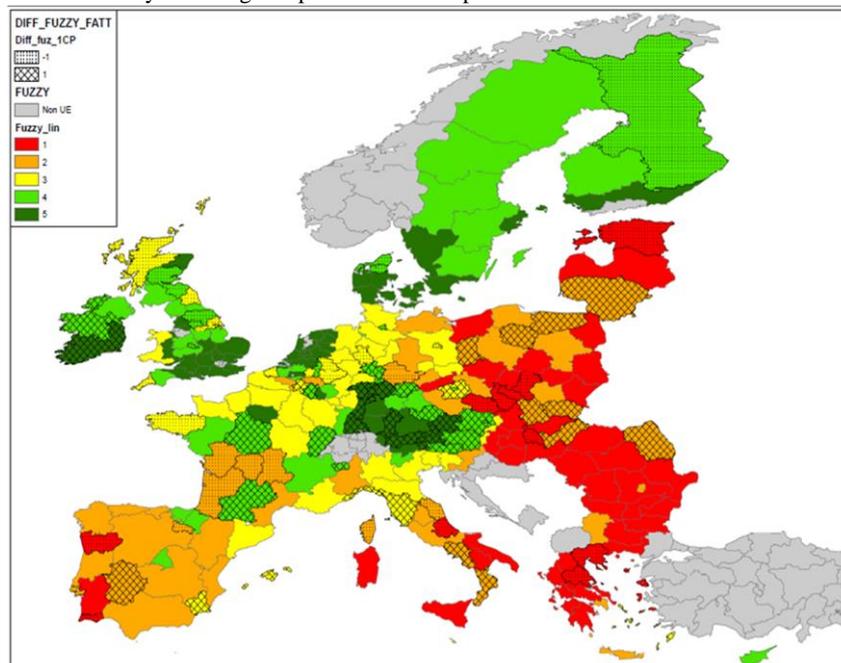


Fig. 1 Map of the values of the fuzzy composite indicator (dark green = best areas; red = worst areas): the regions in a different percentile class according to the scores of the first PC are highlighted by the two types of grid.

Another comparison criterion is the classification of the fuzzy indicator values and the scores of the first PC into a contingency table, considering for each indicator the partition corresponding to 5 classes of percentiles (Table 1).

Table 1: Contingency table of the values of the fuzzy and the 1st PC indicators

		Percentile classes of 1st PC					Total
		1	2	3	4	5	
Percentile classes of the fuzzy indicator	1	39	12	0	0	0	51
	2	11	32	9	0	0	52
	3	1	8	30	12	0	51
	4	0	0	12	31	9	52
	5	0	0	0	9	42	51
Total		51	52	51	52	51	257

Most of the regions (67,7%) are in the same percentile class with the two criteria, i.e. the two indicators show similar but not equal results (Kendall's tau = 0.847).

We have applied *k*-means cluster analysis to the 16 standardized variables, selecting 5 groups (for comparison reason with the previous partitions), ranked according to the average of the values of the fuzzy indicators of the regions in each cluster (Table 2). We also present the average of the scores of the first and second PC. The 65 regions in cluster n. 4 are the ones with the highest well-being evaluated both by fuzzy and by PC indicators.

Table 2: Five clusters of regions obtained by *k*-means method.

Number of the Cluster	Number of regions	Fuzzy values average	First PC scores average	Second PC scores average
3	61	0,089	-1,240	0,178
1	29	0,143	-0,906	0,258
2	37	0,309	0,254	0,974
5	65	0,331	0,323	-1,286
4	65	0,491	1,100	0,437

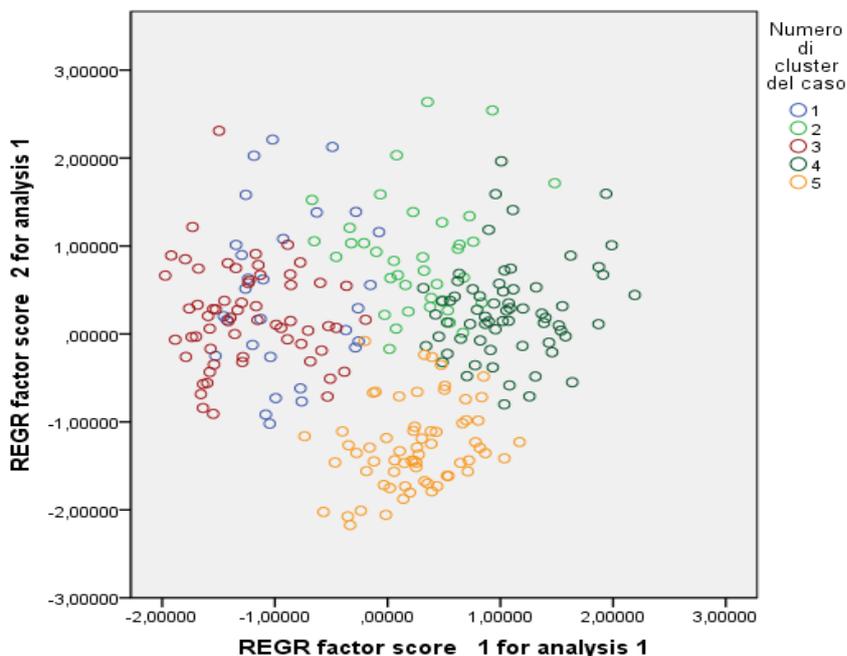
Finally we present in a scatterplot the scores of the first and second PC, labeling the region points with different colors according to their cluster (Figure 2).

The best regions (dark green) are in the right side of the figure (high scores of the first PC) but they are scattered along the second axes, i.e. they show different demographic situations. The clusters n. 2 (green) and n. 5 (yellow) have scores of the first PC approximately in the same range, but they are very different for the scores of the second PC. The regions in cluster n. 3 (red) are the worst and they show partial overlapping with the units of cluster n. 1 (blue). More homogenous clusters could be obtained considering a larger number of groups.

References

1. Annoni, P., Weziak-Bialowolska, D.: Quality of life at the sub-national level: an operational example for the EU. Publications Office of the European Union, Luxembourg (2012).
2. Balamoune-Lutz, M.: On the measurement of human well-being: fuzzy set theory and Sen's capability approach. In: M. McGillivray and M. Clarke (eds) Understanding Human Well-being. United Nation University Press, New York (2006).
3. Berziera, L., Milioli M.A., Zani S.: A fuzzy approach to measure well-being in European regions, SIS Conference "Advances in Latent Variables", contributed papers, Brescia, (2013).
4. Bley, B., Beyond GDP: Classifying alternative measures for progress, Social Indicators Research 109, 355-376 (2012).
5. CNEL-ISTAT, Rapporto BES 2013: il benessere equo e sostenibile in Italia, Roma (2013).

Figure 2 - Scatterplot of the scores of the first and second PC, classified in 5 groups by *k*-means cluster analysis.



6. Facchinetti, G. *et al.*: A fuzzy approach to face the multidimensional aspects of well-being. Fuzzy Information Processing Society, Annual Meeting of the North American. Ieee Conference Publications, Berkeley (USA), 351-356 (2012).
7. Lazim M.A., Abu Osman M.T.: A new Malaysian Quality of Life index base on fuzzy sets and hierarchical needs, *Social Indicators Research* 94, 499-508 (2009).
8. Okulicz-Kozaryn, A.: Income and Well-being Across European Provinces. *Social Indicators Research*, 106, pp. 371-392 (2011).
9. Paruolo P., Saisana M. Saltelli A.: Ratings and rankings: voodoo or science?, *J. R. Statist. Soc. A*, 176, 2, (2013).
10. Pittau, M.G., Zelli, R., Gelman, A.: Economic Disparities and Life Satisfaction in European Regions. *Social Indicators Research* 96, pp.339-361 (2010).
11. Stiglitz, J., Sen, A., Fitoussi, J-P.: Report by the Commission on the Measurement of Economic Performance and Social Progress, Paris (2009).

Comparing text clustering algorithms from a multivariate perspective

Michelangelo Misuraca and Maria Spano

Abstract Text Clustering is an automatic process to organize a large quantity of unlabeled documents into a smaller number of meaningful and coherent clusters similar in content. The main issue of clustering methods is the lack of a shared methodology to evaluate results. A wide variety of validation indices has been proposed in literature, but each of them focuses the attention on a given aspect. Following the idea underlying a composite indicator a multivariate strategy to compare several clustering methods on the basis of different validation measures is proposed.

Key words: text clustering, cluster validity, factorial analysis

1 Introduction

Clustering is a family of unsupervised classification procedures fundamental to Data Mining as well as to Text Mining (TM). The objective is finding the different categories to which a set of objects belong without *a priori* information concerning the number and the composition of the corresponding groups. In the specific frame of TM this approach is suitable for grouping texts with respect to their own topic.

In an explorative perspective it is suitable to refer to Clustering when the aim is to organise the knowledge base underlying a collection of data without any subjective bias, in order to satisfy a specific informative need. Moreover, data structure is taken into account and the grouping task is data-driven. This procedure can be also seen, in a more general TM process, as a first step for selecting the most useful

Michelangelo Misuraca
Università della Calabria, Via P. Bucci - Cubo 0C/1C, I-87036 Arcavacata di Rende (CS), e-mail:
michelangelo.misuraca@unical.it

Maria Spano
Università di Napoli Federico II, Via Cinthia - Complesso Monte Sant' Angelo, I-80126 Napoli
e-mail: maria.spano@unina.it

information with respect to the analysed phenomenon.

In literature a wide variety of Clustering techniques have been proposed for different domains and applications, and many proposals are specifically devoted to texts [1]. The majority of algorithms defines a partitioning of a dataset on the basis of assumptions rather than searching the optimal partitioning that fits the underlying data structure. In many cases applying several algorithms on the same dataset as well as performing the same algorithm with different input parameters can lead to different results. Furthermore, the different algorithms are usually evaluated on toy examples, simulated datasets or benchmark datasets (e.g., Reuters-21578 in Text Mining), while real applications present peculiar characteristics and structures.

One of the main issues of Clustering methods is the lack of a shared methodology to validate the results. In the usual validation task, the effectiveness of a Clustering algorithm is tested by considering different criteria. Each validity dimension points out the attention on a particular aspect such as intra-cluster quality, inter-cluster separation and geometry of the clusters. In this frame it is really hard to compare different solutions and choose the best performing ones.

Following the idea underlying a composite indicator, in this paper we propose to approach the problem of evaluating different Clustering algorithms in a multivariate perspective. The strategy consists in performing a factorial analysis on several algorithms for which a set of validity measures are calculated. After introducing the main characteristics of text clustering algorithms and validation indices, the different steps of our proposal are described and discussed.

2 Methodological framework

There are many different clustering algorithms but they can be classified into few basic types. In literature It is possible to find several reviews on the approaches devoted to text clustering, starting from basic traditional methods to new trends such as fuzzy based, genetic, co-clustering, heuristic oriented algorithms [2].

By considering the structure it can be possible to classically divide the different approaches into *non hierarchical* and *hierarchical* methods. Most algorithms of the first type are usually iterative and improve the initial grouping by reallocating the objects assigned to the clusters in each step. Because the relations among clusters are often undetermined it is possible to consider a hierarchical clustering in order to have an easier interpretation of the emerging classification.

Another important distinction among the different algorithms is whether they carry on a *hard* or a *soft* clustering. In the first one each object is assigned to one and only one group, in the second one a degree of membership is achieved and membership to different clusters is also allowed.

The assessment of clustering quality and the selection of the appropriate method are still open challenges. Since clustering algorithms define clusters that are not known ahead, irrespective of grouping criterion, the final partition of data requires some kind of evaluation. The procedure of evaluating the results of a clustering

algorithm is known as *cluster validity*. The two main approaches to validate a partition are *external* and *internal* validation [5]. The difference is whether or not *a priori* information is used for the validation process. Since external validation indices measure how much the identified clusters correspond to class labels externally provided, they are mainly used for choosing an optimal clustering algorithm on a specific dataset. Internal validation indices evaluate the goodness of a clustering structure without any additional information, and they can be used to choose the best clustering algorithm as well as the optimal cluster number. When class labels are not available internal indices are the only option to evaluate cluster results.

Internal validation indices are based on the following two criteria:

- *Compactness*, which measures how closely related the objects in a cluster are;
- *Separation*, which measures how distinct or well-separated a cluster is from other clusters.

Some of the most well known internal validation measures are the Davies-Bouldin Index, the Calinski-Harabasz Index, the Silhouette Width Criterion, the Dunns Index, just to mention a few [3].

The great variety of possibilities both in terms of clustering methods and measures to assess the quality of the results makes difficult the decision process. For this reason a strategy considering multiple alternatives at the same time can help users in choosing the most suitable solution.

3 A multivariate strategy for clustering validation

The logic of a composite indicator is that different elementary dimensions are summarised in a single complex measure. Each dimension evaluates a peculiar aspect of a multi faced phenomenon. The use of factorial techniques for constructing a composite indicator is well considered in the reference literature [4]. Those techniques allow to overcome some of the main problems related to the choice of an appropriate weighting scheme or data transformation. Moreover, individual indicators selected in arbitrary manner often confuse and mislead researchers and final users.

The proposed strategy for comparing the overall performances of different clustering methods consists in a sequential procedure, as described in terms of pseudo-code in Algorithm 1.

Starting from a documental repository, a set of k methods suitable for clustering textual data and a set of q internal validation measures are considered. After performing a preprocessing process in order to normalise the texts and select the content bearing features, a lexical matrix $\mathbf{T}(n \times p)$ which contains the occurrences of the p terms in n documents is obtained. By applying each clustering method a partition of the documents into different clusters is carried out, and for each partition the set of q internal validation measures are calculated. The final result is a matrix $\mathbf{V}(k \times q)$, where the generic element v_{ij} is the value assumed by the j th validation index for the i th partition.

In order to compare the performances of the different clustering methods with respect to the aspects measured by the internal validation indices, the core step of the strategy consists in performing a factorial analysis on the matrix V . From a geometric viewpoint it is possible to visualize both the different algorithms in the space of the measures and the measures in the space spanned by the different algorithms reaching a double information. Concerning the first representation it is possible to evaluate the results achieved by different methods looking at the relative position on the factorial map, highlighting which of them perform in a similar fashion. At the same time the dual representation of the measures can show which dimension is eventually redundant in evaluating the global validity, having in mind that high correlated indexes express the same kind of information.

Algorithm 1

Given: a set $X = (x_1, \dots, x_n)$ of n objects
 k clustering methods
 q validation measures
for $i = 1$ to k **do**
 apply the clustering criterion $c_i(X)$
 obtain the partition $P_{(c_i)}$
 for $j = 1$ to q **do**
 compute the measure $m_{ij}[P_{(c_i)}]$
 obtain the vector $M_i = (m_{i1}, \dots, m_{iq})$
 end for
 concatenate by row vector M_i
return the matrix $V(k \times q)$
end for
run Factorial Analysis on the matrix V

The effectiveness of the proposed strategy will be reported and discussed elsewhere by performing on both benchmark corpora and *ad hoc* collected documents.

References

1. Aggarwal, C.C., Zhai, C.: A Survey of Text Clustering Algorithms. In: Aggarwal, C.C., Zhai, C. (Eds.) Mining Text Data, pp. 77–128. Springer, Heidelberg (2012)
2. Balbi, S.: Beyond the curse of multidimensionality: high dimensional clustering in text mining. *Statistica Applicata - Italian Journal of Applied Statistics*, **22(1)**, 53–63 (2010)
3. Halkidi, M., Vazirgiannis, M., Batistakis, I.: Quality Scheme Assessment in the Clustering Process. In: Proceedings of PKDD, Lyon, France (2000)
4. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S.: Tools for Composite Indicators building. EUR 21682 EN, European Commission - Joint Research Centre (2005)
5. Theodoridis, S., Koutroubas, K.: Pattern Recognition. Academic Press, Burlington (2003)

Mixture models for ranked data classification

Cristina Mollica and Luca Tardella

Abstract Analysis of ranking data is required in several research fields. In the present work we review statistical models for random rankings and propose an original generalization of a popular parametric distribution, that we name *extended Plackett-Luce model*, to account for the order of the ranking elicitation process. We illustrate the validity of the novel model with its successful maximum likelihood estimation from the real data set of the Large Fragment Phage Display (LFPD) experiment, where the epitope mapping of a specific human protein is the main goal. In particular we address the heterogeneous nature of the experimental units via a finite mixture model approach and compare the performances when alternative ranking models are employed as mixture components.

Key words: Ranking data, Plackett-Luce model, Mixture models, EM algorithm.

1 Introduction

Ranked data arise in several contexts, especially when objective and precise measurements of the phenomena of interest can be impossible or deemed unreliable and the observer gathers ordinal information in terms of orderings, preferences, judgements, relative or absolute ranking among competitors.

Cristina Mollica
Dipartimento di Scienze statistiche, Sapienza Università di Roma
Piazzale A. Moro 5 00185 Roma,
e-mail: cristina.mollica@uniroma1.it

Luca Tardella
Dipartimento di Scienze statistiche, Sapienza Università di Roma
Piazzale A. Moro 5 00185 Roma,
e-mail: luca.tardella@uniroma1.it

Research fields where the analysis of ranked data are very often required are social and behavioral sciences, where studies often ask a sample of N people to rank a finite set of K items according to certain criteria, typically their personal preferences or attitudes. For an introduction on the topic see [4].

Formally, a *full* (or *complete*) *ranking* is a bijective mapping of a finite set $I = \{1, \dots, K\}$ of labeled *items* into a set of *ranks* $R = \{1, \dots, K\}$, that is

$$\pi : I \rightarrow R.$$

Thus a ranking $\pi = (\pi(1), \dots, \pi(K))$ is a sequence in which the generic $\pi(i)$ must be read as the rank attributed to the i -th item. The underlying convention is that if $\pi(i) < \pi(i')$, then item i is ranked higher than (hence preferred to) item i' . The inverse of a ranking $\pi^{-1} = (\pi^{-1}(1), \dots, \pi^{-1}(K))$ is called *ordering* with $\pi^{-1}(j)$ indicating the item ranked in the j -th position.

When a judge proceeds from the elicitation of her best choice (rank 1) up to the worst one (rank K), we have the so-called *forward ranking process*; the inverse ranking procedure is named *backward ranking process*. This formal definition has been originally introduced in [1] but, to our knowledge, the rank assignment scheme has not received an explicit consideration in a model setup in the attempt to improve the description of ranked data. Obviously, any other order for the rank assignment process is admissible and potentially leads to different results. This aspect has inspired us to expand an existing and well-known parametric ranking model and to employ such a new class in the analysis of the LFPD data, in order to verify whether and how the reference order can influence the inferential results and the final model-based clustering.

2 Statistical models for random ranking

2.1 The Plackett-Luce model

The *Plackett-Luce model* (PL) is a very popular parametric distribution for rank data, whose name arises from both contributions supplied by [3] and [5]. Its probabilistic expression moves from the decomposition of the ranking process in independent stages, one for each rank that has to be assigned, combined with the underlying assumption of standard forward procedure on the ranking elicitation. For this reason, the PL is said to belong to the family of *multistage ranking models*. Specifically, the PL probability distribution is completely specified by the so-called *support parameters* vector $\underline{p} = (p_1, \dots, p_K)$, where $p_i \geq 0$ for all $i = 1, \dots, K$ and $\sum_{i=1}^K p_i = 1$. The generic parameter p_i expresses the probability that item i is selected at the first stage of the ranking (multiple comparison) process and hence preferred among all other items. The probability to choose item i at lower preference levels $j > 1$ is proportional to its support value p_i . Taking into account that the set of available items in the sequence of random selections is reduced by one element after each step, the

computation of the choice probabilities for the assignment of the actual rank requires suitable normalization of the support probabilities w.r.t. the set of remaining items at that stage. In this sense the PL can be regarded as a by-product of an urn sampling scheme without replacement where the vector \underline{p} describes the inclusion probabilities of each item-labelled ball. It follows that under the PL the probability of the random ordering π^{-1} is

$$\mathbf{P}(\pi^{-1}|\underline{p}) = \prod_{t=1}^K \frac{p_{\pi^{-1}(t)}}{\sum_{v=t}^K p_{\pi^{-1}(v)}} \quad \pi^{-1} \in \mathcal{S}_K. \quad (1)$$

2.2 Extension of the Plackett-Luce model

Multistage ranking models, as the PL and related extensions proposed in the literature, implicitly suppose that preferences are expressed with the canonical forward procedure. As explained by [1], this is just a preliminary assumption and other reference orders can be contemplated but, to our knowledge, this aspect has not been addressed in the ranking theory. Indeed, even the individual experience in choice problems suggests the plausibility of alternative paths for the ranking elicitation. For example, one can think of situations where the judge has a clearer perception about her most- and least-liked items first but only a vaguer idea relative to middle ranks; alternatively again the ranker can build up her best alternatives following an exclusion process starting with the final position, which would be described by a backward model. Besides the motivation to characterize typical behaviors in real choice/selection problems, we aim at obtaining a more general tool in order to improve the description of observed phenomena collected in the form of ordered data. Hence, we propose to extend the PL in this way: rather than fixing *a priori* the stepwise order leading the judge to her final ranked sequence, we would like to represent it with a specific free parameter ρ in the model and let data guide inference about the reference order followed in the rank assignment scheme. It turns out that the reference order $\rho = (\rho(1), \dots, \rho(K))$ is the result of a bijection between the stages set S and the ranks set R , i.e.,

$$\rho : S \rightarrow R,$$

where the entry $\rho(t)$ indicates the rank attributed at the t -th stage of the ranking process. Then, ρ identifies a discrete parameter taking values in \mathcal{S}_K . The composition of an ordering with a reference order,

$$\eta^{-1} = \pi^{-1}\rho, \quad (2)$$

yields the sequence listing the items selected at each stage, i.e., $\eta^{-1}(t) = \pi^{-1}(\rho(t))$ is the item chosen at step t and receiving rank $\rho(t)$. The probability of a random ordering under the novel *extended Plackett-Luce model* (EPL) can be written as

$$\mathbf{P}_{EPL}(\pi^{-1}|\underline{\rho}, \underline{p}) = \mathbf{P}_{PL}(\pi^{-1}\underline{\rho}|\underline{p}) = \prod_{t=1}^K \frac{P_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^K P_{\pi^{-1}(\rho(v))}} \quad \pi^{-1} \in \mathcal{S}_K. \quad (3)$$

2.3 Finite mixture modeling for ranked data

We verified the validity of the EPL with an application to the real LFPD data set, which collects the binding measurements of human blood exposed to 11 partially overlapping fragments of the HER2 oncoprotein. Raw quantitative outcomes have been obtained from $N = 67$ samples of human blood taken from three different disease groups: healthy patients, patients diagnosed with breast cancer at an early stage and patients diagnosed with metastatic breast cancer. For reasons due to numerical instability of the measurements and to the absence of universally accepted methods of rescaling the original data, we have verified the possible usefulness of the ranking profiles as a more robust and unambiguously-defined evidence, still capable to capture the sample heterogeneity. Hence, we compared the performance of the EPL in a mixture model setting with the one of alternative and well-studied probability distributions for rankings: the distance-based model and the PL with known forward and backward reference order. Maximum likelihood estimation have been conducted with the implementation of the EM algorithm, or with hybrid versions of it based on the Minorization/Maximization algorithm, following the approach detailed in [2]. BIC values for the mixture of EPL are significantly smaller than those of the competitor models. This indicates the EPL as the best model and proves the successful introduction of the discrete parameter ρ , which drastically improves the fitting of the data. Moreover, the mixture of EPL exhibits a very good accuracy in the discrimination of sample units w.r.t. the real disease status. Finally, our work suggests that even when quantitative data are available in a bioassay experiment, statistical analysis of the underlying ordinal information may provide a more parsimonious and robust tool for the description of the outcome, allowing to partially overcome difficulties related to the preliminary choice of an appropriate normalization technique for the raw numerical responses.

References

1. Fligner, M. A., Verducci, J. S.: Multistage ranking models. *J. Amer. Statist. Assoc.* **83**, 892–901 (1988)
2. Gormley, I. C., Murphy, T. B.: Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A* **169**, 361–379 (2006)
3. Luce, R. D.: *Individual choice behavior: A theoretical analysis*. John Wiley & Sons, New York (1959)
4. Marden, J. I.: *Analyzing and modeling rank data*. Monographs on Statistics and Applied Probability, vol. 64. Chapman & Hall, London (1995)
5. Plackett, R. L.: The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* **24**, 193–202 (1975)

Cluster analysis of three-way atmospheric data

Isabella Morlini and Stefano Orlandini

Key words: Functional data analysis, Climate change, Clustering, Precipitation

1 Introduction

When studying climate change in a spatial area, we may search for typical patterns, common to some time periods, describing the underlying atmospheric process. The analysis and the comparison of these different patterns may give an insight into the long-period changes in meteorological variables, such as rain and temperature. These typical patterns may be thought of as centroids of homogeneous clusters, where the units to be classified are years over a long period of time and the variables are measurements of rain and temperature in different occasions (for example, days or months). Classification of these three-way data (unit \times variables \times occasions) should consider the functional form of the multivariate time series. Indeed, salient features of atmospheric measurements, such as extreme values, maxima or minima, may result shifted in the different series. The transformation of time, that is the warping function from one series to another, must be estimated, before computing the dissimilarity between pairs of series. This function permits a fruitful alignment of the two sequences of measurements. As an example, Fig. 1 reports the daily values of rainfall intensity in the years 1839 and 1841, in the province of Modena (Northern Italy). The two sequences show a great similarity, considering that both years have a peak around 30 mm in March, three days with more than 20 mm in

Isabella Morlini

Department of Economics, University of Modena and Reggio Emilia, Via Berengario 51, 41100 Modena, Italy. Tel. +39-059-206728. e-mail: isabella.morlini@unimore.it

Stefano Orlandini

Department of Mechanical and Civil Engineering, University of Modena and Reggio Emilia, Strada Vignolese 905, 41125 Modena, Italy. Tel. +39-059-2056105 e-mail: stefano.orlandini@unimore.it

the period May-June and, in particular, a very rare event such as a daily value near 80 mm in October. The timing of this very rare event is shifted of 13 days in the two years (it occurs the 16th of October in the year 1839 and the 29th of October in the year 1841). Cross sectional similarities, which compare measurements gathered in the same day, produce pessimistic values for these two series. A more comprehensive similarity should align similar events that occur in nearby days. Even the simplest data analysis, such as computing a mean, can require that features be first aligned by a time transformation, a process that is called time series registration.

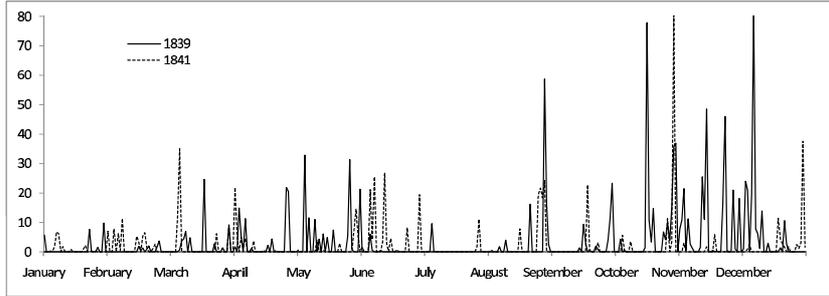


Fig. 1 Daily values of rainfall in Modena in the years 1839 and 1841.

To register time series, in this paper we use the dynamic time warping algorithm (dtw) (Chu *et al.*, 2002; Keogh, 2002; Keogh & Pazzani, 2000; 2001, Wang & Gasser, 1997; 1999). In its original formulation, the dtw estimates a ‘warping path’ for aligning one series to another and minimize a measure of ‘discrepancy’ between the two series which is called dynamic time warping cost. However, if we modify one of the constraints in the classical formulation, the algorithm estimates a path which is a discrete time warping function and minimize an objective function which is a dissimilarity measure between the two registered series.

2 The warping function and the measure of dissimilarity

Suppose we have n observed values x_{ijt} , $i = 1, \dots, n$, of variable j ($j = 1, \dots, p$) at time t ($t = 1, \dots, T$). In its original formulation, given the p -dimensional vector-valued series \mathbf{x}_{1t} and \mathbf{x}_{2t} , where $\mathbf{x}_{it} = [x_{i1t}, \dots, x_{i1t}, \dots, x_{ipt}]$, $i = 1, 2$ and $t = 1, \dots, T$, the dtw first implies the construction of a $T \times T$ square lattice D , in which the element $d(r, c)$ ($r, c = 1, \dots, T$) is the distance $d(\mathbf{x}_{1r}, \mathbf{x}_{2c})$ between the values of series 1 at time r and the values of series 2 at time c . Any distance may be used in the construction of the square lattice D . Each element $d(r, c)$ corresponds to the alignment between points \mathbf{x}_{1r} and \mathbf{x}_{2r} in the p -dimensional Euclidean space. The dynamic time warping cost (dtwc) is defined as follows:

$$dtwc = \min \sqrt{\frac{\sum_{k=1}^K d_k}{K}}, \quad (1)$$

where $T \leq K \leq (2T - 1)$, K is determined by the optimization process of the algorithm and the d_k are elements of D subject to the following constraints:

1. Boundary condition: $d_1 = d(1, 1) = d(\mathbf{x}_{11}, \mathbf{x}_{21})$ and $d_K = d(T, T) = d(\mathbf{x}_{1T}, \mathbf{x}_{2T})$.
2. Continuity constraint: given $d_k = d(\mathbf{x}_{1r}, \mathbf{x}_{2c})$ then $d_{k-1} = d(\mathbf{x}_{1r'}, \mathbf{x}_{2c'})$ where $(r - r') \leq 1$ and $(c - c') \leq 1$.
3. Monotonicity constrain: given $d_k = d(\mathbf{x}_{1r}, \mathbf{x}_{2c})$ then $d_{k-1} = d(\mathbf{x}_{1r'}, \mathbf{x}_{2c'})$ where $(r - r') \geq 1$ and $(c - c') \geq 1$.

In order to define a dissimilarity measure and a warping function, we use a modified parameterized path. This path is characterized by a weaker continuity constraint, defined as follows:

Continuity constraint: given $d_k = d(\mathbf{x}_{1r}, \mathbf{x}_{2c})$ then $d_{k-1} = d(\mathbf{x}_{1r'}, \mathbf{x}_{2c'})$ where $(r - r') \leq 2$ & $(c - c') < 2$ or $(r - r') < 2$ & $(c - c') \leq 2$.

With this continuity constraint, the classical boundary condition and the monotonicity constraint, the dtw algorithm estimates a $wd_i(t)$ warping function, with the following properties:

1. $t_1 < t_2 \Rightarrow wd_i(t_1) \leq wd_i(t_2)$ (the function is monotonic increasing but not strictly increasing and it is not smooth),
2. $wd_i(0) = 0$,
3. $wd_i(T) = T$.

and a dtwc dissimilarity measure, satisfying the following conditions:

1. $dtwc(ii') \geq 0$, $i, i' = 1, \dots, N$ (non negativity);
2. $dtwc(ii) = 0$, $i = 1, \dots, N$ (this a condition weaker than the identity condition required for distance measures),
3. $dtwc(ii') = dtwc(i'i)$, $i, i' = 1, \dots, N$ (symmetry).

The dtwc dissimilarity matrix may be used for classifying time series with hierarchical methods like the complete, the single and the average linkage.

3 Classification of meteorological time series

We perform a cluster analysis of atmospheric measurements gathered by an historical weather station in the urban area of the province of Modena, in the Emilia Romagna Region (Northern Italy). We consider three measurements: X_1 : minimum air temperature (in Celsius degree), X_2 : maximum air temperature (in Celsius degree), X_3 : total rainfall (in mm). We cluster 148 sequences: the years 1861 to 2008. We consider the minimum temperature, the maximum temperature and the total rainfall for the month. We will refer to these data, with $T = 12$, as monthly values

of X_1 , X_2 and X_3 . We set $u = 2$, allowing for a maximum shift of 2 months. Before computing the dtw dissimilarity measure, data are standardized so that, in each t , each variable has 0 mean and unit variance. On the basis of the ratio between the within variance and the total variance (which has a relatively high increase from partition in 6 clusters to partition in 5 clusters) we consider the classification in 6 groups. Cluster means of this partition, obtained with the complete linkage, are reported in Fig. 3. Cluster 1 and 3 are small groups with anomalous years. Cluster 1 contains together former years (the most recent one is 1985), characterized by low maximum temperatures in quite all months, by a very dry summer season and dry months in the second part of autumn. This kind of climate is completely absent in the two last decades. A similar pattern characterizes group 5, in which are clustered several years from 1962 to 1989. In this group, the minimum temperatures are very low, the summer season is dry but the autumn months are extremely wet. Cluster 3 contains together 6 years with a large amount of rain in the summer season and relatively dry spring months. Cluster 6 groups many of the most recent years and the cluster means may be considered as representative of the actual climate situation. This group is characterized by high maximum and minimum temperatures and by a relatively large amount of rain in summer, in autumn and in the beginning of the winter season. The time series of the group means (as long as the composition of the clusters) lead to the evidence that a climate change is present, at the beginning of the 20th century. Both the minima and the maxima temperatures are higher, all over the years, and the seasonality in the rain is less evident, since the average amount of rain shows less variability across months.

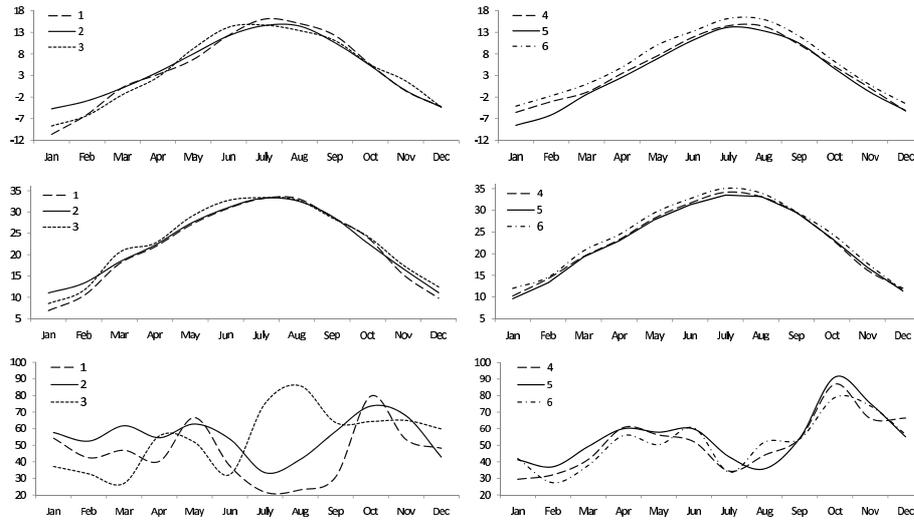


Fig. 2 Data set with $n = 148$: group means in the partition in 6 clusters. Variable X_1 is reported in the top, variable X_2 in the middle and variable X_3 in the bottom.

In order to gain insights into the climate change, we perform a second analysis, considering sequences of 3 years. Each series has 36 monthly values ($T = 36$) and $n = 49$. The first series is the triennium 1861-1863, the last series is the triennium 2005-2007. The label of each series is the second year (for example, for the first series the label is 1862 and for the last series the label is 2006). We consider triennium in order to allow a larger shift in the warping function, and to allow the shift for the month of January (for the second and the third year) and for December (for the first and the second year). Indeed, considering series of 1 year, the warping in the winter months of January and December is not possible. We set $u = 3$ (the same length of a season). We consider the partition in 6 groups obtained with the complete linkage. The cluster means of this partition are shown in Fig. 4 and the group memberships are as follows:

- Cluster 1: {1861, 1900, 1912, 1915, 1918, 1921, 1924, 1930, 1933, 1936, 1939, 1951, 1954, 1960, 1963, 1969, 1972, 1975, 1978}
 Cluster 2: {1864, 1888, 1891, 1897, 1903, 1942, 1945, 1957, 1966, 1984, 1987, 1990}
 Cluster 3: {1879}
 Cluster 4: {1867, 1870, 1885, 1948}
 Cluster 5: {1873, 1876, 1894, 1906, 1909, 1927}
 Cluster 6: {1882, 1981, 1993, 1996, 1999, 2002, 2005}

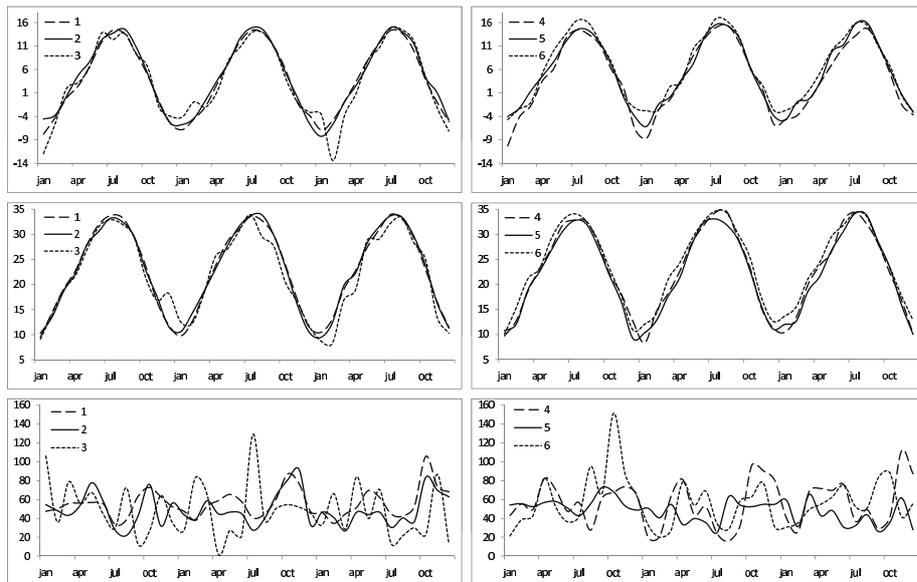


Fig. 3 Data set with $n = 49$, $T = 12$ and $p = 3$: group means in the partition in 6 clusters. Variable X_1 is reported in the top, variable X_2 in the middle and variable X_3 in the bottom.

The triennium 1878-1880 is characterized by extreme (both very high and very low) values in the temperatures and in the rain. Group 4 contains early years and is

characterized by very low temperatures in winter and large amounts of rain in spring and autumn. Groups 1, 2 and 5, contain non-recent years. The time series of the average values of these groups are smoother than the other series: the seasonality in the temperatures is more evident and the amount of rain across months presents less variability. Group 6 contains recent years. The average values of the temperatures (both the minima and the maxima) are higher than the values in the other groups. In particular, the minima temperatures are much higher than in the other groups. The amount of rain is greatly variable across months and shows anomalous peaks in the first year of the triennium. In general, the amount of rain is higher around April and October and the summer months are wetter than in other groups.

The climate change is more evident in this second analysis, since all recent years (after 1991) are clustered together. The group containing these years remains isolated until the last level of the dendrogram and the times series of the average values show peculiar patterns.

References

- Chu, S., Keogh, E., Hart, D. and Pazzani, M. (2002) Iterative Deepening Dynamic Time Warping. *Second SIAM International Conference on Data Mining*.
- Keogh, E. (2002) Exact indexing of dynamic time warping. *28th International Conference on Very Large Data Bases*. Hong Kong. pp 406-417.
- Keogh, E. and Pazzani, M. (2000) Scaling Up Dynamic Time Warping for Data Mining Applications. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, August 20-23. Boston, USA. pp 285-289.
- Keogh, E. and Pazzani, M. (2001) Dynamic Time Warping with Higher Order Features. *First SIAM International Conference on Data Mining (SDM'2001)*, Chicago, USA.
- Wang, K. and Gasser, T. (1997) Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25, 3: 1251-1276.
- Wang, K. and Gasser, T. (1999) Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27, 2: 439-460.

Productivity transition probabilities: A micro level data analysis for Italian manufacturing sectors (1998-2007)

Roberto Nardecchia*, Roberto Sanzo*, Margherita Velucchi[□], Alessandro Zeli[□]

Abstract Italian productivity in the last decade has been declining and widely jeopardized across firms, investment and internationalization modes. This paper analyzes the labour productivity determinants (1998-2007), at firm level, identifying the factors that mainly influenced it, using an original database provided by INSI. We use, firstly, transition matrices to classify firms and, secondly, we run panel regressions to identify the determinants of the success and the failure of Italian firms. From our results emerges the key role of human capital in fostering the Italian productivity.

1 Introduction

In the last decade, Italian productivity was stagnant if not decreasing. Estimates from the Italian National Statistical Institute (ISTAT, 2007) show negative performance of labour productivity, with increasing working hours and low value added performance (Dosi et al., 2012). While several analyses focused on the low level in ITC investments (Bassanetti et al., 2003), confirming the relationship between productivity and ITC investments, others stressed the role of inter-sectoral differences due to small firms' size (Milana et al. 2013). Dosi et al. (2000) stress the importance of basic competences in the firms as well as productivity differentials, strictly linked to internationalization processes (Castellano and Zanfei, 2007). This paper aims at analysing the labour productivity

[□]Corresponding author (margherita.velucchi@unier.it). Università Europea di Roma

[□] Alessandro Z., Istat, Servizio Studi econometrici e previsioni economiche, zeli@istat.it,

* Nardecchia R., e Sanzo R., Istat, Servizio Statistiche strutturali sulle imprese e le istituzioni; email: nardecch@istat.it, sanzo@istat.it

determinants in 1998-2007, at firm level, identifying the factors that mainly influenced it using an original database provided by INSI.

2 Data and the Model

The database used in this paper is a balanced retrospective panel at a micro level developed by ISTAT for Italy (1998-2007). It is a catch-up prospective database, based on cross-sectional enterprises surveys microdata with the integration of administrative source for ensuring the matching of items over time and of eventual non-respondents². At methodological level, we considered the effect of a set of predetermined variables on the labour productivity, analyzing the productivity changes in some specific years (1998, 2002, 2006, 2007). Firstly, we focused on factors that may affected the probability of a change in labour productivity (either positive or negative) in 10, 5, or 1 year, using a transition matrix among quartiles of the productivity distribution (in 1998 and 2007). Secondly, we used an econometric model to study the determinants of labor productivity in the period considered. The model is:

$$lprod_{i,t} = \alpha_{i,t} + \beta_{1,t} MOLFATT_{i,t} + \beta_{2,t} WC_TOT_{i,t} + \beta_{3,t} CWCB_{i,t} + \beta_{4,t} CLDIP_{i,t} + \beta_{5,t} INVMADD_{i,t} + \beta_{6,t} INVMADD(t-2)_{i,t} + \beta_{7,t} INVMADD_{i,t}(t-3) + \beta_{8,t} INVMADD_{i,t}(t-5) + \beta_{9,t} ICTINV_{i,t} + \beta_{10,t} ICTINV_{i,t}(t-2) + \beta_{11,t} ICTINV_{i,t}(t-3) + \beta_{12,t} ICTINV_{i,t}(t-5) + \beta_{13,t} EXP_{i,t} + \varepsilon_{i,t}$$

where: $lprod$ = log of value added per worker, $MOLFATT$ = MOL/sales, $WCTOT$ = white collars/ workers, $CWCB$ = white collars labor cost/blue collars labor cost, $CLDIP$ = labor cost/workers, $INVMADD$ = phisical investment/workers, $ICTINV$ = ITC investments/workers, EXP = *dummy* on export activity.

3 Results

Table 1 reports the estimated transition probability among the quartiles of the labour productivity. As expected and according to the literature (Dosi et al, 2012), the persistence is very high. However, the probability of moving to lower quartiles from 1998 to 2007 is also very high, showing how the labour productivity has declined in the period considered.

Table 1: Transition Matrix (1998 -2007) on quartiles of manufacturing productivity distribution (%)

	1	2	3	4
07				
98				
1	55.0	30.5	10.0	4.7

²For a detailed description of the panel and a deep discussion of its characteristics, see Nardecchia et al. (2010).

2	27.5	33.9	27.9	11.6
3	11.3	24.0	39.2	26.1
4	6.0	11.6	23.0	57.6

The persistence is particularly high in the tails of the distribution: very high and low productive firms do not change much. This is particularly interesting for low productive firms that had low probability of increasing their productivity over this period. Second and third quartiles show higher probabilities and dynamic performance. We may use these results to classify firms as: late-comers (in the lowest two quartiles in both years), dynamic firms (in the lowest two quartiles in 1998, jumping to the fourth in 2007), leaders (in the highest quartile in both years). This classification allows us to disentangle the three groups performance with different time lags (10,5, 1 years) with respect to 2007. In Table 2, we report results. Leaders (C) show the best performance on the majority of the proxies in all sub-periods, particularly on physical investments (per worker), ITC investments and export (dummy). On this, however, the differences with dynamic firms are less evident. According to Meyer and Ottaviano (2007), leaders pay more and have a higher share of white collars (the double of late-comers). Leaders also pay more than late-comers and a leaders' white collar earns 4 times more than a late-comer' blue collar. However, during the period, leaders slowed down and their differences with late-comers lessened. From our results³, we show that dynamic firms have the highest profitability growth rates during the period, rapidly converging to leaders' levels of profitability.

Table 2 : Determinants of Productivity (1998 – 2007)

Variable	Coefficient	Test	p-value
Constant	3.03	33.91	<.0001
MOL/sales	0.121361	30.68	<.0001
White collars/workers	0.007501	0.69	0.4894
White collars cost/blue collar cost	0.000181	2.88	0.004
Labor cost/ workers	0.018108	61.38	<.0001
Physical investments/workers	0.00025	3.4	0.0007
Physical investments/workers (-2)	0.000288	3.65	0.0003

³We run several regressions that we do not report here for space reasons. However, these results are available from authors upon request.

Physical investments/workers (-3)	0.000141	1.77	0.0762
Physical investments/workers (-5)	0.00009	0.94	0.3449
ITC investments/Total investments	-0.00123	-0.12	0.9028
ITC investments/Total investments (-2)	0.003602	0.32	0.7484
ITC investments/Total investments (-3)	0.012918	1.07	0.2831
ITC investments/Total investments (-5)	0.016662	1.14	0.2525
Export (dummy)	0.007719	0.97	0.3309

The results of the panel model in Table 2, show the importance of the composition and competences of the labor force (see the positive and significant effect of white over blue collars cost and the competences of workers on productivity). The relationship between productivity and profitability (MOL/sales) is strong and significant. Physical investments have a large effect that lasts over time. From these results, the most important policy implication, independently of the type of firms, is to invest in human capital. Then, the mix between human and physical capital is peculiar of the type of firms: more human capital for leaders and a weighted average of the two for catching up firms.

References

1. Bassanetti A., Iommi M., Jona Lasinio C., Zollino F. (2003), "The Slow Italian Growth in the 1990s: Is the Gap in Information Technology the Story?", ISTAT – Statistics Finland Workshop "Productivity, Competitiveness and the New Information Economy" Roma , 26 – 27 June.
2. Castellano M, Zanfei A. (2007) "Internationalisation, innovation and productivity: how do firms differ in Italy?" *The World Economy*, 30.
3. Dosi G., Nelson R.R., Winter S. eds (2000). "*The Nature and Dynamics of Organizational Capabilities*". Oxford University Press. Oxford (UK).
4. Dosi G., Grazzi M., Tomasi C., Zeli A. (2012) "Turbulence underneath the big calm? The micro-evidence behind Italian productivity dynamics" *Small Business Economics* vol. 39 November. p. 1043-1067.
5. ISTAT (2007), "*Misure di produttività (anni 1980-2006)*" Statistiche in breve, 5 Ottobre.
6. Mayer, T., Ottaviano, G. M. (2007) The happy few: the internationalization of European firms: new facts based on firm-level evidence. www.bruegel.org
7. Nardecchia R., Sanzo R., Zeli A. (2010) "*La costruzione di un panel retrospettivo di micro-dati per le imprese italiane con 20 addetti ed oltre dal 1998 al 2004*" Documenti Istat n.7.

Interviewers, co-operation and data accuracy: is there a link?

Andrea Neri and Giuseppe Ilardi

Abstract The objective of the paper is to present evidence on the correlation between interviewers' ability in gaining cooperation and their contribution to measurement error bias. We draw on the data from the Italian survey household and income. Our preliminary results suggest that on average, interviewers who are good at recruiting respondents are also good in collecting quality data. Nonetheless, they also tend to spend less time than average to complete an interview. Since data accuracy and interview length are positively associated, some measures like training or a different set of the interviewer's incentives could be used to improve data accuracy.

Key words: Total Survey Error, Interviewer Effect, Non-Response, Measurement Error Bias

1 Introduction

Interviewers are known to contribute to the quality of survey data in a number of ways. First, they vary in their ability to gain cooperation. As a consequence, also their contribution to the nonresponse bias may vary. Moreover, interviewers also vary in their willingness to follow the prescribed interviewing procedure (such as reading questions, recording answers, probing ambiguous responses, providing definitions, etc.). As a consequence, their contribution to measurement error may vary.

Usually, survey agencies tend to evaluate interviewer's effect mainly on the basis of their ability to get high response rates. Other aspects such as data quality are less likely to be used probably because they are more difficult to measure and to extract

Andrea Neri
Bank of Italy, Via Nazionale 91 -00184 Rome, e-mail: andrea.neri@bancaditalia.it

Giuseppe Ilardi
Bank of Italy, Via Nazionale 91 -00184 Rome e-mail: giuseppe.ilardi@bancaditalia.it

quickly enough to be useful in a survey field work. Indeed, to improve survey data quality it would be useful to evaluate interviewers' contribution to total survey error along more than one dimension at the same time. An interviewer may contribute in different directions to the various sources of error. For instance some may be able to gain household co-operation but they could not be good at collecting accurate data during the interview. For others the reverse might be true. Investigating those aspects could have important implications for the recruitment, training and evaluation of interviewers, in order to reduce the total survey error.

Despite the relevance of this topic, only few papers address the role of the interviewer as a potential source of correlation between different types of error. In a recent paper, Brunton-Smith, Sturgis, and Williams (2012) study the association between success in gaining co-operation and interviewer variance in the British Crime Survey. They find that interviewers that are least successful on both contact and co-operation measures exhibit considerably larger variances than the other interviewers.

The objective of the present study is to provide evidence on the correlation between interviewers' ability in gaining cooperation and their contribution to measurement error bias, rather than variability. To better understand the difference with Brunton-Smith, Sturgis, and Williams (2012), let's assume that some interviewers record systematically a "No" answer to a filter question to skip follow-up questions. As a result, interviewers are introducing bias, but their contribution to a variance increase is modest. By only looking at the interviewers' effect on the variability of the variable of interest, no problem would probably emerge.

2 Data and methodology

This study draws on the data from the survey on household income and wealth (SHIW thereafter), conducted by Banca d'Italia (the Italian central bank) to study the economic behaviors of Italian households. The focus of the survey is collecting detailed information on household income, wealth and (to a lesser extent) consumption. Because of the sensitiveness of the issues of the survey, measurement error and unit nonresponse are the main components of total survey error.

The sample size comprises about 8000 households selected from population registers in two stages. Data collection is entrusted to a specialized company using some 190 interviewers with at least two years of experience in face to face interviews. The majority of them (around 60 percent) work in at least two municipalities. Data are collected mainly with the aid of computers, using the Computer-Assisted Personal Interviewing (CAPI) technique (85 percent of cases).

All contact attempts are recorded together with some additional information on all the sampled units. Interviewers are paid on the basis of the number of completed interviews that have passed some quality checks.

We use three variables to classify interviewers' contribution to data quality. The first variable y_1 is the percentage of low-quality responses out of the total responses.

It is based on the following indicators: item nonresponse, number of edits, number of "Don't know / no answer", number of answers in which the respondent provides a qualitative response rather than a punctual one in questions regarding the amount of financial assets held. These measures are proxies of data accuracy (Eurostat, 2007).

A second variable y_2 is the length of the interview. Interviews that are too short or that require a lot of time to complete may suggest data quality problems.

The third variable y_3 is the response rate, defined as the number of complete interviews over contacted households.

In order to disentangle the interviewer's contribution from other influences such as respondent's specific effect, we consider the following model: for an household $h = 1, \dots, n$ and for the interviewer $i = 1, \dots, I$:

$$y_{1hi} = \alpha_{1i} + X'_{1hi}\beta_1 + y_{2hi}\gamma + \varepsilon_{1hi} \quad (1a)$$

$$y_{2hi} = \alpha_{2i} + X'_{2hi}\beta_2 + \varepsilon_{2hi} \quad (1b)$$

$$y_{3hi} = \alpha_{3i} + X'_{3hi}\beta_3 + \varepsilon_{3hi}, \quad (1c)$$

The first equation models the proxy of measurement error y_{1hi} as a function of some characteristics of the interviewer and the household (X_{hi1}), the length of the interviewer y_{2hi} and a random effect at the interviewer's level α_{1i} . The random effect is a measure of the interviewer average distance in terms of data accuracy (and thus in terms of measurement bias) from the others. It can be interpreted as his/her latent ability to collect good quality data.

The second response variable y_{2hi} is a measure of the length of the interview. It is modeled as a function of some observed characteristics X_{2hi} and a random effect α_{2i} . It measures the distance between the time an interviewer spends to complete the interview and the average time required by the others (after controlling for household characteristics). It can be interpreted in terms of patience and ability to take all the time needed to collect good data.

Note that The first two equations are observed only for respondents.

Finally, the third response variable y_{3hi} is a dichotomous variable that takes the value of one if the household participate to the survey and zero otherwise. For identification purposes, the matrix X_{3hi} contains two instrumental variables: the predicted probability of interviewing the selected household and the number of inhabitants of the municipality. This also enables us to take into account that usually survey agency selects the most experienced interviewers in the municipalities where the perceived difficulty for obtaining household cooperation is higher. By including the predicted probability of success, the interviewers performances are compared with the expected response rate of the households he or she has been assigned to. The random effect α_{3i} is a measure of the interviewer's ability to contact and gain household co-operation.

The vectors of unobserved heterogeneity are assumed to be normally distributed. Our interest lies in the estimation in the parameters of the variance-covariance matrix: $\sigma_1, \sigma_2, \sigma_3, \sigma_{31}, \sigma_{32}, \sigma_{21}$. Its components measure the association among the mentioned interviewers' latent abilities. For instance, σ_{31} measures the association

between the ability to gain household co-operation and the ability of collecting accurate data.

The model is completed by assuming that the errors $\boldsymbol{\varepsilon}_{hi} = (\varepsilon_{1hi}, \varepsilon_{2hi}, \varepsilon_{3hi})'$ are distributed according to a tri-variate Gaussian $\mathcal{N}_3(0, \Omega)$, where Ω is an restricted symmetric positive definite matrix. These errors can be interpreted as the effect of factors non related to the interviewer, such as the respondent-specific effects.

3 Results

We find that interviewers play a crucial role in explaining overall variability for each of the three variables considered. The average response rate per interviewer is about 0.55. Some 17 percent of interviewers has a response rate lower than 40 percent, while some 27 percent is higher than 70 percent. Overall interviewers explain about 22 percent of total variability of response rate. A similar situation holds for the data accuracy indicator. The overall average is 5 percent, but about one third of the interviewers have a value below 0.4 percent while for some 7 percent the figure jumps over 7 percent. They are estimated to contribute to 21 percent of total variability. The average length of the interview is about 54 minutes. About 16 percent of interviewers does not spend more than 30 minutes for interviewing their cases, while some interviewers spend more than 70 minutes. About 35 percent of the variability is due to interviewers.

We find that interviewers' latent traits are correlated. Interviewers' ability to gain cooperation is negatively associated with the proxy of measurement error ($\hat{\sigma}_{31} = -0.15$). On average, interviewers who are good at recruiting respondents are also good in collecting good quality data. Moreover, we found the ability to gain cooperation to be negatively associated with the interviewer's patience ($\hat{\sigma}_{31} = -0.22$). Interviewers who are good at recruiting people tend to spend less time in completing the interview. We also find that on average, the willingness to spend more time to complete the interview is also negatively associated with the measurement error proxy ($\hat{\sigma}_{31} = -0.12$). The results suggest that there is room for further improvements in data accuracy, for instance using specific training to increase their degree of patience or by changing their economic incentives.

References

- BRUNTON-SMITH, I., P. STURGIS, AND J. WILLIAMS (2012): "Is Success in Obtaining Contact and Cooperation Correlated With the Magnitude of Interviewer Variance?," *Public Opinion Quarterly*, 76(2), 265–286.
- EUROSTAT (2007): "Handbook on data quality assessment methods and tools," *Working paper*.

Nonhierarchical Asymmetric Cluster Analysis

Akinori Okada and Satoru Yokoyama

Abstract A new procedure of cluster analysis for dealing with asymmetric proximities is introduced, where the similarity from one object to the other object is not necessarily equal to the similarity from the latter to the former. The procedure analyzes one-mode two-way asymmetric proximities among objects to classify objects into clusters, where the number of clusters is determined in advance. Each cluster has its dominant (central) object, and comprises a dominant object and the other objects which belong to the cluster or is dominated by the dominant object. In the present procedure the asymmetry is adjusted by the sum of two corresponding proximities so that the asymmetry is more largely evaluated when the similarity between two objects is larger. The present procedure is applied to car switching data among car segments.

Key words: asymmetry, cluster analysis, k -means clustering, nonhierarchical, similarity

1 Introduction

Relationships among objects are not always symmetric. While asymmetry sometimes can be important to understand relationships among objects, asymmetry has been long ignored in the analysis of data on relationships. Several researchers have paid attention on asymmetry, and have introduced several procedures to analyze and represent asymmetry. There have been developed two sorts of approaches of

Akinori Okada
Graduate School of Management and Information Sciences Tama University, 4-1-1 Hijirigaoka
Tama-shi Tokyo Japan 206-0022, e-mail: okada@rikkyo.ac.jp

Satoru Yokoyama
Department of Business Administration Faculty of Economics Teikyo University, 359 Otsuka Hachioji City Tokyo Japan 192-0395 e-mail: satoru@main.teikyo-u.ac.jp

analyzing asymmetric relationships. One is the procedure based on multidimensional scaling (Borg & Groenen, 2005, Chapter 23). The other is the procedure based on cluster analysis (Okada & Iwamoto, 1996; Takeuchi et al., 2007). While most of procedures based on cluster analysis are agglomerative, and thus are hierarchical necessarily. Olszewski (2011, 2012) introduced nonhierarchical cluster analysis procedures for analyzing asymmetric proximities. Two procedures adopt different approaches. Olszewski (2011) utilized the distance itself, and Olszewski (2012) utilized asymmetric coefficient to deal with asymmetry. Vicari (2012) used two different sorts of clusters for symmetric and skew-symmetric components of the data. In the present study, the asymmetry is evaluated differently compared with earlier studies, and a nonhierarchical clustering procedure based on the evaluation is introduced .

2 The Procedure

The present procedure is a sort of an extension of k -means cluster analysis. Each cluster has its own dominant (central) object and objects which is dominated by the dominant object (Okada & Imaizumi, 2007; Olszewski, 2011, 2012). The number of clusters should be determined beforehand.

Let s_{ik} is the similarity from objects i to k , where s_{ik} is not necessarily equal to s_{ki} . Two terms $s_{ik} - s_{ki}$ and $s_{ki} - s_{ik}$ represent the difference of two similarities having opposite directions, They have the same absolute value, and have opposite signs. When $s_{ik} > s_{ki}$, object k dominates over object i , and when $s_{ki} > s_{ik}$, object i dominates over object k (Tversky, 1977). In the present study, the difference of two similarities $s_{ik} - s_{ki}$ is evaluated after multiplied by $(s_{ik} + s_{ki})$

$$(s_{ik} - s_{ki}) \times (s_{ik} + s_{ki}).$$

Let N is the number of objects, and K is the number of clusters. The problem is to find K objects as dominant objects, and to classify the other $(N - K)$ objects to one of K clusters. Each of $(N - K)$ objects is classified into K clusters. Object i is assigned to the cluster whose dominant object is object k which satisfies

$$\max_{k=1, \dots, K} (s_{ik} - s_{ki}) \times (s_{ik} + s_{ki}).$$

$(s_{ik} - s_{ki})$ shows the skew-symmetry between objects i and k . $(s_{ik} - s_{ki}) \times (s_{ik} + s_{ki})$ shows that the larger the $(s_{ik} + s_{ki})$ is, the skew-symmetry between objects i and k is weighted more. This means that object k dominates over object i more than object ℓ does, when object i is more similar to object k than to object ℓ even if $(s_{ik} - s_{ki}) = (s_{i\ell} - s_{\ell i})$.

The problem is to find K clusters which maximize the goodness of fit (GOF)

$$\text{GOF} = \sum_{k=1}^K \sum_{\substack{i \in \text{cluster } k \\ i \neq k}}^{N_k} \text{signum}(s_{ik} - s_{ki}) [(s_{ik} - s_{ki})(s_{ik} + s_{ki})]^2, \quad (1)$$

where $\text{signum}(s_{ik} - s_{ki}) = 1$ when $(s_{ik} - s_{ki}) > 0$, $\text{signum}(s_{ik} - s_{ki}) = 0$ when $(s_{ik} - s_{ki}) = 0$, and $\text{signum}(s_{ik} - s_{ki}) = -1$ when $(s_{ik} - s_{ki}) < 0$. The number of clusters K is determined in advance, and N_k is the number of objects in cluster k . The method of finding K clusters is (a) to choose all combinations of K objects from N objects, (b) assign each of $(N - K)$ objects to the cluster where $(s_{jk} - s_{kj}) \times (s_{jk} + s_{kj})$ is largest, (d) find clusters which give the largest GOF among ${}_N C_K$ combinations.

3 An Application

The present procedure is applied to car switching data among 16 car segments (Harshman et al., 1982). The data are represented by a 16×16 table. The (i, j) element of the table shows the number of cars corresponding to car segment i which was traded in to purchase cars in car segment j . The table was rescaled by multiplying a rescaling constant to the row and the column so that the sum of row plus column elements is equal over all 16 sums (Okada & Imaizumi, 1987). The 16 car segments are: 1 subcompact domestic (SUBD), 2 subcompact captive imports (SUBC), 3 subcompact imports (SUBI), 4 small specialty domestic (SMAD), 5 small specialty captive imports (SMAC), 6 small specialty imports (SMAI), 7 low price compact (COML), 8 medium price compact (COMM), 9 import compact (COMI), 10 midsize domestic (MIDD), 11 midsize imports (MIDI), 12 midsize specialty (MID), 13 low price standard (STDL), 14 medium price standard (STDM), 15 luxury domestic (LUXD), and 16 luxury imports (LUXI) (Harshman et al., 1982).

The rescaled car switching data were analyzed by the present procedure for $K=2, 3$, and 4. Obtained GOFs for $K=2, 3$, and 4 are; 399.0×10^{15} , 425.0×10^{15} , and 441.3×10^{15} respectively. The three cluster result ($K = 3$) was adopted as the solution from the standpoint of the interpretation and the increment of GOF, and is shown below. The dominant object of each cluster is written in boldface at the front.

Cluster 1; **SUBD**, SMAC, COML, COMM, STDL
 Cluster 2; **SUBI**, SUBC, COMI, MIDI, LUXD, LUXI
 Cluster 3; **MIDS**, SMAD, SMAI, MIDD, STDM

SUBD is the dominant object in Cluster 1. Cluster 1 consists of domestic car segments which seem smaller or less expensive than those in Cluster 3. SUBI is the dominant object in Cluster 2. Cluster 2 consists mainly of imports. One domestic and one captive import car segments (SUBC and LUXI) are in Cluster 2. MIDD is the dominant object in Cluster 3. Cluster 3 consists of mainly domestic car segments, while one import (SMAI) is in the cluster.

4 Discussion

A procedure of nonhierarchical cluster analysis of one-mode two-way asymmetric proximities was introduced, and was applied to car switching data. The present result seems to be compatible with earlier studies. Three dominant objects have the smallest, second, and fourth smallest radii obtained in the asymmetric multidimensional scaling analysis (Okada & Imaizumi, 1987), suggesting the three car segments are dominant in the analysis. It seems that (a) Cluster 1 corresponds with Dimension 1, (b) Cluster 2 corresponds with Dimension 2, (c) Cluster 3 corresponds with Dimension 3 of the analysis. Three clusters are compatible with the (unconstrained) configuration of Zielman & Heiser (1993) as well. SUBI and SUBD seem to be dominant segments in the configuration. The dominant objects SUBD in Cluster 1, SUBI in Cluster 2, and MIDS in Cluster 3 seem to be the smallest and least expensive car segments in each cluster.

The present method of obtaining clusters giving maximized GOF is simple but inefficient, because ${}_N C_K$ combinations should be examined. When the number of objects are large, the present method is not practical. It is preferable to develop a more efficient method.

References

1. Borg, I., Groenen, P.J.F.: *Modern Multidimensional Scaling: Theory and Applications* 2nd edn.. Springer, New York (2005)
2. Harshman, R.A., Green, P.E., Wind, Y., Lundy, M.E. : A model for the analysis of asymmetric data in marketing research. *Marketing Science*. **1**, 205–242 (1982)
3. Okada, A., Imaizumi, T.: Nonmetric multidimensional scaling of symmetric proximities. *Behaviormetrika*. **No. 21**, 81–96 (1987)
4. Okada, A., Imaizumi, T.: Multidimensional scaling of asymmetric similarities with a dominance point. In: Baier, D., Decker, R., Lenz, H. -J. (eds.) *Advances in Data Analysis*, pp. 307-318. Springer, Heidelberg (2007)
5. Okada, A., Iwamoto, T.: University enrollment flow among the Japanese prefectures: A comparison before and after the joint first stage achievement test by asymmetric cluster analysis. *Behaviormetrika*. **23**, 169–185 (1996)
6. Olszewski, D.: Asymmetric k -means algorithm. In: Dobnikar, A., Lotrič., Šter, B. (eds.) *ICANNGA 2011, Part II. LNCS*, 6594, pp. 1-10. Springer, Heidelberg (2011)
7. Olszewski, D.: k -means clustering of asymmetric data. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, Graña, M., Cho, S.-B. (eds.) *Hybrid Artificial Intelligent Systems, Part II. LANI*, 7209, pp. 243-254. Springer, Heidelberg (2012)
8. Takeuchi, A., Saito, T., Yadohisa, H.: Asymmetric agglomerative hierarchical clustering algorithms and evaluations. *Journal of Classification*. **24**, 123–143 (2007)
9. Tversky, A.: Features of similarity. *Psychological Review*. **84**, 327–352 (1997)
10. Vicari, D.: Partitioning asymmetric dissimilarity data. In: *Book of Short Papers of jcs-cladag2012, Analysis of Modeling of Complex Data in Behavioural and Social Sciences*. Retrieved May 10, 2013, from www.jcs-cladag12.tk, (2012)
11. Zielman, B., Heiser, W.J.: Analysis of asymmetry by a slide-vector. **58**, 101–114 (1993)

SuRF: Subspace Ridge Finder

Marco Perone Pacifico

Abstract This paper deals with a nonparametric method for estimating the ridges of a density function. Ridge estimation is useful for understanding the structure of a density. It can also be used to find hidden structure in point cloud data: when the data are noisy measurements of a manifold, under mild conditions the ridges are close and topologically similar to the hidden manifold. We propose a new estimation procedure called SuRF and study its rate of convergence.

Key words: ridges, manifold learning, density estimation, mean shift.

1 Motivation and summary

We present here some results of a paper in collaboration with Chris Genovese, Isabella Verdinelli and Larry Wasserman [6]

A ridge of a density function is a low dimensional set (such as a curve or a surface in a three dimensional space) where the density is sharply peaked in one direction and smooth in the perpendicular dimension. In many problems, multivariate data have some intrinsic low dimensional structure. Examples are clusters, filaments, sheets and so forth. See Figure 1. These features may show up in the density as modes, ridges and hyper-ridges that need to be estimated.

Ridge finding is also relevant in the manifold learning context: if the data are obtained by sampling on an unknown manifold M and adding noise, the manifold can only be estimated at a logarithmic rate (see [5]). We show that the ridge is a *surrogate* for M —meaning that it is close to M and has a similar topology— that can be efficiently estimated.

The goal is to provide a theoretical framework and a practical method for estimating ridges of the unknown density that generated the data.

Marco Perone Pacifico
Sapienza University of Rome, Italy, e-mail: marco.peronepacifico@uniroma1.it

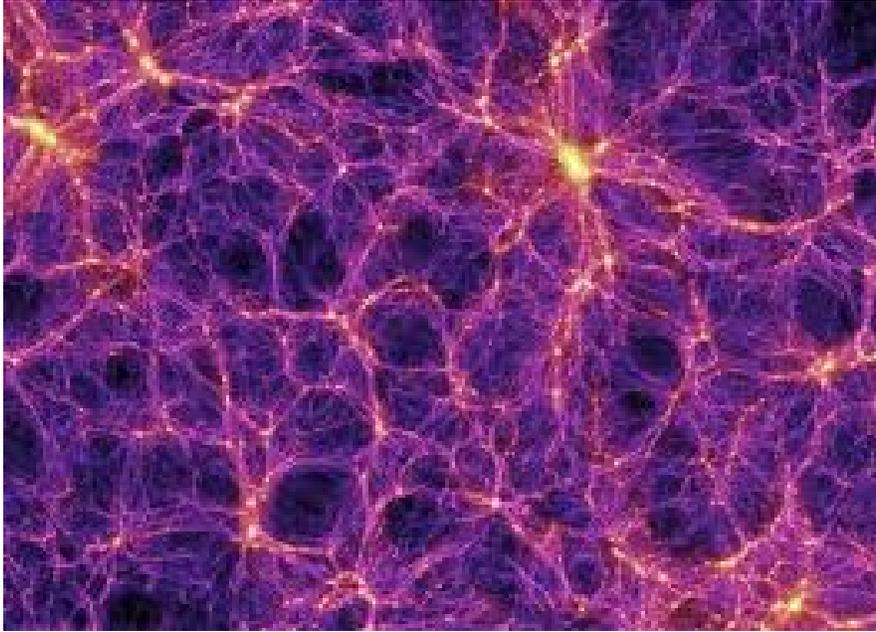


Fig. 1 The Cosmic web. The matter in the universe is concentrated around lower dimensional features: zero dimensional clusters, one dimensional filaments, two dimensional sheets

In the special case when the ridge is zero dimensional, ridge estimation reduces to modes finding and it can be effectively approached using the mean shift clustering algorithm (see [3], [2],[1]). The algorithm starts with a grid of points and moves the points along gradient ascent trajectories towards local maxima.

The mean shift is also the first step of a procedure used in [4] for estimating filaments (one dimensional ridges).

In a recent paper [7], Ozertem and Erdogmus proposed a modified version of the mean shift algorithm, called the subspace constrained mean shift (SCMS), for locating ridges of arbitrary dimension. The SCMS mimics the mean shift algorithm but at each step it replaces the gradient with its projection towards the maximum curvature.

In [7] it is assumed that the underlying density function is known. We, instead, assume that the density is unknown: the only information available is a finite sample of observations generated by the density.

For estimating ridges of an unknown density we propose the SuRF (Subspace Ridge Finder), that consists of applying SCMS to a kernel density estimate. It can be viewed as a sort of “mean shift within local principal components”: at each step points are shifted in the direction of the projection of the gradient in the local principal component space. Figure 2 shows our SuRF estimator at work in the “stylized

cosmic web” example, consisting of intersecting line segments with background clutter.

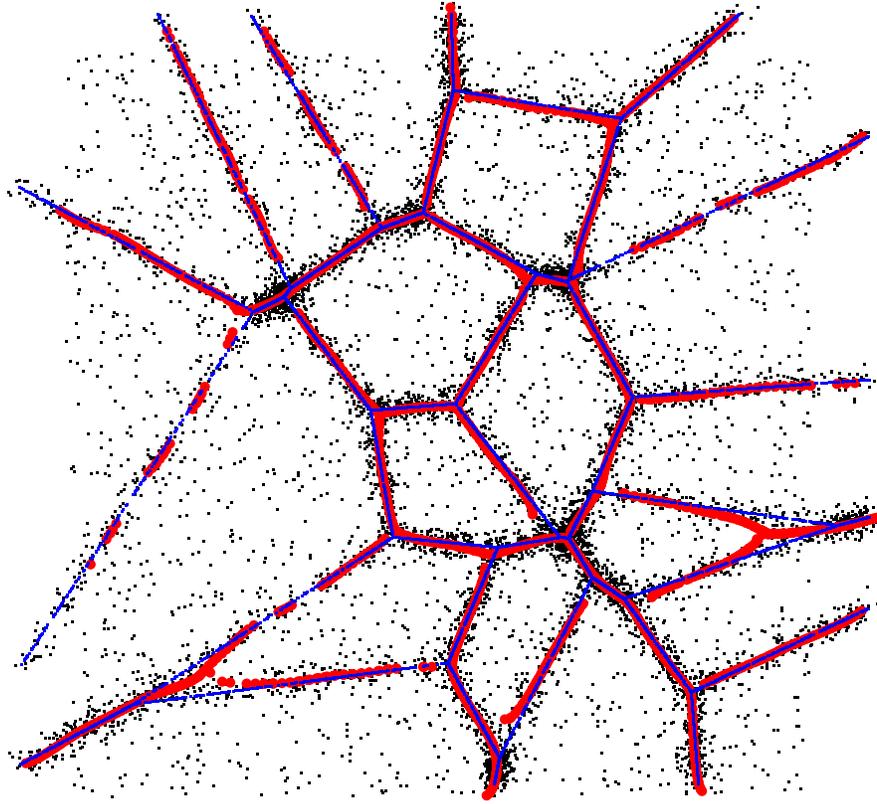


Fig. 2 True (thin blue lines) and estimated (thick red lines) ridges in a two-dimensional dataset generated from a stylized cosmic web. For density estimation we used the Silverman rule (see [8]). The starting points for SuRF are taken on a regular grid. To remove background clutter, we excluded all grid points at which the estimated density is below 5% of the maximum estimated density.

Under suitable smoothness conditions, our main teoretical results are:

1. We give a stability theorem for ridges: we show that if two functions are sufficiently close together, then their ridges are also close together. As a consequence, ridges of an accurate density estimate are close to ridges of the unknown density.
2. We show that there is an estimator \hat{R} such that its Haudorff distance from the true ridge R is

$$Haus(R, \hat{R}) = O_P \left(\left(\frac{\log n}{n} \right)^{\frac{2}{D+8}} \right)$$

where D is the dimension of the space. Further, \widehat{R} is topologically similar to R .

3. We construct an estimator \widehat{R}_h such that

$$Haus(R_h, \widehat{R}_h) = O_P \left(\left(\frac{\log n}{n} \right)^{\frac{1}{2}} \right)$$

where R_h is a smoothed version of the true ridge R .

4. In the manifold learning context, if the noise has small variance σ^2 , then the resulting density has a ridge R_σ such that

$$Haus(M, R_\sigma) = O(\sigma^2 \log^3(1/\sigma))$$

and R_σ is topologically similar to M . It then follows that

$$Haus(M, \widehat{R}) = O_P \left(\left(\frac{\log n}{n} \right)^{\frac{2}{D+8}} \right) + O(\sigma^2 \log^3(1/\sigma)).$$

References

1. Chacón, J.E.: Clusters and water flows: a novel approach to modal clustering through Morse theory. arXiv preprint: arXiv:1212.1384 (2012) <http://arxiv.org/abs/1212.1384>
2. Comaniciu, D. and Meer, P.: Mean shift: a robust approach towards feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**, 603–619
3. Fukunaga, K. and Hostetler, L.D.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, **21**, 32–40 (1975)
4. Genovese, C., Perone Pacifico, M., Verdinelli, I., Wasserman, L.: On the path density of a gradient field. The Annals of Statistics, **37**, 3236–3271 (2009)
5. Genovese, C., Perone Pacifico, M., Verdinelli, I., Wasserman, L.: Manifold estimation and singular deconvolution under Hausdorff loss. The Annals of Statistics, **40**, 941–963 (2012)
6. Genovese, C., Perone Pacifico, M., Verdinelli, I., Wasserman, L.: Nonparametric Ridge Estimation. arXiv preprint: arXiv:1212.5156 (2012) <http://arxiv.org/abs/1212.5156>
7. Ozertem, U. and Erdogmus, D.: Locally defined principal curves and surfaces. Journal of Machine Learning Research, **12**, 1249–1286 (2011)
8. Scott, D.: Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York (1992)

Robust clustering of EU banking data

Andrea Pagano, Francesca Torti, Jessica Cariboni, Domenico Perrotta

Abstract In this paper we present an application of robust clustering to the European Union (EU) banking system. Banking institutions may differ in several aspect, such as size, business activities and geographical location. After the latest financial crisis, it has become of paramount importance for the European regulators to identify common features and issues in the EU banking system and address them in all Member States (or at least those of the Euro area) in a harmonized manner. A key issue is to identify activities, in particular *trading*, which may impact on the stability of the whole EU banking sector. In this paper we show the use of robust clustering for this purpose. The data discussed cover the total volumes and relative share of the trading activities. Data, extracted from the SNL database, includes 245 banks from all EU27 countries, but Estonia, plus a Norwegian bank. At first glance, data appear not showing clear patterns and with standard clustering techniques, such as *k-means*, it is difficult to identify a separation even in this two-dimensional space. Robust clustering, in particular *TCLUST*, has allowed to get a better insight of the dataset.

Andrea Pagano

EC Joint Research Centre, Ispra site, Italy, e-mail: andrea.pagano@jrc.ec.europa.eu

Francesca Torti

EC Joint Research Centre, Ispra site, Italy, e-mail: francesca.torti@jrc.ec.europa.eu

Jessica Cariboni

EC Joint Research Centre, Ispra site, Italy, e-mail: francesca.torti@jrc.ec.europa.eu

Domenico Perrotta

EC Joint Research Centre, Ispra site, Italy, e-mail: domenico.perrotta@ec.europa.eu

1 Problem

The latest financial crisis is driving the scientific community to develop suitable tools to address, among many others, the problem of financial stability. Most of the analyses carried in the academia need to be based on quite large amount of publicly available data. The difficulty of identifying those variables which may be indicators of the fragility of individual institutions, as well as of the system as a whole, asks for the use of statistically sound techniques and procedures. In particular, in this paper we face the issue of classifying bank's riskiness in terms of trading activities based on the assumption that the "modern" banking business model is heavily mixing trading and retail businesses (universal banking model). In several countries (United States, United Kingdom, France, Germany) banking authorities are trying to implement specific regulations aiming at separating trading activities from retail ones, hence one needs to assess the issues of

1. Properly defining trading activities *TradAct*.
2. Evaluating their share over the total assets *ShareTradAct*.
3. Setting suitable thresholds which will divide the 2-dimensional space (*TradAct*, *ShareTradAct*) into two separate zones (banks going under possible structural separation versus banks not going in this direction).

Here we focus on the third issue by means of robust clustering tools.

The banks sample used is extracted from the SNL database, covering 245 EU banks for the years 2006-2011, for which consolidated balance sheet data have been considered. We define trading activities as

$$TradAct = AFS + DA + TSA$$

AFS are total securities designated as available for sale where *DA* are derivatives with positive replacement values not identified as hedging or embedded derivatives and *TSA* are assets part of a portfolio managed as a whole and for which there is evidence of a recent actual pattern of short-term profit-taking, excluding derivatives.

2 Robust clustering of SNL data: first findings

A clustering problem like our is traditionally addressed in the Model-Based Clustering framework i.e. with a finite mixture of distributions, where each mixture component corresponds to a group in the data. A common reference model for the mixture components is the multivariate Gaussian distribution, estimated using the EM algorithm in the popular MCLUST (Fraley and Raftery, 2002). Then, each observation is assigned to the group to which it is most likely to belong. The determination of the right number of groups is still today an outstanding unsolved problem, usually approached with the BIC or AIC criteria.

For our data such models are insufficient, because they do not account for the presence of outliers, which may occur as noise-like structures or as a small tight group of observations in specific areas of the space. In both cases, the presence of outliers can considerably bias the estimation of the centroids and shape (covariance structure) of the groups and seriously affect the final clustering. For this reason, we opted for a robust counterpart of the Normal Mixture Modeling known in the literature as *Robust Trimmed Clustering* or TCLUS (Garcia-Escudero, Gordaliza, Matran and Mayo-Iscar, 2008). The robustness capacity of TCLUS comes from the trimming approach, i.e. the possibility to leave a proportion α of observations, hopefully the most outlying ones, unassigned.

The TCLUS approach is defined through the search of k centers m_1, \dots, m_k and k shape matrices U_1, \dots, U_k solving the double minimization problem:

$$\arg \min_{\mathbf{Y}} \min_{\substack{m_1, \dots, m_k \\ U_1, \dots, U_k}} \sum_{j=1, \dots, k} (x_i - m_j)' U_j^{-1} (x_i - m_j) \quad i = 1, \dots, n \quad (1)$$

where \mathbf{Y} ranges on the class of subsets of size $\lfloor n(1 - \alpha) \rfloor$ within the sample $\{x_1, \dots, x_n\}$. To run the method on our data we used a MATLAB implementation developed in the framework of the FSDA project¹ (Riani, Perrotta and Torti, 2012). The original implementation by the TCLUS authors², available in R, has been also applied with almost identical results.

The choice of the number of groups k is crucial in our case. We have therefore used two approaches. One, recommended by the authors of TCLUS, is the so called Classification Trimmed Likelihood Curves plot (Garcia-Escudero, Gordaliza, Matran and Mayo-Iscar (2011)). The same tool can be also used to refine the choice of the trimming proportion α that, however, should be primarily selected on the basis of the knowledge or prior information one has on the problem. For example, in our case a reasonable choice is $\alpha = 0.04$. The second approach for the choice of k is based on the Forward Search of Atkinson et al. (2004). Originally introduced for detecting masked outliers, the Forward Search is not yet applicable as a fully automatic clustering tool. However, it can be used to infer k by repeating searches from many different randomly chosen subsets. The repeated process reveal the presence of multiple populations as separated peaks in plots monitoring the trajectory of the values of the minimum Mahalanobis distance of observations from the data centroids.

¹ The inclusion of robust clustering tools in the FSDA toolbox for MATLAB is work in progress. The toolbox can be downloaded at these web addresses: <http://www.riani.it> and <https://fsda.jrc.ec.europa.eu>.

² <http://cran.r-project.org/web/packages/tclus/index.html>, CRAN R-package for TCLUS.

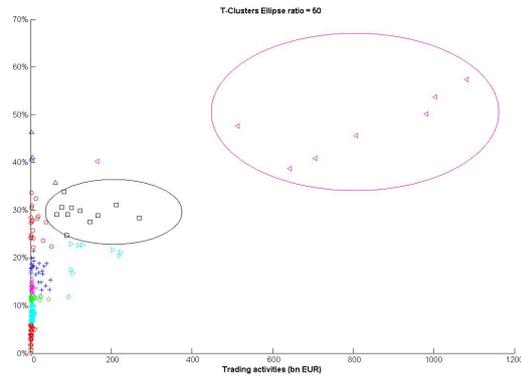


Fig. 1 TCLUS T on SNL data: an example for $k = 8$ groups.

Figure 1 shows the clustering obtained on *TradAct* vs *ShareTradAct* with TCLUS T, using $k = 8$ groups selected on the basis of the inspection of the CTL curves and the Forward Search multiple start plot. The main remarks are:

- There is a subset of banks which are clearly separated from the others.
- There is a tendency of having many banks with similar share values, but different level of trading activities.
- Different choices for the number of clusters to be identified, lead to conclusions slightly different from those in Figure 1.

References

- Atkinson A.C., Riani M. and Cerioli A. (2004). *Exploring Multivariate Data with the Forward Search*, New York: Springer.
- Fraley, C. and Raftery, A. E. (2002). *Model-based clustering, discriminant analysis, and density estimation*, in *Journal of the American Statistical Association*, 97:611–631.
- Riani, M. and Perrotta, D. and Torti, F. (2012). *FSDA: A MATLAB toolbox for robust analysis and interactive data exploration*, in *Chemometrics and Intelligent Laboratory Systems*, 116:17–32.
- Garcia-Escudero L.A., Gordaliza A., Matran C. and Mayo-Iscar A. (2008). *A General Trimming Approach to Robust Cluster Analysis*, *Annals of Statistics*, Vol.36, 1324–1345.
- Garcia-Escudero L.A., Gordaliza A., Matran C. and Mayo-Iscar A. (2011). *Exploring the number of groups in robust model based clustering*, *Statistics and Computing*, Vol 21 (4), 585–599, 2011.

On depth functions for directional data

Giuseppe Pandolfo and Giovanni C. Porzio

Abstract Statistical depth functions are aimed at providing a ranking of the data in order of centrality. The concept of depth can also be extended to directional data, according to a proper notion of center. This work first reviews the few notions of statistical depth functions defined for directional data within the literature. Then, a new class of distance-based depth functions for directional data is introduced.

Key words: Arc distance depth, Circular distance, Rotation invariance.

1 Introduction

Directional data arise in many fields, such as astronomy, earth science, biology and meteorology. Directions are typically in two or three dimensions, but extensions to d -dimensional hyperspheres are of some interest in multivariate analysis as well.

In analogy with depth measures for points in \mathbb{R}^d , data depth for directional data measures the degree of centrality w.r.t. a directional distribution and provides a center-outward ordering on hyperspheres C_{d+1} . Recently, Agostinelli and Romanazzi (2012a) highlighted the great contribution that data depth can give to directional statistics, where no standard ordering is available. They illustrate how a non-parametric analysis of directional data can be obtained using directional depth measures. The main point is that data depth allows to deal with the peculiar aspects of directional data such as the lack of zero-direction and the wrap-around effect.

Giuseppe Pandolfo

Department of Economics and Law, University of Cassino and Southern Lazio, Cassino,
e-mail: g.pandolfo@unicas.it

Giovanni C. Porzio

Department of Economics and Law, University of Cassino and Southern Lazio, Cassino,
e-mail: porzio@eco.unicas.it

Three notions of depth for directional data are available within the literature. Liu and Singh (1992) proposed the *angular Tukey's depth* in analogy with the half-space Tukey depth for multivariate distributions, by replacing halfspaces with hemispheres. They also extended the simplicial depth function to directional data, obtaining the *angular simplicial depth*.

Both these directional depth functions have been extensively studied and used within the literature. Rousseeuw and Struyf (2004) focused on the properties of the angular Tukey's depth, while Agostinelli and Romanazzi (2012b) worked out a local version of the angular simplicial depth. *R* packages allow the computation of both of them in up to three dimensions (www.r-project.org).

However, both the angular Tukey's and the angular simplicial depths are computationally heavy, especially in high dimensions. This is why Ley et al. (2013) recently suggested a new concept of quantiles for directional data. However, it works properly only for a restricted class of distributions (that is, for antipodally symmetric distributions).

With this work, we argue that the *arc distance depth* (Liu and Singh, 1992), the third angular depth available in the literature, has been unduly neglected and it is worth of further investigation. Its computational feasibility make it of considerable practical importance in applications, while offering potential solutions for some interesting inferential procedure (e.g., to test multivariate asymmetry of multidimensional distributions). All this motivates us to introduce a class of depth functions which is a generalization of the arc distance depth. The class is based on distances between points on hyperspheres. By choosing easy-to-compute distances, depth that are computational feasible even in high dimensions are obtained.

Finally, we note that not only distance-based depth functions for multivariate distributions in \mathbb{R}^d already exist, but also that they are experiencing an increasing interest within the literature (see e.g. the recent Dutta and Ghosh, 2012). Their computational feasibility make them a strong competitor with respect to more well-known and used functions.

2 A class of depth functions for directional data

Statistical depth functions measure the degree of centrality of data and provide a ranking according to their values so that a center-outward ordering is defined. While many notions have been presented in the literature, a few are available for directional data: the angular Tukey's, the simplicial, and the arc distance depth.

Here, in analogy with distance-based depth functions for multivariate distributions in \mathbb{R}^d , we introduce a class of distance-based depth functions for directional data. The class is defined as follows.

Definition 1. A directional distance-depth function of the point θ w.r.t. a distribution H on the hypersphere is defined as

$$DD(\theta, H) := \max_{\theta, \phi} \{d(\theta, \phi)\} - E(d(\theta, H)), \quad (1)$$

where $d(\cdot)$ is a directional distance measure, and ϕ is any point on C_{d+1} .

The sample version DD_n 's are obtained by replacing H in (1) with its corresponding empirical version \hat{H}_n .

Many choices for $d(\cdot)$ are available. By choosing Mardia's distance measure $d_M(\theta, \phi) = \pi - |\pi - |\theta - \phi||$, we obtain a depth function that attains its maximum at Mardia's median by definition. Alternatively, Rao's distance measure $d_R(\theta, \phi) = 1 - \cos(\theta - \phi)$, can be adopted. We may define, thus, the following directional distance-based depth functions.

Definition 2. The Mardia distance depth function MDD of the point θ w.r.t. a distribution H on the hypersphere is defined as

$$MDD(\theta, H) = \pi - E(\pi - |\pi - |\theta - H||). \quad (2)$$

Definition 3. The Rao distance depth function RDD of the point θ w.r.t. a distribution H on the hypersphere is defined as

$$RDD(\theta, H) = 2 - E(1 - \cos(\theta - H)). \quad (3)$$

Bounded and non-negative functions. Note that $d_M(\theta, \phi) \in [0, \pi]$ and $d_R(\theta, \phi) \in [0, 2]$. Hence, we have that both $MDD(\theta, H)$ and $RDD(\theta, H)$ are non-negative and bounded.

Rotation invariance. Clearly, both $MDD(\theta, H)$ and $RDD(\theta, H)$ are rotation invariant for any rotation ϑ .

Deepest point. Given that the median direction (see e.g. Mardia and Jupp 2000, page 30), is the direction ϕ that minimizes

$$E(\pi - |\pi - |\theta - \phi||), \quad (4)$$

we have that $MDD(\cdot)$ attains its maximum value at the circular median. As a consequence, the MDD deepest point inherits all the properties of Mardia's median (e.g., uniqueness, influence function, and so on...).

Finally, we note that the MDD is equal to the arc distance depth by definition, while the RDD should be more directly related to the von Mises (or circular normal) distribution. In that respect, RDD may be well suited to describe von Mises distributions just as the Mahalanobis depth is suited for multivariate normal distributions.

To provide a first idea of the behaviour of MDD_n and RDD_n , we simulated data from von Mises distributions with mean $\mu = \pi$ and concentration parameter $\kappa \in \{0, 0.5, 2\}$, and computed their value over a grid of points on a circle. Results are offered in Figure 1. When $\kappa = 0$ (that is, for a uniform distribution on the circle), the functions are substantially flat. This corresponds to a desirable property for any directional depth function. For $\kappa > 0$, both the functions attain their maximum value

in π , as expected. Moreover, at first glance it seems that RDD is able to reproduce the shape of the underlying density. This behaviour, that may turn out to be useful in some applications, is definitely worth of further analysis.

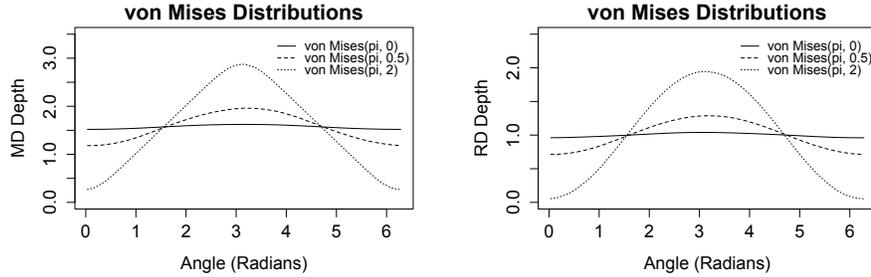


Fig. 1 MDD (*left*) and RDD (*right*) for some von Mises distributions with mean $\mu = \pi$.

At the end, we believe interest should arise for distance-based directional depth functions. A deeper analysis of their properties and of their potential applications, along with a comparison with existing directional functions, is currently under study.

References

1. Agostinelli, C., Romanazzi, M.: Nonparametric analysis of directional data based on data depth. *Ecol Stat* (2012a) doi: 10.1007/s10651-012-0218-z
2. Agostinelli, C., Romanazzi, M.: Depth analysis of directional data. Proceedings of the 46th Scientific Meeting of the Italian Statistical Society, CLEUP, Padova (2012b)
3. Dutta, S., Ghosh, A.K.: On classification based on L^p depth with an adaptive choice of p . Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India (2012)
4. Ley, C., Sabbah, C., Verdebout T.: A new concept of quantiles for directional data. ECARES working paper, No. 23. Brussels, Belgium: Université Libre de Bruxelles (2013). Available from <http://www.ecares.org>
5. Liu, R.Y., Singh, K.: Ordering directional data. Concepts of data depth on circles and spheres. *Ann Stat* **20**, 1468–1484 (1992)
6. Mardia, K.V., Jupp, E.P.: *Directional statistics*. Wiley, Chichester (2000)
7. Rousseeuw, P.J., Struyf A.: Characterizing angular symmetry and regression symmetry. *J Stat Plan Infer*, **122**, 161–173 (2004)

A generalised Silhouette-width measure

Andrea Pastore and Stefano F. Tonellato

Abstract An extension of the Silhouette-width measure, capable of taking into account the different metrics characterising the distinct components of a Gaussian mixture model, is proposed. In a simulated example we show that such dissimilarity measure leads to conclusions different than those suggested by the Silhouette width.

Key words: Gaussian mixture model, silhouette

1 Introduction

Gaussian finite mixture models (GMM) are widely used for clustering purposes. Commonly, maximum likelihood estimates of the parameters are achieved via the Expectation-Maximization (EM) algorithm, whereas the number of components of the mixture is determined via the Bayesian Information Criterion (BIC) [3]. Usually, each cluster is associated with a different component of the fitted GMM, implicitly assuming that the optimal number of groups is equal to the number of components.

Such correspondence has been criticised under different aspects. On one hand, some authors [1] argue that GMM might lead to poorly separated components, and hence to overestimate the number of clusters. On the other hand, it has been proved that the number of modes of a GMM might be greater than the number of components [7]. Therefore, an estimated GMM might have some modes located far from the mean of each of its components. Thus, if clusters are to be located around the modes of a probability distribution, GMM based clustering might underestimate the number of groups, under some particular conditions. These criticisms suggest that

Andrea Pastore

Department of Economics, Ca' Foscari University of Venice, e-mail: pastore@unive.it

Stefano F. Tonellato

Department of Economics, Ca' Foscari University of Venice, e-mail: stone@unive.it

some care should be taken in identifying each cluster with a component of the fitted GMM. A rather common approach to cluster evaluation is based on the definition of suitable measures of within-cluster heterogeneity and of between-clusters separation (see for instance, [4], ch. 3). These measures can be seen either as optimality criteria for a classification problem, or as diagnostic tools for evaluating a given clustering solution. The Silhouette Width (SW) measure [5] is a widely adopted index which evaluates the quality of a partition, accounting for both heterogeneity within each cluster and separation among clusters. The definition and the use of the SW measure for partitions obtained from a GMM are not trivial when the variance matrices of the components are not equal.

In this paper we propose an extension of the SW measure that can be used as a diagnostic tool for evaluating the quality of a GMM-based clustering. The differences with the original SW measure are illustrated with an example.

2 GMM-Silhouette Width

Let $U = \{u_i, i = 1, \dots, n\}$ be the set of objects we want to classify. Moreover, let $\mathcal{P}_g = \{C_1, \dots, C_g\}$ be a generic partition of U in g groups and n_t the cardinality of C_t , $t = 1, \dots, g$. We shall denote by $z_i \in \{1, \dots, g\}$ the i -th membership variable, i.e. $z_i = t \Leftrightarrow u_i \in C_t$. Assume that \mathcal{P}_g represents the clustering solution based on the GMM

$$\sum_{t=1}^g \pi_t N[\mu_t, \Sigma_t],$$

and obtained by applying the maximum a posteriori rule (MAP).

If u_i and u_j have been assigned to the same group, say C_t , then a dissimilarity measure consistent with the model is the squared Mahalanobis distance (SMD) induced by the t -th component in the mixture model:

$$d_{i,j,t} = (x_i - x_j)' \hat{\Sigma}_t^{-1} (x_i - x_j), \quad (1)$$

where $\hat{\Sigma}_t$ represents an estimate of Σ_t . If, instead, $z_i \neq z_j$, the SMD would make sense only if $\hat{\Sigma}_{z_i} = \hat{\Sigma}_{z_j}$. Let us define the following average dissimilarities between u_i and C_t :

$$\bar{d}_{i,t} = \begin{cases} 0 & \text{if } t = z_i \text{ and } n_t = 1 \\ \frac{1}{n_t - 1} \sum_{j: z_j = t} d_{i,j,z_i} & \text{if } t = z_i \text{ and } n_t > 1 \\ \frac{1}{n_t} \sum_{j: z_j = t} d_{i,j,z_i} & \text{if } t \neq z_i \end{cases} \quad \text{and} \quad \tilde{d}_{i,t} = \frac{1}{n_t} \sum_{j: z_j = t} d_{i,j,z_j}$$

measured with the metric induced by the z_i -th and the z_j -th components respectively. We can then define a Gaussian mixture model silhouette-width (GMM-SW) $s_i = \frac{b'_i - a'_i}{\max(a'_i, b'_i)}$, where $a'_i = \bar{d}_{i,z_i}$ and $b'_i = \min_{t \in \{1, \dots, g\} \setminus \{z_i\}} \tilde{d}_{i,t}$. The GMM-SW is a gen-

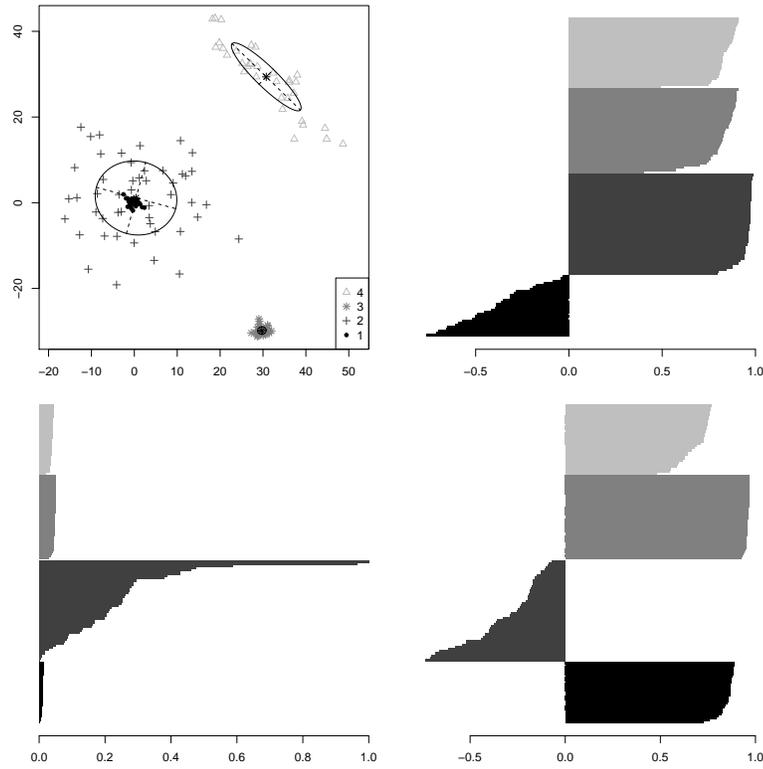


Fig. 1 Simulated dataset: data scatterplot, GMM-SW, dbs and SW measure

Table 1 Simulated dataset: some statistics for GMM-SW, SW and dbs measure

group	GMM-SW		SW		dbs	
	mean	median	mean	median	mean	median
1	-0.40	-0.39	0.85	0.87	0.01	0.01
2	0.95	0.97	-0.32	-0.25	0.24	0.21
3	0.82	0.84	0.96	0.96	0.05	0.05
4	0.83	0.83	0.69	0.73	0.04	0.04
overall	0.63	0.86	0.47	0.75	0.10	0.04

eralization of the SW, since it coincides with the SW when $\Sigma_t = I$ (I denotes the identity matrix) for $t = 1, \dots, g$.

3 An example

We generated a sample of size 150 from a four-component bivariate GMM with $\pi_1 = \pi_2 = \pi_3 = \pi_4$, $\mu_1 = \mu_2 = (0, 0)'$, $\mu_3 = (30, -30)'$, $\mu_4 = (30, 30)'$, $\Sigma_1 = \Sigma_3 = I$, $\Sigma_2 = 100 \cdot \Sigma_1$, whereas Σ_4 has diagonal elements equal to 100 and off-diagonal elements equal to -95 . The top-left panel of Figure 1 shows the scatterplot of the simulated dataset with the profile of the estimated variance matrices. For this dataset, we computed the GMM-SW, the SW, using the euclidean distance as dissimilarity measure, and the density based silhouette measure (dbs) [6]. The dbs is an alternative to the SW for density-based clustering techniques. Let $\tau_{i,t}$ be the posterior probability, provided by the fitted GMM, that the i -th object belongs to the t -th group, and let $q_i \in \{1, \dots, g\}$ be the group label maximising $\tau_{i,t}$, subject to the constraint $t \neq z_i = \arg \max\{\tau_{i,s}, s = 1, \dots, g\}$. The dbs for u_i is:

$$dbs_i = \frac{\log\left(\frac{\tau_{i,z_i}}{\tau_{i,q_i}}\right)}{\max_{j \in \{1, \dots, n\}} \left| \log \frac{\tau_{j,z_j}}{\tau_{j,q_j}} \right|}.$$

The other panels of Figure 1 show the GMM-SW plot (top-right), the dbs plot (bottom-left) and the SW plot (bottom-right), while Table 1 gives some numerical results. In this case, the SW, based on the euclidean distance, suggests that the group associated with the second component is very heterogeneous and poorly separated from the group associated with the first component. Moreover, the groups associated with the first and the third components have a quite similar silhouette, and seem to be better characterised, in terms of internal homogeneity and separation from the other clusters, than the group associated to the fourth component. The GMM-SW, taking into account the estimated variance matrices of the different components, leads to opposite conclusions, accordingly with the dbs measure.

References

1. Biernacki, C., Celeux, G., Govaert, G.: Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood, *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 719–725 (2000)
2. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: *Cluster Analysis*, 5th ed. Wiley, New York (2011)
3. Fraley, C., Raftery, A.E.: Model-Based Clustering, Discriminant Analysis, and Density Estimation, *Journal of the American Statistical Association*, **97**, 611–631 (2002)
4. Gordon, A.D.: *Classification* (2nd ed.). Chapman & Hall, Boca Raton (1999)
5. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York (1990)
6. Menardi, G.: Density-based Silhouette diagnostics for clustering methods. *Statistics and Computing*, **21**, 295–308 (2011)
7. Ray, S., Ren, D.: On the upper bound of the number of modes of a multivariate normal mixture, *Journal of Multivariate Analysis*, **108**, 41–52 (2012)

Hospital efficiency under two competing panel data models

Fulvia Pennoni and Giorgio Vittadini

Abstract Motivated by an application to a longitudinal dataset regarding hospitals in Lombardy we propose a model to measure hospital efficiency which incorporates observed covariates and time-varying latent hospital effects. The distribution of the latter may be continuous-valued or discrete-valued. In the first case it is a mixture of auto-regressive AR(1) processes with specific mean values and correlation coefficients and common variances. In the second case it is based on a first order homogeneous Markov chain with a fixed number of states. Maximum likelihood estimation of the model parameters is performed by using the expectation-maximization algorithm and the Newton-Raphson algorithm. The effect of the different formulations is evaluated in terms of the estimated efficiency scores.

Key words: diagnosis-related groups, heterogeneity, latent Markov model, latent auto-regressive model

1 Introduction

In recent years several models have been proposed to address the problem of evaluating hospital efficiency. One important feature of the healthcare system is the introduction of a new perspective hospital reimbursement regulating the hospital compensations for the different treatments provided which is based on the diagnosis-related groups (DRG) [1] [2]. According to this system hospitals receive a fixed rate

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca (IT) e-mail:
fulvia.pennoni@unimib.it

Giorgio Vittadini
Department of Statistics and Quantitative Methods, University of Milano-Bicocca (IT) e-mail:
giorgio.vittadini@unimib.it

for each admission depending on a patient's diagnosis. As a consequence, hospitals face an increased pressure on their financial performance and a risk of insolvency.

Among the models developed to address the problem of efficiency the stochastic frontier model proposed by [3] has been extended in several directions (see among others [4]). The model proposed by [5] has the advantage to allow for time-varying efficiency components. Recent reviews of such models may be found in [6] and [7].

Motivated by an applications concerning hospitals in Lombardy to examine if the hospitals have efficiency gains during the period 2008-2011 we propose a model based on two different formulations of the distribution of the latent process in order to properly model the unobserved heterogeneity. The latter is due to the fact that the hospital general manager is indeed lawfully responsible for all the activity performed in her/his hospital. In such context experience, ability and skills are important features to be considered. The proposed flexible structure for the latent process allows to get efficiency scores which are less biased compared to other methods given that they vary in a fashion way across time capturing those events in the hospitals which are not observed through the covariates.

2 The data

The data derive from a large administrative data base provided by the Lombard Health Care Department regarding hospital's features. The data cover the full population of patients for the general medicine ward which is one with the highest discharges and number of beds compared to the other wards in the region. They are related to 120 hospitals and cover the period 2008-2011.

One response variable is the number of outpatient discharges which considers the existing relation between patients and inputs. Another response variable linked with the previous one is the yearly revenues from discharges. The latter includes the monetary incentives coming from admitting patients. It is given by the product of the DRG tariff times the yearly number of discharges in that diagnosis related group. Since the DRG tariff is a function of the treatment's complexity, revenues take also into account the severity of health care procedure provided to the patient.

Table 1 shows some descriptive statistics about each response variable and of the available covariates. The latter are the DRG weight capturing the treatment complexity, the number of beds, the number of working hours for physicians, nurses, and other employees and the hours of activity of the surgery rooms.

3 The proposed model

With reference to a sample of n hospitals observed at T time occasions, let y_{it} be the response variable for hospital i at occasion t and let x_{it} be a corresponding column vector of covariates, with $i = 1, \dots, n$ and $t = 1, \dots, T$. We also denote by

Variable	Year			
	2008	2009	2010	2011
discharges (number)	1,374.91	1,353.87	1,348.33	1,250.68
revenues (Euro)	4,036,772.73	4,070,066.67	4,246,381.74	4,050,475.38
DRG weight	1.05	1.04	1.05	1.05
beds (number)	45.07	44.88	44.28	43.61
physicans (hours)	243,269.59	244,657.00	211,976.51	205,590.11
nurses (hours)	475,809.23	479,607.26	394,430.63	342,241.89
others (hours)	456,053.75	454,783.34	306,586.58	155,523.67
surgey rooms (hours)	7,639.82	7,624.77	8,091.62	7,882.73

Table 1 Distribution of variables over the time occasions.

$\mathbf{y}_i = (y_{i1}, \dots, y_{iT})$ the vector of response variables and by $\mathbf{X}_i = (\mathbf{x}_{i1} \cdots \mathbf{x}_{iT})$ the matrix of time-varying covariates for hospital i .

The model we formulate is based on the assumption that $y_{it} = G(y_{it}^*)$, where y_{it}^* follows the model

$$y_{it}^* = \alpha_{it} + \mathbf{x}_{it}'\boldsymbol{\beta} + \eta_{it}, \quad i = 1, \dots, n, t = 1, \dots, T,$$

with η_{it} being independent error terms with a standard logistic distribution, and $G(\cdot)$ is a link function which models the relationship between each response variable y_{it} and the corresponding latent variable α_{it} and the vector of covariates \mathbf{x}_{it} . The distribution of the latent variable may be based on a continuous or on a discrete latent process. The discrete latent process formulation assumes that, for all i , $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iT})$ follows a first-order homogenous Markov chain with k states denoted by ξ_1, \dots, ξ_k . This chain has initial probabilities $\boldsymbol{\pi}_h$ and transition probabilities $\boldsymbol{\pi}_{h_1 h_2}$, with

$$\boldsymbol{\pi}_h = p(\alpha_{i1} = \xi_h), \quad h = 1, \dots, k, \quad (1)$$

$$\boldsymbol{\pi}_{h_1 h_2} = p(\alpha_{i,t-1} = \xi_{h_1}, \alpha_{it} = \xi_{h_2}), \quad h_1, h_2 = 1, \dots, k, t = 2, \dots, T. \quad (2)$$

It is assumed that every α_{it} is conditionally independent of $\alpha_{i1}, \dots, \alpha_{i,t-2}$ given $\alpha_{i,t-1}$, but apart from this assumption, the distribution of $\boldsymbol{\alpha}_i$ is unconstrained. In such case a latent Markov (LM) model with covariates results; see [8] for a review.

The continuous latent variable formulation assumes instead the existence of a discrete latent variable u_i with k support points and mass probabilities $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_k$ such that when $u_i = h$ the latent process has a distribution given by a mixture of AR(1) processes. We assume that

$$\alpha_{i1} = \xi_h + \varepsilon_{i1}, \quad i = 1, \dots, n,$$

and that

$$\alpha_{it} = \xi_h + (\alpha_{i,t-1} - \xi_h)\rho_h + \varepsilon_{it}\sqrt{1 - \rho_h^2}, \quad i = 1, \dots, n, t = 2, \dots, T,$$

where $\varepsilon_{it} \sim N(0, \sigma^2)$ for all i and t and (ξ_h, ρ_h) are parameters which for $h = 1, \dots, k$ are estimated jointly with the common variance.

Maximum likelihood estimation of the model parameters is performed by a joint use of the expectation-maximization algorithm and of the Newton-Raphson algorithm as proposed by [9] and standard errors for the parameter estimates are obtained by exploiting the observed information matrix. The number of mixture components is selected by considering the BIC index and another criterion which takes into account the level of separation of the mixture.

Acknowledgements We acknowledge “Finite mixture and latent variable models for causal inference and analysis of socio-economic data” (FIRB - Futuro in ricerca) funded by the Italian Government (RBFR12SHVV).

References

1. Berta, P., Callea, G., Martini, G., Vittadini, G. (2010). The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: a population-based investigation. *Economic Modelling*, **27**, 812-821.
2. Herwartz, H., Strumann, C. (2013). Hospital efficiency under prospective reimbursement schemes: an empirical assessment for the case of Germany. *European Journal of Health Economics*, doi: 10.1007/s10198-013-0464-5.
3. Aigner, D.J., Lovell, C.A.K. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, **6**, 21-37.
4. Green, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, **126**, 269–303.
5. Battese, G.E., Coelli, T.J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, **20**, 325–332.
6. Green, W. (2009). The econometric approach to efficiency analysis Ch 2, in H.O., Fried, C.A.K. Lovell, and S.S. Schmidt (Eds.), *The measurement of productive efficiency - Techniques and Applications*, 92–251, Oxford University Press, Oxford.
7. Kumbhakar, S.C., Lien, G. and Brian J. (2012). Technical efficiency in competing panel data models: a study of Norwegian grain farming. *Journal of Productivity Analysis*, doi: 10.1007/s11123-012-0303-1.
8. Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press: Boca Raton.
9. Bartolucci F., Bacci S., Pennoni F. (2013). Longitudinal analysis of self-reported health status by mixture latent autoregressive model. *Journal of the Royal Statistical Society - Series C*, in press.

The Interval-Wise Control of the Family-Wise Error Rate for Testing Functional Data

Alessia Pini and Simone Vantini

Key words: Functional Data, Inference, Family Wise Error Rate, Permutation Test.

1 Introduction

In many current fields of research, due to the development of data acquisition and storage technologies, the problem of the analysis of high-dimensional data have enhanced. An example of this situation is constituted by functional data, i.e., random functions lying in an infinite-dimensional space [4]. The major issue for the statistical analysis of this kind of data is represented by the fact that many classical inferential tools are not generally suited for the analysis of functional data, as they require the number of sample units to be greater than the dimension of the space in which inference has to be carried out, which is, in this case, infinite.

In order to deal with this problem, we observe that, if the functional space is a separable Hilbert space (e.g., L^2 , which is often assumed to be the space in which data are defined [4]), hypotheses to be tested can always be stated as intersections of a countable infinity of marginal hypotheses pertaining components. Indeed, it is always possible to choose a countable functional basis, and express the hypothesis test in terms of tests on the basis coefficients. Following this approach, the methods dealing with this problem may be classified in two different categories, depending on the type of error control, and subsequent decision, that has to be provided.

As a first approach to the problem, the control of the global level of the test, coinciding with the weak control of the Family Wise Error Rate (FWER) may be required. In this case, a unique global test is performed, and the result is the decision of whether there is sufficient evidence to reject the global null hypothesis.

Alessia Pini and Simone Vantini

MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, - 20133 Milan, Italy e-mail: alessia.pini@polimi.it; simone.vantini@polimi.it

A more stringent requirement is the strong control of the FWER. Techniques that focus on this type of control are based on the multiplicity correction of component-specific univariate statistics, assuring the control of the level of the test for each possible set of true null hypotheses (e.g., Bonferroni-Holm correction [1], Closed Testing Procedure [2]). These procedures, unlike the ones based on a global test, enable the selection of the significant components. Nevertheless, they are generally not suited to deal with high-dimensional data. Indeed, as the dimension increases, their computational cost might explode and/or their power can become very low.

The Interval Testing Procedure (ITP) that we briefly present here and is described in detail in [5], lies in between these two different approaches. Indeed, the type of control provided is intermediate between the weak and strong control of the FWER, that is, an interval-wise control of the FWER. The leading idea is to develop a new methodology that assures a statistically convenient error control which is stronger than the weak control of the FWER, has a power which is comparable with the one provided by the global inference techniques, and is computationally affordable.

2 The Interval Testing Procedure

The ITP is a non-parametric procedure that is developed in a very general framework. In particular, it enables to test for differences between two functional populations, to test the central tendency of one functional population, or to test differences among several functional populations. The procedure is constituted by three steps:

- A high-dimensional functional basis is selected (e.g., Fourier, B-splines) and data are represented by means of the coefficients of the basis expansion.
- The significance of each basis component is tested with a univariate permutation test on the associated coefficients. All tests are performed jointly, i.e., the permutations are the same on each component.
- The results of the joint univariate tests are used to build multivariate permutation tests on all intervals of basis components. In particular, the significance of each closed interval is tested, and the resulting p -value heatmap is used to correct the univariate p -values, to obtain suitably adjusted p -values.

This approach provides the selection of the statistically significant basis components. Indeed, the interval-wise control of the FWER allows to control, for all possible closed intervals of basis components, the probability of rejecting at least one null hypothesis if all the hypotheses of the interval were true. This type of control is more focused on intervals of components than on sparse subsets, and it is thus particularly suited for functional data. Indeed, all functional bases commonly used in FDA present a natural ordered structure: Fourier components are ordered according to frequency, B-spline components according to the abscissa, wavelet components according to the abscissa and the frequency, Taylor components (or polynomial-inspired components in general) according to roughness, functional principal components according to variance.

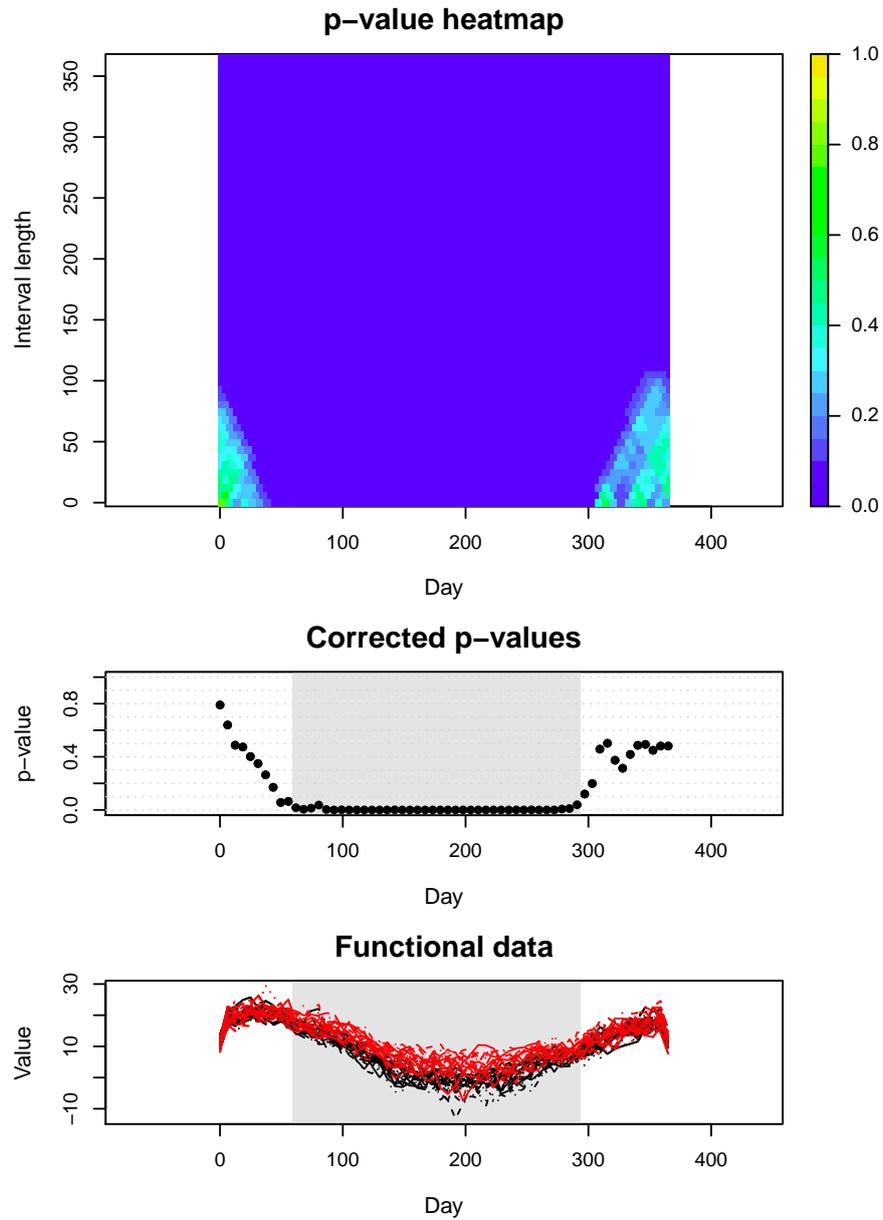


Fig. 1 Results of the B-spline based ITP on the NASA daily temperatures data: p -value heatmap (top); plot of the corrected p -values of each basis component (middle); functional data associated to the Milan (black) and Paris (red) data (bottom). The shaded part in the middle and bottom graphs is the area selected as significant at a 5% level from the ITP.

As an example, we hereby report an application where the B-spline representation is used. In this case, because of the compact support of the B-spline basis elements, we have the control of the FWER on intervals of the domain. Thus, this property allows to select, in case of rejection, the significant intervals of basis functions. Figure 1 reports the result of the B-spline-based ITP to test the difference between two populations on a case study. Data (displayed on the lower panel of Figure 1) are mean daily temperatures in Milan and Paris registered from July 1983 to June 2005 and stored in the database NASA *Earth Surface Meteorology for Solar Energy*. In the application reported, we identified the 22 years as sample units and the 365 records available for each year as 365 point-wise evaluations of the functional data. We tested for differences between Milan (black curves) and Paris (red curves) temperature profiles in the coupled scenario. The upper panel of Figure 1 shows the p -value heatmap resulting from the third phase of the ITP, and the middle panel of the same Figure reports the corrected p -values. The shaded part represents the area rejected by the ITP at a 5% level, that is, the components with a corrected p -value lower than 5%.

The test results shows a significant difference between the two populations in the autumn and winter seasons, when the temperature in the Milan area is lower than the one in the Paris area. On the other hand, in the spring and summer period, no difference is detected, meaning that the average temperatures on the two areas are in this period the same. To have a clearer idea of the action of the interval-wise control of the FWER in the practice, here it allows to state that, if there were no differences in the autumn and winter period, we would have erroneously selected this period as significant with a 5% probability. Hence, we may state that we have enough statistical evidence to trust the test results, i.e., to state that the two populations are statistically different on the autumn-winter period. In addition, this last condition holds for any closed period of the year, e.g., any season, month, week, and in general for every time interval.

References

1. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* **6**, 65-70 (1979).
2. Marcus, R., Eric, P., Gabriel, K.R.: On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika* **73** No.3, 655–660 (1976).
3. Pesarin, F., Salmaso, L.: *Permutation Tests for Complex Data*. John Wiley & Sons (2010).
4. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer, New York (2005).
5. Pini, A., Vantini, S.: The Interval Testing Procedure: Inference for Functional Data Controlling the Family Wise Error Rate on Intervals. Technical Report **13**, MOX (2013).

Detecting differences between primary schools in mathematics and reading achievement by using schools added-value measures of performance

Mariano Porcu and Isabella Sulis

Abstract This paper try to jointly identify plausible factors which influence fifth grade pupils' achievement in math and reading according to the results of the tests carried out in 2010 by the Italian Institute for the Evaluation of the School System. The work has four main aims: (i) to detect performance potential confounding factors observed at pupils' and school's level; ii) to assess how confounders shape performances in reading and math; (iii) to suggest value-added indicators of the contribution that schools give to students' competencies; (iv) to detect schools characterized by outstanding performances.

Key words: multilevel, IRT, INVALSI

1 Introduction

A long series of quantitative papers has looked into the relationship between the socio-cultural background of individuals, their school environment and their educational attainment.

Many studies made use of individuals (or families) participation in the so called highbrow cultural activities [5], others consider individuals' reading attitudes [9], or cultural resources at home (e.g., books, computers, etc.), or involvement in extra-curricular activities (e.g., foreign language classes, music classes, etc.) [2]. Researches have been also carried out to study which factors make a school successful in terms of policies, resources and practices [3].

Mariano Porcu
Università degli Studi di Cagliari, e-mail: mrporcu@unica.it

Isabella Sulis
Università degli Studi di Cagliari e-mail: isulis@unica.it

Using data provided by the Italian Institute for the Evaluation of the School System we try to evaluate the effect of the Potential Confounding Factors (PCF) [4] on the pupils' achievement in maths and reading considering the hierarchical structure of the data. After performing an explorative analysis of the sources of heterogeneity using *variance components models* to evaluate the intensity of the 'intra-class' variability of pupils' achievement at class, school, county, regional and macro-regional areas, we assess the effect of both relevant pupils' details and socio-cultural characteristics, as well as schools compositional variables [7] that should be specifically considered in the joint analysis of students' achievement in math and reading literacy tests. Section 2 discusses methods applied to model the data presented in Section 3. Section 4 contains the results of the analysis.

2 Methods

We adopt a bivariate two-level model with heteroskedastic random term at level-2 in order to jointly analyze the effect of pupils, families and schools' covariates on pupils' achievement in mathematics and reading.

Indicating with $Y_{ij}^{d=1,2}$ the achievement (calculated through an Item Response Theory – IRT – scoring model) of pupil i in school j in math and in reading, we model the joint density function as follows:

$$Y_{ij}^{(d)} = \alpha^{(d)} + x_{ij}^{(d)} \beta^{(d)} + z_j^{(d)} \gamma^d + \zeta_j^{(d)} + \varepsilon_{ij}^{(d)} \quad (1)$$

where d indicates to which test the score the variable refer to ($d = 1$ math, $d = 2$ reading), $\varepsilon_{ij} \sim N(\mathbf{0}, \mathbf{\Omega})$ is a bivariate vector of random terms which take into account of within-school individual variability in mathematics and reading test, whereas $\zeta_j \sim N(\mathbf{0}, \mathbf{\Phi})$ is a multivariate normal vector of random terms which take into account of factors which can affect the between-school variability in both test scores, and x a z are two vectors of covariates at individual-level and school-level. In the simplest hypothesis in which at level-two it is considered just a random intercept (to take into account for between-school variability), the multivariate random term $\zeta_j \sim N(\mathbf{0}, \mathbf{\Phi})$ is specified as bivariate. The model has been fitted with *Stata* using the *runMLwiN* routine implemented by Leckie and Charlton [10] and adopting the Iterative Generalized Least Squares estimation method. The values of the random terms are estimated in a second step from their posterior distribution after plugging-in in the likelihood function the values of $\hat{\beta}$, $\hat{\gamma}$ and of the parameters of the variance-covariances of the random terms ($\hat{\Phi}$, $\hat{\Omega}$). At school-level, the expected posterior prediction of the two random-intercepts $\tilde{\zeta}_j^{(d)}$ is considered the adjusted indicator of school added-value on the d th discipline (Reading, Mathematics). The assessment of the performance of a school with respect to the overall average performance is made by considering whether or not the confidence interval of $\tilde{\zeta}_j^{(d)}$ overlaps with the value 0 [6]. Whenever the interest of the evaluators is mainly addressed to compare the performance of two schools with each other (rather than

with respect to the average), the criterion for evaluating if the two schools are significantly different is to verify if their confidence intervals overlap [6].

3 Data

Data here analyzed come from the sample survey carried out on May 2010 by the Italian Institute for the Evaluation of the School System (INVALSI). The survey is administered yearly to assess the skills and knowledge of the Italian students enrolled in private or public schools. The INVALSI survey makes use of two different supports to collect the information needed: the pupil's *biographic form* and the pupils's questionnaire

In the following analysis only data referred to fifth-level pupils have been considered. Specifically, we retained just the records containing information in the fields concerning pupils' family background and both math and reading tests. The final dataset contains 34,554 records (pupils). For each student we have considered from the *biographic form* the following information: gender (SEX: 1=F); whether or not the pupil is an Italian native (NATIVE: 1=Native Italian); parents' education level (MEDU, FEDU: 1=Primary, 2=Secondary, 3=Tertiary); language spoken at home (HLANG: 1=Italian); number of siblings (SIBLIN: 0, 1, 2 or more – ordinal); whether or not the pupil begun primary school after the regular age (REGULAR: 1=Regular); the attendance of a preschool education program (PRESCH: 1=Yes); whether or not the pupil attends a *full-time* school program (TIME: 1=No – i.e. up to 30 hours per-week). The pupils' questionnaire gathers the responses to the test sheets in maths and reading. Maths test sheet contains 44 items while the reading one contains 69. A score for both math and reading has been calculated using a scaling procedure based on Item Response Theory (IRT) model. Higher values of both indicators indicates better performances in the two disciplines. Likewise, using IRT models we summarized the responses provided to dichotomous and ordinal indicator variables contained in the biographic form and addressed to collect information on: *Pupil's home resources* – HOMRES; *Pupil's perception of test climate* – TESTCL; *Pupil's self confidence* – SELCON; *Pupil's self commitment in school activity*; *Pupil's attitude towards mathematics and reading* – MATHLIKE - READLIKE; *Pupil's perception of school environment* – NENVSC. Higher values on these indexes indicate a greater availability of home resources, a higher self-confidence, a better test climate; a greater attitude towards math and reading; a more negative school environment. Considering the responses of pupils to the questionnaire the following variables have been considered: the number of book at home (BOOK: 1=0-10, 2=11-25, 3=26-100, 4=100+) along with the variable related to the time spent in reading for amusement (JOYREAD: 1=Never, 2= less than 1h/day, 3=1-2 h/day, 4=more than 2h/day). Finally, at school level, we have defined the following compositional variables: *Rate of pupils native Italian* – SCHNATIVE; *Mean of NENVSC variable observed for the school* – SCHNENV; *Rate of pupils with at least one parent's education level at ISCED 5* – SCHIS5; *Rate of pupils with both*

parents' education level at ISCED 3 – SCHIS3; *Mean size of the pupils' families* – SCHFAMSIZE; *Mean number of books available at home for the school pupils* – SCHBOOK. This last indicator variable has been built by averaging the mean value of each category of the variable BOOK.

4 Application

In order to detect relevant levels in the hierarchical structure of the data we carried out an explorative analysis of the sources of heterogeneity to evaluate the intensity of the intra-cluster variability of pupils' achievement at class, school, county, region and macro-region levels. Results shows that the highest level of heterogeneity in both math and reading scores is due to the clustering of pupils in schools (respectively 23.7% for math and 14.2% for reading). The geographical components at district (province) and macro-region (Main Islands, South, Center, North-East, North-West) levels do not seem to explain a significant amount of variability on the mean of pupils' score. However, as outlined by previous works carried out on the INVALSI data on math for 2009 by Grilli and Sani[7] and Agasisti and Vittadini [1], if we allow the differences in the average score at school-level to differ across geographical areas, it arises that the between-school within-area variability in the South is much stronger than in the North (at least three times bigger). This result holds for both tests, but it is stronger for the math one. Thus, we relaxed the assumption of homoscedasticity of the variance of the random terms across geographical areas, allowing the second level error $\zeta_j^{(g)} \sim N(0, \psi^{(g)})$ to have different variability across the areas [7]. Specifically, considering the alike values assumed by the parameters in North-East and North-West as well in South and Main Islands, these four areas have been aggregated in North and South. The multivariate normal random term at level two has been specified as a six-variate vector of random terms which take into account of the between-schools variability in mathematics and reading test within the three main geographical area.

Throughout a stepwise selection procedure we have selected the covariates that significantly influence pupils' performances at pupils' and school' level to consider in the Bivariate Multilevel Model with Heteroscedastic Random terms at Level-Two. Results are reported in Table 4. The estimates of the coefficient parameters have to be interpreted considering that the range of variation of the two indicators of pupils' performances in math and reading (summarized by the IRT tools) is, respectively $[-3.488 \div 2.425]$ and $[-4.197 \div 2.714]$, and that both variables are approximately normal distributed. The most interesting evidence of the results is that most of the pupils' socio-cultural characteristics in almost one of the two measures of performance are significant whereas some of the compositional variables built up at level of school to contextualize the school are not (or, surprisingly, they change the direction of their effect). The coefficients related to the demographic characteristics show that females perform worst than male in math, whereas no significant differences arise in reading competences [11]. Being a native (NATIVE) Italian and to use

the Italian as language in family (HLANG) positively influence both tests. If we look to the covariates related to the educational background it arises that pupils which attended a preschool program (PRESCH) and that did not delay their entrance at school (REGULAR) perform better (the effect of being 'regular' is stronger on reading). As expected, all the proxies of the cultural resources of the families, such as the number of books available at home (BOOK), the parents' education (MEDU, FEDU), the attitude to read books for amusement (JOYREAD) positively influence pupils' performances. The home resources in terms of facilities and access to the technologies, expressed throughout the variable HOMRES, have a significant but weaker effect and only on reading performances. It is interesting to see that the confidence intervals of the different levels of mother and father's education do not overlap; furthermore the confidence interval of the effect of having a graduated mother lays completely above the upper bound of the confidence interval for the effect of having a graduated father. The joint effect of having both parents graduated increases the score in math and reading of 0.615 and 0.403, respectively. Moreover, if we consider the effect of having both parents graduated together with the effect of the availability at home of more than 100 books and of the habit of reading more than 2 hours per day, we can observe an increasing of 1.318 on reading and of 0.796 on math performances. Another interesting results is that the size of the family, measured by the number of the siblings, has a significant negative effect just on reading test. In making comparisons across institutions on the basis of the schools' added value measures, the introduction in the model of the covariates such as READLIKE, MATH- LIKE, SELCON, COMMIT, and NENVSC allows to control the effect of some PCF which are external to the process under evaluation. Considering the variables observed at school level we note a significant effect of the variable SCHBOOK (a proxy of the family cultural resources), whereas the rate of family with at least a graduated parent (SCHIS5) lost is significant effect once that we control it at level-1. Thus, if we compare the two schools with the minimum and maximum observed value of the variable SCHBOOK (7-160), a pupil from the latter would perform an average score in reading 0.31 higher. It is interesting to pay attention to the sign of the school level covariates SCHNATIVE and SCHFAMSIZE which seems to deploy a negative effect just on the reading performances. Noteworthy is the (strong) and unexpected negative sign of SCHNATIVE coefficient: likely it masks the effect of geographical factors for which we do not have information (rural vs urban areas).

References

1. Agasisti, T. and Vittadini, G.: Regional Economic Disparities as Determinants of Students' Achievement in Italy. *Research in Applied Economics*. **41**, 33–54 (2012)
2. Covay, E., Carbonaro, W.: After the bell: participation in extracurricular activities classroom behavior, and academic achievement. *Sociology of Education*, **83**, 1, 20–45 (2010)
3. Dobbie, W., Fryer, Roland G.Jr.: Getting Beneath the Veil of Effective Schools: Evidence from New York City. NBER Working Papers, 17632 (2011)
4. Draper, D., Gittoes, W.: Statistical analysis of performance indicators in UK higher education. *J. R. Stat. Soc. A*, **167**, 3, 449–474 (2004)

Table 1 Bivariate multilevel model with two random effects at level-two

covariates	READING			MATH		
	$\hat{\beta}$	p-value	95% CI	$\hat{\beta}$	p-value	95% CI
cons	-1.363	0.000	-1.720 -1.007	-1.203	0.000	-1.570 -0.833
SEX	0.017	0.155	-0.006 0.041	-0.150	0.000	-0.169 -0.129
NATIVE	0.323	0.000	0.271 0.376	0.187	0.000	0.143 0.230
HLANG	0.062	0.000	0.031 0.093	0.049	0.000	0.023 0.074
PRESCH	0.156	0.001	0.066 0.245	0.130	0.001	0.056 0.204
REGULAR	0.257	0.000	0.163 0.352	0.124	0.002	0.045 0.202
TIME	0.051	0.005	0.015 0.086	-0.056	0.001	-0.080 -0.024
JOYREAD.2	0.088	0.000	0.057 0.119	0.063	0.000	0.030 0.088
JOYREAD.3	0.195	0.000	0.159 0.231	0.072	0.000	0.042 0.102
JOYREAD.4	0.253	0.000	0.206 0.299	0.070	0.000	0.031 0.108
MEDU.2	0.172	0.000	0.144 0.200	0.139	0.000	0.110 0.161
MEDU.3	0.320	0.000	0.277 0.363	0.243	0.000	0.207 0.279
FEDU.2	0.132	0.000	0.104 0.160	0.088	0.000	0.064 0.111
FEDU.3	0.232	0.000	0.188 0.276	0.160	0.000	0.123 0.196
BOOK.2	0.127	0.000	0.083 0.171	0.118	0.000	0.080 0.154
BOOK.3	0.285	0.000	0.242 0.328	0.239	0.000	0.202 0.275
BOOK.4	0.383	0.000	0.337 0.429	0.323	0.000	0.284 0.361
SIBLIN.2	-0.046	0.005	-0.079 -0.014	0.032	0.018	0.005 0.059
SIBLIN.3	-0.114	0.000	-0.151 -0.077	-0.004	0.793	-0.034 0.026
HOMRES	0.027	0.007	0.007 0.046	0.013	0.099	-0.000 0.029
NENVSC	-0.162	0.000	-0.181 -0.143	-0.143	0.000	-0.159 -0.126
READLIKE	0.051	0.000	0.034 0.068	-0.055	0.000	-0.069 -0.041
MATHLIKE	0.149	0.000	0.132 0.166	0.306	0.000	0.291 0.320
SELCON	0.080	0.000	0.053 0.106	0.077	0.000	0.055 0.099
COMMIT	0.202	0.000	0.178 0.227	0.135	0.000	0.115 0.155
SCHNENV	-0.163	0.021	-0.300 -0.025	-0.338	0.000	-0.483 -0.192
SCHBOOK	0.002	0.001	0.001 0.003	0.001	0.005	0.000 0.003
SCHISS	-0.216	0.182	-0.534 0.102	-0.061	0.721	-0.392 0.271
SCHNATIVE	-0.454	0.000	-0.691 -0.216	-0.111	0.387	-0.358 0.139
SCHFAMSIZE	-0.140	0.005	-0.239 -0.042	0.023	0.672	-0.082 0.127

Random effects Parameters					
	Level 2: School		Level 2: Pupils		
	$\hat{\Phi}$	p-value	$\hat{\Omega}$	p-value	
$\sigma_{d_1}^2$	0.040	0.005	$\sigma_{d_1}^2$	0.738	0.007
$\sigma_{d_1}^2$ ₈₁	0.086	0.012	σ_{d_1, d_2}	0.327	0.004
$\sigma_{d_1}^2$ ₈₂	0.166	0.014	$\sigma_{d_2}^2$	0.504	0.004
$\sigma_{d_1}^2$ ₈₃	0.051	0.005			
$\sigma_{d_2}^2$ ₈₁	0.129	0.016			
$\sigma_{d_2}^2$ ₈₂	0.284	0.022			
$\sigma_{d_2}^2$ ₈₃	0.024	0.004			
$\sigma_{d_{12}}^2$ ₈₁	0.076	0.012			
$\sigma_{d_{12}}^2$ ₈₂	0.135	0.015			
$\sigma_{d_{12}}^2$ ₈₃					

5. Flere, S. *et al.*: Cultural Capital and intellectual ability as predictors of scholastic achievement: a study of Slovenian secondary school students. *British Journal of Sociology of Education*, **31**, 1, 47–58 (2010)
6. Goldstein H., *Multilevel Statistical Models*, 4th ed., Wiley, Chichester (2011)
7. Grilli, L., Sani, C.: Differential Variability of Test Scores Among Schools: a Multilevel Analysis of the Fifth-grade INVALSI Test Using Heteroscedastic Random Effects. *Journal of Applied Quantitative Methods*, **6**, 4, 88–99 (2011)
8. INVALSI: Servizio Nazionale di Valutazione a.s. 2009-10. Rilevazione degli apprendimenti. Scuola Primaria. Prime Analisi. http://www.invalsi.it/download/rapporti/snv2010/Rapporto_\SNV_09_10.pdf. Cited 8 Feb 2013 (2011)
9. Jæger, M.M., Holm, A.: Does parents' economic, cultural, and social capital explain the social class effect on educational attainment in the Scandinavian mobility regime?. *Social Science Research* **36**, 7, 19–44 (2007)
10. Leckie, G., Charlton, C.: runMLwiN - A Program to Run the MLwiN Multilevel Modelling Software from within Stata. *Journal of Statistical Software*, **52**, 11, 1–40 (2013)
11. Matteucci, M., Mignani, S.: Gender differences in performance in mathematics at the end of secondary school in Italy. *Learning and Individual Differences*, **21**, 543–548 (2011)

Outlier Detection via Contaminated Mixture Distributions

Antonio Punzo, Paul D. McNicholas, Katherine Morris and Ryan P. Browne

Abstract Contaminated mixture distributions have are parameterized to indicate the proportion of outliers and the degree of contamination. By their nature, they present a natural method for outlier detection and are very attractive for mixture model-based clustering and classification. The first contribution of this paper is to introduce a mixture model whereby each mixture component is itself a contaminated Gaussian distribution. To introduce parsimony, a family of fourteen mixtures of contaminated Gaussian distributions is developed by applying constraints to eigen-decomposed component covariance matrices. This approach is, amongst other things, an effective alternative to trimmed clustering. An expectation-conditional maximization (ECM) algorithm is used to find maximum likelihood estimates of the parameters and thereby give classifications for the observations. The second contribution of this paper is to introduce a mixture model whereby each mixture component is itself a shifted asymmetric Laplace distribution. This approach allows the possibility to carry out robust clustering when there is skewness present in the data. Again, an ECM algorithm is used for parameter estimation. Our novel approaches are applied to artificial and real data in order to illustrate some of the advantages. Amongst them, and in contrast to the trimmed clustering approach, we have: 1) each observation has a posterior probability of belonging to a particular group and, inside each group, of being an outlier or not, 2) the models do not require pre-specification of quantities such as the proportion of observations to trim, 3) the approach can be easily used in high dimensions, 4) model-based classification is permitted in addition to clustering, and 5) (in the second contribution only) we can account for non-elliptical clusters.

Key words: Clustering, mixture models, outliers, robust.

A. Punzo
University of Catania, Catania, Italy. e-mail: antonio.punzo@unict.it

P.D. McNicholas, K. Morris and R.P. Browne
University of Guelph, Guelph, Ontario, Canada.
e-mail: {pmcnicho,kmorri09,rbrowne}@uoguelph.ca

1 Introduction

Finite mixtures of distributions are commonly employed in statistical modelling with two different purposes (Titterington *et al.*, 1985, pp. 2–3). In *indirect applications*, they are used as semiparametric competitors of nonparametric density estimation techniques (see Titterington *et al.*, 1985, pp. 28–29, McLachlan and Peel, 2000, p. 8, and Escobar and West, 1995). On the other hand, in *direct applications*, finite mixture models are considered as a powerful device for clustering, classification, and discriminant analysis by assuming that one or more mixture components represent a group (or class or cluster) within the original data (see McLachlan and Basford, 1988 and Fraley and Raftery, 1998).

For continuous multivariate random variables, attention is commonly focused on mixtures of Gaussian distributions because of their computational and theoretical convenience. Unfortunately, real data are often “contaminated” by outliers that affect the estimation of the component means and covariance matrices (see, e.g., Bock, 2002). Thus, the detection of these outliers, and the development of robust methods of parameter estimation insensitive to the presence of outliers, are important practical problems. Following Gallegos and Ritter (2009), the mixture modelling literature on this topic can be summarized as follows (for the alternative trimmed clustering approach see, e.g., García-Escudero *et al.*, 2008, 2010).

1. Campbell (1984), McLachlan and Basford (1988, Section 2.8), and De Veaux and Krieger (1990) use M-estimates of the means and covariance matrices of the Gaussian components of the mixture model.
2. McLachlan and Peel (1998) and Peel and McLachlan (2000) introduce mixtures of t -distributions (see also Greselin and Ingrassia, 2010 and Andrews and McNicholas, 2011).
3. Fraley and Raftery (2002) add, to the mixture of Gaussian distributions, a uniform component on the convex hull of the data in order to accommodate outliers.

The performance of these aforementioned methods is analyzed in Hennig (2004).

4. Browne *et al.* (2012) introduce a mixture model whereby each mixture component is itself a mixture of a Gaussian and a uniform distribution.

However, these mixture-based approaches have some drawbacks. In direct applications, the first two methods do not allow for the direct detection of outliers. The approaches considering the uniform distribution, if used for discriminant analysis, cannot recognize a new noisy observation (an observation that has not been used to fit the model) if it lies outside the support defined by the fitted uniform distribution(s); this is paradoxical because the new observation should be the strongest available outlier in the philosophy of the corresponding model. Finally, in indirect applications, mixtures having one or more uniform distributions do not provide an overall smooth density, which is a fundamental requirement in the nonparametric paradigm (Silverman, 1981).

To overcome these problems, a mixture of contaminated Gaussian distributions is proposed in Section 2.1. A contaminated Gaussian distribution is a two-component

Gaussian mixture in which one of the components, with a large prior probability, represents the “good” observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the “bad” observations (Aitkin and Wilson, 1980). It represents a common and simple theoretical model for the occurrence of outliers. Furthermore, parsimonious variants of the proposed model are introduced in the fashion of Celeux and Govaert (1995) by imposing constraints on eigen-decomposed component covariance matrices. The most general model-based classification framework is then considered and an expectation-conditional maximization (ECM) algorithm for parameter estimation is outlined. Then, the shifted asymmetric Laplace distribution is introduced and an analogous approach is taken for mixtures of asymmetric Laplace distributions. The Bayesian information criterion (BIC; Schwarz, 1978) and the integrated completed likelihood (ICL; Biernacki *et al.*, 2000) are compared for model selection for our novel families of mixtures of contaminated distributions. Applications on artificial and real data are presented and a comparison with trimmed clustering, as implemented in the `tclust` package (Fritz *et al.*, 2012) of R (R Core Team, 2013), is discussed.

2 Mixtures of Contaminated Gaussian Distributions

2.1 The general model

The distribution of a p -variate random vector X , according to a parametric finite mixture model with k components, can be written as

$$p(x; \Psi) = \sum_{j=1}^k \pi_j f(x; \vartheta_j), \quad (1)$$

where π_j is the weight (mixing proportion) of the j th component, with $\pi_j > 0$ and $\sum_{j=1}^k \pi_j = 1$, $f(x; \vartheta_j)$ is the parametric (with respect to ϑ_j) distribution associated with the j th component, and $\Psi = \{\pi, \vartheta\}$, with $\pi = \{\pi_j\}_{j=1}^k$ and $\vartheta = \{\vartheta_j\}_{j=1}^k$, contains all of the parameters of the mixture. As usual, model (1) implicitly assumes that the component distributions should all belong to the same parametric family.

In this paper, as component density in (1), we adopt the *contaminated Gaussian distribution*

$$f(x; \vartheta_j) = \alpha_j \phi(x; \mu_j, \Sigma_j) + (1 - \alpha_j) \phi(x; \mu_j, \eta_j \Sigma_j),$$

where $\alpha_j \in [0, 1]$, $\eta_j > 0$, $\vartheta_j = \{\alpha_j, \mu_j, \Sigma_j, \eta_j\}$, and

$$\phi(x; \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} \delta(x, \mu; \Sigma)\right\} \quad (2)$$

is the distribution of a p -variate Gaussian random vector with mean μ and covariance matrix Σ . In (2),

$$\delta(x, \mu; \Sigma) = (x - \mu)' \Sigma^{-1} (x - \mu)$$

denotes the squared Mahalanobis distance between x and μ with covariance matrix Σ . The result is the *mixture of contaminated Gaussian distributions*, given by

$$p(x; \Psi) = \sum_{j=1}^k \pi_j \left[\alpha_j \phi(x; \mu_j, \Sigma_j) + (1 - \alpha_j) \phi(x; \mu_j, \eta_j \Sigma_j) \right], \quad (3)$$

where $\Psi = \{\pi, \alpha, \vartheta\}$, with $\alpha = \{\alpha_j\}_{j=1}^k$. Previous work on mixtures of Gaussian mixtures can be found, for example, in Orbanz and Buhmann (2005) and Di Zio *et al.* (2007).

References

- Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, **22**(3), 325–331.
- Andrews, J. L. and McNicholas, P. D. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**(3), 361–373.
- Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(7), 719–725.
- Bock, H. (2002). Clustering methods: From classical models to new approaches. *Statistics in Transition*, **5**(5), 725–758.
- Browne, R. P., McNicholas, P. D., and Sparling, M. D. (2012). Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(4), 814–817.
- Campbell, N. (1984). Mixture models and atypical values. *Mathematical Geology*, **16**(5), 465–477.
- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.
- De Veaux, R. D. and Krieger, A. M. (1990). Robust estimation of a normal mixture. *Statistics & Probability Letters*, **10**(1), 1–7.
- Di Zio, M., Guarnera, U., and Rocci, R. (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data analysis*, **51**(5), 2573–2585.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.
- Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, **41**(8), 578–588.

- Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.
- Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012). **tclust**: an R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47**(12), 1–26.
- Gallegos, M. T. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhyā: The Indian Journal of Statistics, Series A*, **71**(2), 164–220.
- García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.
- García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**(2), 89–109.
- Greselin, F. and Ingrassia, S. (2010). Constrained monotone EM algorithms for mixtures of multivariate t distributions. *Statistics and Computing*, **20**(1), 9–22.
- Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, **32**(4), 1313–1340.
- McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.
- McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t -distributions. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 658–666. Springer, Berlin - Heidelberg.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.
- Orbanz, P. and Buhmann, J. M. (2005). SAR images as mixtures of Gaussian mixtures. *IEEE International Conference on Image Processing*, **2**, 209–212.
- Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Silverman, B. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B: Methodological*, **43**, 97–99.
- Titterton, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

New perspectives for the *RDI* index in social research fields

Emanuela Raffinetti and Pier Alda Ferrari

Abstract The great interest in quantitative social research has led to the development of specific statistical techniques suitable in dealing with dependence between variables also in connection with ordinal data. A new index, hereafter called “Rank-based Dependence Index” (*RDI*), was recently provided as a monotonic dependence measure. Due to its properties and specific features, *RDI* appears as an overcoming of the Pearson’s correlation coefficient, since it captures not only linear dependence relationships, and is specifically useful when the independent variable is ordinal. The *RDI* adequacy is validated by simulation studies assessing its performance with respect to the ones of other main competitors, such as the Pearson’s and Spearman’s correlation coefficients. An application in case of ordinal data is also illustrated.

Key words: dependence relationship, ordinal data, Monte Carlo simulations

1 Background

Typically, when dealing with social issues the need of introducing and discussing descriptive indicators arises in order to easily identify features of a society which can be measured, vary over time and taken as revealing some underlying aspect of social reality. In such context, the most commonly used indicators are derived from official statistics and include unemployment figures, health, gender and mortality data and crime rates.

The purpose of this paper is presenting a new monotonic dependence index. This index is based on a previous proposal. In fact, it was firstly employed as an index of “equity in a taxation process” (see e.g. Muliere, 1986) and subsequently used as a novel measure of goodness of fit for a multiple linear regression model (see e.g. Raffinetti and Giudici, 2012). The new version of the index was studied and developed by Ferrari and Raffinetti (2012), in order to provide both an extension of such measure to any real-valued variable and to attribute a new interpretation in terms of monotonic dependence relationships. For this reason, hereafter the index will be called “Rank-based Dependence Index” and denoted with the acronym *RDI*.

Some recalls to *RDI* characteristics are needed, such as, for instance its expression and its properties. Let Y and X be two variables (Y numerical, X numerical/ordinal) and let us consider a simple linear regression model between a

Emanuela Raffinetti · Pier Alda Ferrari

Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano (Italy), e-mail: emanuela.raffinetti@unimi.it, pier-alda.ferrari@unimi.it

response variable Y and a covariate X . Let us denote with $y_{(i)}$ the ordered (in non-decreasing sense) Y values and with y_i^* the same Y values re-ordered according to the ranks of the Y values estimated by regression model, obtaining n pairs $(y_{(i)}, y_i^*)$ ($\forall i = 1, \dots, n$). A general *RDI* formula is given by:

$$RDI = \frac{2 \sum_{i=1}^n i(y_i^* - y_0) - n(n+1)(M_Y - y_0)}{2 \sum_{i=1}^n i(y_{(i)} - y_0) - n(n+1)(M_Y - y_0)}, \quad i = 1, \dots, n \quad (1)$$

where $y_0 = \min(0, y^-)$, with y^- representing the minimum response variable Y if a negative value, and M_Y is the Y mean value. The proposed measure ranges between -1 and $+1$, is equal to zero in case of independence and since is suitable to assess monotonic dependence of Y from X , it seems worth analyzing its properties in comparison to other similar measures. Here, this problem is faced and discussed.

Additional studies are carried on in order to detect similarities and dissimilarities of *RDI* with respect to its main competitors, i.e. Pearson's (ρ) and Spearman's (ρ_S) correlation coefficients. For such purpose, a Monte Carlo simulation study was run and the related findings are shown and discussed in Sect. 2. In particular, *RDI* results also as an overcoming of ρ , since it captures linear dependence relationship as well as any monotonic dependence one. Due to such role, *RDI* performance is specifically useful in case of discretization process involving the independent variable, as highlighted in Sect. 3, where a real application of our proposal is illustrated and the performance of the *RDI* compared with ρ is considered.

2 A comparison of *RDI*, ρ and ρ_S through Monte Carlo simulations

For a more deep analysis of *RDI* behavior, a comparison between its performance and that of its main competitors is provided by running Monte Carlo simulations based on generating samples from the family of MEP distributions, being this family one of the possible generalizations of normal distributions in terms of ellipsoidal departures. For more details about MEP distributions, see e.g. Solaro (2004): however, what is basic to point out is that MEP distributions depend on a specific parameter, denoted with κ and expressing the “non-normality” condition. For $\kappa < 2$ and $\kappa > 2$, respectively leptokurtic and platikurtic distributions are obtained, while for $\kappa = 2$ a normal distribution is derived. In such paper the simulation study is carried on by choosing values $\kappa = \{1, 2, 8\}$ for respectively describing leptokurtic, normal and platikurtic bivariate distributions. Through the illustrated procedure, the sampling distribution of *RDI*, ρ and ρ_S for variables generated by MEP distributions are obtained under different experimental conditions. For this purpose, a pair of continuous variables, whose marginal distributions are defined according to the following pairwise correlation coefficients $\rho = \{0.2, 0.4, 0.6, 0.8\}$, were generated defining four different scenarios. Under each of such scenarios, we drew samples of size 100, 500 and 1,000 and we iterated these steps 10,000 times. For the sake of shortness, here we report only results corresponding to the scenario with a pairwise correlation coefficient equal to 0.6, stressing that similar findings can be reached also with

respect to the other pairwise correlation levels. Simulation results are shown by the boxplots in Fig. 1, 2 and 3: above each boxplot the median value is specified.

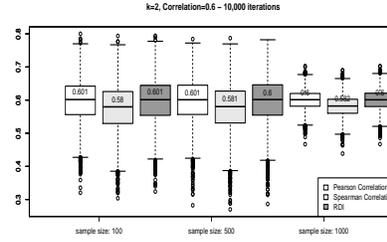
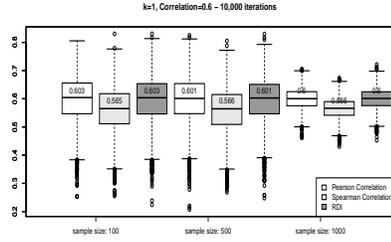


Fig. 1 Pairwise correlation $\rho = 0.6$ and $\kappa = 1$ **Fig. 2** Pairwise correlation $\rho = 0.6$ and $\kappa = 2$

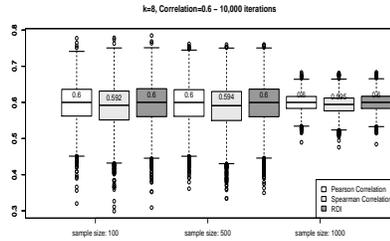


Fig. 3 Pairwise correlation $\rho = 0.6$ and $\kappa = 8$

In case of MEP distributions with $\kappa = 2$, the dependence and correlation coincide. In fact, as shown in Fig. 2 the boxplots representation at 0.6 pairwise correlation coefficient, highlights this expected result but also the better performance of *RDI* with respect to ρ_S , which in turn is built on ranks. In particular, even if the corresponding boxplots are not reported here, as gradually the pairwise correlation coefficient increases, the distance between the *RDI* and ρ_S grows stressing that with variables characterized by strong linear dependence linkages, the *RDI* is a more suitable measure than ρ_S . Similar conclusions arise in case of MEP distributions with $\kappa \neq 2$. More precisely, even if with platikurtic MEP distributions ρ_S sampling distribution approaches to that of *RDI* and ρ when the pairwise correlation coefficient is quite small (i.e. 0.2 or 0.4), ρ_S always provides an underestimation of ρ . In conclusion the *RDI* performance is similar compared to that of ρ . Better performance of this index with respect to ρ is verified when other bivariate variables, for which dependence and correlation do not coincide, are involved. For the sake of brevity, the results are not here reported.

3 Further investigations in application contexts

In this section, we introduce an example of possible application of the proposed dependence index *RDI* with the purpose to show how our method can lead to efficient results in terms of monotonic dependence relationship especially in case of ordinal

variables. In order to validate in practice our index, let us consider an SPSS Data file “Employee Data.sav” located in folder C:\Program Files\SPSS. Some brief details about the considered variables are needed. Since our proposal is suitable when the response variable is continuous and the independent one is almost of ordinal nature, the two analyzed variables are represented by the individual current salary (Y) and the individual beginning salary (X), both by considering the observed X values and discretizing them according to five ordered categories obtained by sharing the beginning salary into intervals of equal width. The corresponding frequency distribution is illustrated in Table 1.

Groups	1	2	3	4	5
Absolute Frequencies	112	95	109	72	85

Table 1 Frequency distribution of the five selected ordered categories

In particular, we aim at studying the effect of the discretization process on RDI and ρ .

Independent continuous variable: Beginning Salary in dollars	
RDI	0.8902
ρ	0.8802
Independent discretized variable: Beginning Salary grouped in five ordered categories	
RDI	0.8465
ρ	0.7375

Table 2 RDI and ρ in case of continuous and discretized independent variable

According to results shown in Table 2, the discretization process translates into a shrinkage of ρ (0.7375). Such condition does not occur for RDI which achieves a value (0.8465) very close to that of the original correlation coefficient (0.8802) computed with respect to the beginning salary variable expressed in dollars.

Acknowledgements The authors wish to acknowledge financial support from the European Social Fund Grant (Lombardy Region, Italy).

References

1. Ferrari, P.A., Raffinetti, E.: An extension and a new interpretation of the Rank-based Concordance Index. In: Analysis and modeling of complex data in behavioural and social sciences. Cleup, Padova (2012)
2. Muliere, P.: Some notes about the horizontal equity of a taxation (in Italian), in honor of Francesco Brambilla. Bocconi Communication (Ed.), Vol. 2, Milan (1986)
3. Raffinetti, E., Giudici, P.: Multivariate Ranks-based concordance indexes. In: Advanced Statistical Methods for the analysis of large data-sets, pp. 465-473, Springer (2012)
4. Solaro, N.: Random variate generation from Multivariate Exponential Power distribution. Statistica & Applicazioni, II(2) (2004)

Mixture models for ordinal data: a pairwise likelihood approach

Monia Ranalli and Roberto Rocci

Abstract We propose a latent Gaussian mixture model to classify ordinal data. The observed data are considered as a discretization of an underlying latent mixture. A pairwise likelihood approach is used to evaluate a multidimensional integral that cannot be written in a closed form. The model is estimated within the expectation-maximization framework.

Key words: Mixture models, Ordinal data, Pairwise likelihood, EM algorithm

1 Introduction

In behavioural, social and health sciences, the variables aimed at measuring attitudes, abilities or opinions are typically categorical, frequently of ordinal type. Modelling such variables is quite challenging, due to the lack of metric properties; a possible way is to consider the observed responses as a categorical manifestation of continuous latent variables. Two main approaches exist in this context: the Underlying Response Variable (URV), (see e.g. [8, 9, 10]), and the Item Response Theory (IRT), (see e.g. [1]). However, if the data arise from an heterogeneous population, the most common model based-clustering method for categorical data is the latent class analysis; but it does not account for the dependence within the groups because of the local independence assumption. Some earlier attempts have aimed to assume a latent Gaussian mixture underlying the categorical variables. Everitt presents a mixture model for mixed data with some constraints on the mixture parameters,

Monia Ranalli
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: monia.ranalli@uniroma1.it

Roberto Rocci
Dipartimento di Economia e Finanza, Università di Roma Tor Vergata, Roma
e-mail: roberto.rocci@uniroma2.it

[5]. Nevertheless, the estimation of the model by maximum likelihood requires the numerical computation of a multidimensional integral and thus it can include only one or at most two categorical variables [6]. More recently, latent variable models to cluster mixed [2, 3] and categorical [7] data under the conditional independence assumption have been proposed. Here, following the UVR approach we assume that the observed data are generated by thresholding an underlying latent Gaussian mixture. The model differs from the existing ones in relaxing the assumption of the local independence and including an arbitrary number of ordinal variables: it allows us to capture both the unobserved heterogeneity (cluster structure) and the dependence within the groups. Moreover, we impose some identification constraints letting us to estimate both the means and the covariance matrices of the latent mixture. Since parameter estimates through the full information maximum likelihood becomes computationally demanding and thus infeasible as the number of variables increases (typically, when it is greater than 5), we adopt a pairwise maximum likelihood approach belonging to the composite maximum likelihood methods ([11, 12] give an extensive overview). Based on the existing results in literature, the pairwise likelihood estimation seems to be a good balance between statistical and computational efficiency and its estimators have the desired properties of being asymptotically unbiased, consistent and normally distributed, under regularity assumptions.

2 The standard underlying response variable approach

The URV approach treats the ordinal variables as metric by assuming underlying continuous variables, which are partially observed through their ordinal counterparts. Let x_1, x_2, \dots, x_P be ordinal variables and $c_i = 1, 2, \dots, C_i$ the associated categories for $i = 1, 2, \dots, P$. There are $\prod_{i=1}^P C_i$ possible response patterns, which have the following form $\mathbf{x}_r = (x_1 = c_1, x_2 = c_2, \dots, x_P = c_P)$. The probability $\pi_r(\boldsymbol{\theta})$ of \mathbf{x}_r is assumed to be a function of a parameter vector $\boldsymbol{\theta}$, subject to $\pi_r(\boldsymbol{\theta}) > 0$ and $\sum_r \pi_r(\boldsymbol{\theta}) = 1$. For a random i.i.d. sample of size N the log-likelihood is given by

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \sum_{r=1}^R n_r \log \pi_r(\boldsymbol{\theta}), \quad (1)$$

where n_r is the observed sample frequency of response pattern r and $\sum_r n_r = N$. Under URV, there is $\mathbf{y} \sim N_P(\mathbf{0}, \mathbf{R})$ underlying \mathbf{x} . The latent relationship between \mathbf{x} and \mathbf{y} is defined by the following threshold model,

$$x_i = c_i \Leftrightarrow \gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)}. \quad (2)$$

Hence, the probability of a response pattern r is given by

$$\pi_r(\boldsymbol{\theta}) = Pr(x_1 = c_1, x_2 = c_2, \dots, x_P = c_P; \boldsymbol{\theta}) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_P-1}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi_P(\mathbf{y}; \mathbf{0}, \mathbf{R}) d\mathbf{y}. \quad (3)$$

3 An extended underlying response variable approach to mixture

Now, we extend the model presented above assuming that the data arise from an heterogeneous population. The aim is to cluster individuals into their unobservable groups. To accommodate both cluster structure and dependence within the groups, let \mathbf{y} be a latent variable distributed as a Gaussian finite mixture underlying the observed data. The ordinal variables are considered as a discretization of the heteroscedastic latent Gaussian mixture \mathbf{y} and the main assumption is that the thresholds γ are the same over the groups. For a random i.i.d. sample of size N the log-likelihood is

$$\ell(\Psi; \mathbf{x}) = \sum_{r=1}^R n_r \log \left[\sum_{g=1}^G p_g \pi_r(\theta_g, \gamma) \right], \quad (4)$$

where $\Psi = \{p_1, \dots, p_{G-1}, \theta_1, \dots, \theta_G, \gamma\}$, p_g is the probability of belonging to group g subject to $p_g > 0$ and $\sum_{g=1}^G p_g = 1$, while $\pi_r(\theta_g, \gamma)$ is the probability of the response pattern r in the cluster g ,

$$\pi_r(\theta_g, \gamma) = Pr(x_1 = c_1, x_2 = c_2, \dots, x_P = c_P; \theta_g, \gamma) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_{P-1}}^{(P)}}^{\gamma_{c_P}^{(P)}} \phi_P(\mathbf{y}; \mu_g, \Sigma_g) d\mathbf{y}.$$

4 Estimation and Identifiability

The maximization of the log-likelihood defined in (4) over the parameter vector Ψ requires the evaluation of a P -dimensional integral, which cannot be written in a closed form. Furthermore its numerical computation is infeasible with $P > 5$. Thus, we adopt a pairwise maximum likelihood approach. The pairwise log-likelihood is

$$p\ell(\Psi; \mathbf{x}) = \sum_{i=1}^{P-1} \sum_{j=i+1}^P \ell(\Psi; (x_i, x_j)) = \sum_{i=1}^{P-1} \sum_{j=1}^P \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} n_{c_i c_j}^{(ij)} \log \left[\sum_{g=1}^G p_g \pi_{c_i c_j}^{(ij)}(\theta_g, \gamma) \right], \quad (5)$$

where $n_{c_i c_j}^{(ij)}$ is the observed frequency of a response in category c_i and c_j for variables x_i and x_j respectively, while $\pi_{c_i c_j}^{(ij)}(\theta_g, \gamma)$ is the corresponding probability under the model obtained by integrating the density of a bivariate normal distribution with parameters (μ_g, Σ_g) between the corresponding threshold parameters. Parameter estimation is carried out using an expectation-maximization (EM) algorithm [4]. According to (5), the group memberships are considered as missing. Let \mathbf{z} denote the group membership matrix $\left(\left(\sum_{i=1}^{P-1} \sum_{j=1}^P c_i \times c_j \right) \times G \right)$, where $z_{c_i c_j; g}^{(ij)} = 1$ if the cell (c_i, c_j) belongs to component g and $z_{c_i c_j; g}^{(ij)} = 0$ otherwise, for $g = 1, \dots, G$. In the E-step the latent variables \mathbf{y} are not considered missing. In the M-step we maximize the complete pairwise log-likelihood function; as the M-step has not a closed form, its maximization is implemented in Matlab by using the command "fmincon". The

complete pairwise log-likelihood function is

$$p\ell_c(\Psi; \mathbf{x}) = \sum_{i=1}^{P-1} \sum_{j=1}^P \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} \sum_{g=1}^G n_{c_i c_j z_{c_i c_j; g}}^{(ij)} \left[\log \left(\pi_{c_i c_j}^{(ij)}(\boldsymbol{\theta}_g, \boldsymbol{\gamma}) \right) + \log(p_g) \right]. \quad (6)$$

It is clear that the pairwise approach is feasible as it requires the evaluation of bivariate normal distributions, regardless of the number of observed or latent variables. Furthermore the estimation of all parameters is carried out simultaneously.

The model proposed is identified under some constraints. In general, given a $C_1 \times C_2 \times \dots \times C_p$ contingency table, there are $C = \prod_i^p c_i - 1$ essential parameters, since they have to sum to 1; it follows that if the number of model parameters is greater than C , then the model is not identified. This is a necessary condition for the identifiability of a latent class model [1], but here it is not sufficient. In order to estimate both the means and the covariance matrices of the latent variables in each group, the thresholds have not to change across the groups. Furthermore, a group is fixed as a reference group; thus, its mean vector is set to 0, its covariance matrix is a correlation matrix.

References

1. Bartholomew, D., Knott, M., Moustaki, I.: *Latent Variable Models and Factor Analysis: A Unified Approach*, third edn. Wiley Series in Probability and Statistics. Wiley (2011)
2. Browne, R.P., McNicholas, P.D.: Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference* **142**(11), 2976–2984 (2012)
3. Cai, J.H., Song, X.Y., Lam, K.H., Ip, E.H.S.: A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics & Data Analysis* **55**(11), 2889–2907 (2011)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1) (1977)
5. Everitt, B.: A finite mixture model for the clustering of mixed-mode data. *Statistics Probability Letters* **6**(5), 305–309 (1988)
6. Everitt, B., Merette, C.: The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics* **17**(3), 283–297 (1990)
7. Gollini, I., T.B., M.: *Mixture of latent trait analyzers for model-based clustering of categorical data*. Tech. rep., University College Dublin (2012)
8. Jöreskog, K.G.: New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity* **24**(4), 387–404 (1990)
9. Lee, S.Y., Poon, W.Y., Bentler, P.: Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics Probability Letters* **9**(1), 91–97 (1990)
10. Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**(1), 115–132 (1984)
11. Varin, C.: On composite marginal likelihoods. *AStA Advances in Statistical Analysis* **92**(1), 1–28 (2008)
12. Varin, C., Firth, D.: An overview of composite likelihood methods. *Statistica Sinica* **21**(1), 1–41 (2011)

Issues in robust clustering

Marco Riani, Andrea Cerioli and Gianluca Morelli

Abstract It is now widely recognized that the presence of outliers can affect the results of any statistical analysis. This is also the case of cluster analysis methods. Recently, special attention in the robust clustering literature has been devoted to classification methods based on trimming which try to discard most outlying observations when carrying out the clustering process. The idea of trimming, together with the need of considering groups of different sizes and orientation, has led to the suggestion of maximization of very complicated functions with many parameters and a very high computational complexity due to the “combinatorial” nature of the problem and constraints in order to avoid spurious solutions. In this paper we give a general overview about the computational/theoretical problems that recent robust cluster analysis methods necessarily imply and we concentrate on the graphical tools which have been proposed in order to select the optimal trimming proportion and the optimal number of groups.

Key words: Robust clustering, TCLUS, Trimmed likelihood curves

1 Description of the study

Several “mixture modeling” and “crisp clustering” approaches to model-based Clustering can be found in the literature. Mixture modeling approaches assume that

Marco Riani
Department of Economics, Univ. of Parma, Via Kennedy 6, 43125 Parma, e-mail: mriani@unipr.it

Andrea Cerioli
Department of Economics, Univ. of Parma, Via Kennedy 6, 43125 Parma, e-mail: andrea.cerioli@unipr.it

Gianluca Morelli
Department of Economics, Univ. of Parma, Via Kennedy 6, 43125 Parma, e-mail: gianluca.morelli@unipr.it

data at hand y_1, \dots, y_n in R^p come from a probability distribution with density $\sum_{j=1}^k \pi_j \phi(\cdot, \theta_j)$ with $\phi(\cdot, \theta_j)$ being the p -variate (generally multivariate normal) densities with parameters θ_j , $j = 1, \dots, k$. Generally $\theta_j = (\mu_j, \Sigma_j)$ where μ_j is the population mean and Σ_j is the covariance matrix for component j . This leads to likelihoods of the form

$$\prod_{i=1}^n \sum_{j=1}^k \pi_j \phi(y_i; \theta_j). \quad (1)$$

On the other hand, ‘‘crisp’’ (0-1) clustering approaches assume classification likelihoods of the following form

$$\prod_{j=1}^k \prod_{i \in R_j} \phi(y_i; \theta_j), \quad (2)$$

where R_j contains the indexes of the observations which are assigned to group j , with the constraint that $\#\bigcup_{j=1}^k R_j = n$.

In order to discard a fraction of most outlying observations (say equal to α) and to take into account the different sizes of the groups when making the final group assignments, Garcia-Escudero et al. (2008) suggested to maximize the following expression (TCLUST):

$$\prod_{j=1}^k \prod_{i \in R_j} \pi'_j \phi(y_i; \theta_j) \quad (3)$$

with the constraint that $\#\bigcup_{j=1}^k R_j = \lfloor n(1 - \alpha) \rfloor$ where symbol $\lfloor \cdot \rfloor$ denotes the integer part. Note that in equation (3) we have used symbol π'_j to stress that these parameters have a completely different interpretation from the π_j in equation (1). They are intended to take into account the different sizes of the groups when making the final group assignments and they are not the weights of the mixture likelihood. TCLUST method also considers scatter constraints in terms of the group covariance matrices. More specifically, if $\lambda_l(\hat{\Sigma}_j)$ ($l = 1, \dots, p$; $j = 1, \dots, k$) are the estimated eigenvalues of the group covariance matrix $\hat{\Sigma}_j$, TCLUST in each iteration of the maximization routine imposes the constraint:

$$\frac{\max_{l=1, \dots, p} \max_{j=1, \dots, k} \lambda_l(\hat{\Sigma}_j)}{\min_{l=1, \dots, p} \min_{j=1, \dots, k} \lambda_l(\hat{\Sigma}_j)} \leq c. \quad (4)$$

Note that classic k -means procedure is simply obtained putting $\alpha = 0$ and $\pi'_j = 1$ in equation (3) and $c = 1$ in equation (4). The idea of trimming under the eigenvalue constraint ratio of equation (4) can also be applied in the context of the mixture likelihood given in equation (1) with important consequences. In the crisp assignment in each iteration of the maximization process, the selection of the $\lfloor n(1 - \alpha) \rfloor$ units is made taking the $\lfloor n(1 - \alpha) \rfloor$ largest values of ϕ_i^* , where $\phi_i^* = \max_{j=1, \dots, k} \hat{\pi}'_j \phi(y_i; \hat{\theta}_j)$ where $\hat{\pi}'_j$ is estimated using proportion of untrimmed observations which are assigned to each group. Estimates of centers and the covariance matrices use respectively the unweighted sample mean, sample covariance matrices. On the other hand,

in the context of mixture modelling, the quantities

$$\phi_{ij}^* = \frac{\hat{\pi}_j \phi(y_i; \hat{\theta}_j)}{\sum_{j=1}^k \hat{\pi}_j \phi(y_i; \hat{\theta}_j)} \quad i = 1, 2, \dots, n, \quad j = 1, \dots, k$$

are interpreted as posterior probabilities. The criterion for selecting the units to trim remains the same, however, centers and the covariance matrices are updated with the weighted sample mean and weighted sample covariance matrices with the weights given by the posterior probabilities. The posterior probabilities for the α trimmed units are set to 0 for each group. Similarly, the $\hat{\pi}_j$ are updated using $\sum_{i=1}^n \phi_{ij}^* / [n(1 - \alpha)]$.

A feasible algorithm aimed at approximately solving the objective function was given in Garcia-Escudero et al. (2008) and Fritz et al. (2013). These algorithms belong to the family of Classification EM algorithms and, to perform the data-driven trimming, make use of the so called ‘‘concentration’’ steps (as those behind the Fast-MCD algorithm in Rousseeuw and van Driessen 2006). Leaving aside the intricacies and the computational issues of the concentration steps, and the possibility of finding a local minimum together with the choice of the number of initial starting points, the specific aspects of TCLUST concern the optimal choice of α , of c and k , the implications of assuming equal or unequal weights, or the computation of the centroids and covariances with weighted or unweighted arithmetic means. The goodness of TCLUST, together with its theoretical foundations, has been shown in a series of papers by the Valladolid group (see for example García-Escudero et al. (2008) and (2010)). However, a deep simulation study to investigate the implications of the different a priori choices on the parameters mentioned above, has not been done yet.

In order to find optimal values of k and α , García-Escudero et al. (2010) and (2011) have suggested to monitor the trimmed log likelihoods given in equation (2) as a function of α , for different values of k and a prefixed value of c . The idea is to choose the smallest value k such that no great changes in the classification trimmed likelihood curves are found when increasing from k to $k + 1$. In order to provide further insight about this technique we show here some preliminary results of a broad simulation study which has been conducted in order to investigate the implications and feasibility of the use of this criterion.

The two panels of Figure 1 show two representative datasets which have been taken from different simulation scenarios. In both cases the data generating process is made up of three groups. However, while in the first simulation study the noise is concentrated in a particular area, in the second the noise is much more spread. Figure 2 shows the median values of the classification likelihood curves over 100 simulations. It is clear that while in the first simulation study this plot provides us with a guidance about the proper number of groups and the optimal percentage of trimming (in this case $\alpha = 0.05$ and $k = 3$), in the second case the plot is uninformative about the choice of the trimming level. This plot therefore shows that in such cases it is necessary to combine TCLUST with additional robust adaptive proce-

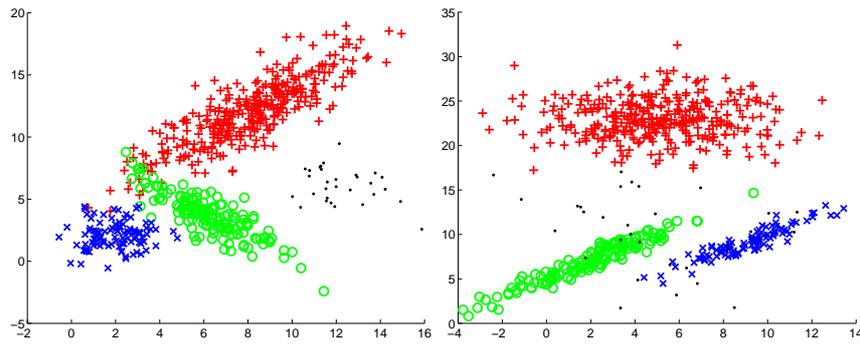


Fig. 1 Typical dataset of first (left panel) and second (right panel) simulation study

dures like the forward search (Riani et al. 2009) to have an idea about the number of observations which have to be assigned to the different groups.

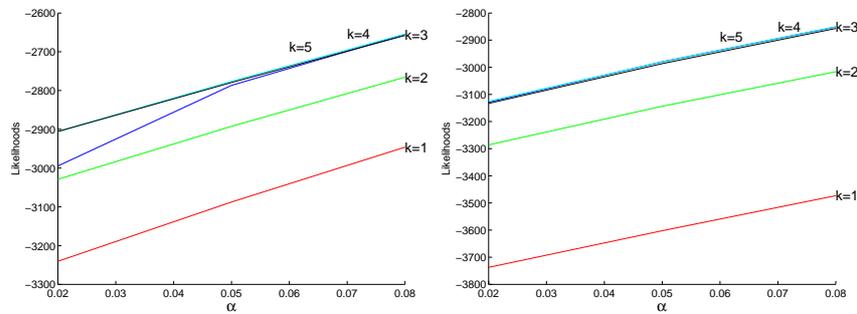


Fig. 2 Median values of trimmed likelihood curves for first (left panel) and second (right panel) simulation study

References

1. Fritz H., García-Escudero L.A., and Mayo-Iscar A.: A fast algorithm for robust constrained clustering. *Comput. Stat. Data Anal.* **61**, pp. 124-136 (2013)
2. García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A.: A general trimming approach to robust cluster analysis, *Ann. Statist.*, **36**, 1324-1345, (2008)
3. García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A.: A review of robust clustering methods, *Advances in Data Analysis and Classification* **4**, pp. 89-109, (2010)
4. García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A.: Exploring the number of groups in robust model-based clustering, *Statistics and Computing*, **21**, pp. 585-599, (2011)
5. Riani, M., Atkinson, A. C., and Cerioli, A.: Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, pp. 447-466, (2009)
6. Rousseeuw, P. J. and K. Van Driessen: Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery* **12**, pp. 2945, (2006)

Deciphering and modeling heterogeneity in interaction networks

Stéphane Robin

Abstract Network analysis has become a very active field of statistics within the last decade. Several models have been proposed to describe and understand the heterogeneity observed in real networks. We will present here several modeling involving latent variable. We will discuss the issues raised by their inference, focusing on the stochastic block model. We will describe a variational approach that allows to deal with the complex dependency structure of this model.

Key words: network, random graph, mixture model, variational inference

1 Introduction

From more than a decade, network analysis has arisen in many fields of application such as biology, sociology, ecology, industry, internet, *etc.* Network is a natural way to describe how individuals or entities interact. An interaction network consists in a graph where each nodes represents an individual and an edge exists between two nodes if the two corresponding individuals interact in some way. Interaction may refer to social relationships, molecular bindings, wired connexion or web hyperlinks, depending on the context. Such interactions can be symmetric or asymmetric, binary (when only the presence or absence of an edge is recorded) or weighted (when a value is associated with each observed edge).

Network analysis raises a series of interesting statistical questions because of the atypical organization of graph-structured data. We focus here on the analysis of the global topology of the graph. It has been observed that real network display various structural characteristics such as hubs (nodes connected to a large number of other nodes), highly imbalanced degree ('scale free') distributions, communities (sets of nodes highly connected between them but with only few connections with outer

S. Robin
UMR518, AgroParisTech / INRA, Paris, France, e-mail: robin@agroparistech.fr

nodes), small diameter ('small world': every node can be reached from any other in a small number of steps with respect to the graph size). All such characteristics result from an heterogeneous behavior of the nodes, that we would like to capture to better understand the network's organization.

A large number of methods have been proposed to analyze the topology of a given graph, that mostly belong to two categories. Algorithmic approaches do not make any assumption about the way the network has been build but propose efficient computational strategies to split it into sub-network or to isolate node with critical topological properties ([17]). Model-based methods rely of some probabilistic model ([20]) that provides easy-to-interpret results but often raise inference issues. We focus here on the latter approach, with a special attention to state space models. We limit ourselves to undirected binary network, although many of these methods can be generalized.

2 State space models for networks

[2] proposed a general framework for heterogeneous random graph model. Their model is a state space model and is defined as follows. Consider n nodes ($i = 1, \dots, n$) and denote the presence of an edge between nodes i and j as $X_{ij} := \mathbb{I}\{i \sim j\}$. A latent (unobserved) variable Z_i is associated with each node, the Z_i 's being iid with distribution π over some space \mathcal{Z} : $(Z_i)_i \text{ iid } \sim \pi$. The edges X_{ij} are drawn independently conditional on the Z_i 's, with Benoulli distribution: $X_{ij}|Z_i, Z_j \sim \mathcal{B}(\gamma(Z_i, Z_j))$ where γ is some mapping of $\mathcal{Z} \times \mathcal{Z}$ onto $[0, 1]$. A series of models that have been proposed in the literature can be casted into this framework; we briefly remind some of them.

Latent space model: [11] define a model where the Z_i 'ss have a d -dimensional normal distribution $\pi = \mathcal{N}_d(\mathbf{0}, \mathbf{R})$ and where the connections are governed by the distances in the latent space: $\text{logit}(\gamma(z, z')) = a + \|z - z'\|$.

Clustering in the latent space: [10] propose an extension of the previous model, accounting for clustering. Keeping a similar function γ , the clustering is modeled via a d -dimensional Gaussian mixture in the latent space: $\pi = \sum_k p_k \mathcal{N}_d(\mathbf{m}_k, \mathbf{R})$.

Stochastic block model (SBM): [18] propose a mixture model where the Z_i 's can only take a finite number of values: $\mathcal{Z} = (1, \dots, K)$ so π is simply multinomial.

W-graph: [14] assume that the latent variable are uniformly distributed over $\mathcal{Z} = [0, 1]$ so $\pi = \mathcal{U}[0, 1]$. The function W defined in the quoted article corresponds to the function γ of this abstract.

3 Variational inference for the stochastic block model

We now focus on the SBM. The aim of the inference is both to estimate the parameter $\theta = (\pi, \gamma)$ and to infer the latent variables Z_i 's.

Regular EM. As SBM is a genuine mixture model, maximum likelihood inference can be attempted via the EM algorithm ([7]). We remind that the E step requires to compute the conditional distribution $p(Z|X)$ of the latent variables $Z = (Z_i)$ given the observed ones $X = (X_{ij})$, or at least some of its moments.

Dependency structure. The E step turns out to be infeasible because of the complexity of this conditional distribution. Intuitively, the issue is due to the moralization phenomenon that is observed in certain graphical models ([13]). The conditional distribution of the latent variables under SBM displays a highly intricate dependency structure that make classical inference infeasible even for medium-size graphs.

Variational EM. Variational approximations are often used for the inference of complex graphical models ([23]). The idea is to replace the calculation of the conditional distribution, $p_\theta(Z|X)$ in the E step by an approximation step defined as

$$q_\theta^*(Z) = \arg \min_{q \in \mathcal{Q}} KL(q_\theta(Z) || p_\theta(Z|X))$$

where \mathcal{Q} is a class of easy-to-handle distributions, e.g. $\mathcal{Q} = \{q : q(Z) = \prod_i q_i(Z_i)\}$. It can be shown that the resulting variational EM (VEM) algorithm aims at maximizing a lower bound of the log-likelihood of the data $\log p_\theta(X)$. Such a strategy can be applied to SBM ([6]) and results in a so-called mean field approximation ([19]).

Validity of the variational approximation. Not much is known in general about the validity of variational approximations except some rather negative results ([9]). However, due to specific asymptotic framework of graphs (the number of edges grows as n^2), the variational approximation turns out to be valid as shown in [3].

4 Alternatives and Extensions

Alternatives to variational EM have been proposed for SBM inference.

Variational Bayes EM (VBEM). A Bayesian counterpart of VEM can be derived in order to get an approximate conditional distribution of both the parameter and the latent variables: $p(\theta, Z|X)$. The problem can be stated as

$$q^*(Z, \theta) = \arg \min_{q \in \mathcal{Q}} KL(q^*(Z, \theta) || p(\theta, Z|X)).$$

If the distributions belong to the exponential family and if conjugate priors are used for θ , explicit update formulas for VBEM can be derived ([1]). [12] applied this strategy to SBM and [8] showed its validity even on medium-size graphs, based on an intensive simulation study.

Spectral clustering. Spectral clustering is an efficient algorithm for graph clustering based on the spectral decomposition of the graph Laplacian ([15]). Although it has not been conceived as a model-based method, [22] show that, combined with a K -means step, it results in consistent estimates for SBM.

Degree distribution. As said before, random graph models possess a specific property, as each new node provides information on all other nodes. Under SBM, the degree of each node conditional on its latent variable has a binomial distribution. In the case of SBM, this specific asymptotic framework results in a fast concentration of the degrees around their means, so that the clustering of the nodes can be achieved only based on the degrees, in a linear time, with consistency guaranties ([4]).

Extension of SBM

Weighted graphs and covariates. The SBM model can be generalized to some valued graph by simply changing the emission distribution $\mathcal{B}(\gamma(z, z'))$ into any parametric distribution such as normal or Poisson. In the framework of generalized linear models, covariates can be accounted for via a regression models, as studied in [16].

Connexion with W -graphs. It can be easily seen that SBM corresponds to a W -graph where the graphon function γ is block-wise constant. SBM can therefore be viewed as a piece-wise constant approximation of W -graph and variational inference can be used to infer the graphon function. As noticed in [5], the W -graph suffers strong identifiability issues, but some network characteristics such as the expected number of occurrence of given subgraphs (also called motifs) are invariant. Such moments can be computed in the framework of SBM ([21]).

References

1. Beal M., J., Ghahramani, Z.: The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayes. Statist.* **7**, 543–52 (2003)
2. Bollobás, B., Janson, S., Riordan, O.: The phase transition in inhomogeneous random graphs. *Rand. Struct. Algo.* **31**(1), 3–122 (2007)
3. Celisse, A., Daudin, J.J., Pierre, L.: Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.* **6**, 1847–99 (2012)
4. Channarond, A., Daudin, J.J., Robin, S.: Classification and estimation in the stochastic block model based on the empirical degrees. *Electron. J. Statist.* **6**, 2574–601 (2012)
5. Chatterjee, S., Diaconis, P.: Estimating and Understanding Exponential Random Graph Models. ArXiv e-prints (2011)
6. Daudin, J.J., Picard, F., Robin, S.: A mixture model for random graphs. *Stat. Comput.* **18**(2), 173–83 (2008)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38 (1977)
8. Gazal, S., Daudin, J.J., Robin, S.: Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation* **82**(6), 849–862 (2012). DOI 10.1080/00949655.2011.560117. URL <http://www.tandfonline.com/doi/abs/10.1080/00949655.2011.560117>

9. Gunawardana, A., Byrne, W.: Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res.* **6**, 2049–73 (2005)
10. Handcock, M., Raftery, A., Tantrum, J.: Model-based clustering for social networks. *JRSSA* **170**(2), 301–54 (2007). doi: 10.1111/j.1467-985X.2007.00471.x
11. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97**(460), 1090–98 (2002)
12. Latouche, P., Birmelé, E., Ambroise, C.: Variational bayesian inference and complexity control for stochastic block models. *Statis. Model.* **12**(1), 93–115 (2012)
13. Lauritzen, S.: *Graphical Models*. Oxford Statistical Science Series. Clarendon Press (1996)
14. Lovász, L., Szegedy, B.: Limits of dense graph sequences. *J. Combin. Theory, Series B* **96**(6), 933–57 (2006). DOI DOI: 10.1016/j.jctb.2006.05.002
15. von Luxburg, U., Belkin, M., Bousquet, O.: Consistency of spectral clustering. *Ann. Stat.* **36**(2), 555–586 (2008). DOI 10.1214/009053607000000640
16. Mariadassou, M., Robin, S., Vacher, C.: Uncovering structure in valued graphs: a variational approach. *Ann. Appl. Statist.* **4**(2), 715–42 (2010)
17. Newman, M., Girvan, M.: Finding and evaluating community structure in networks,. *Phys. Rev. E* **69**, 026,113 (2004)
18. Nowicki, K., Snijders, T.: Estimation and prediction for stochastic block-structures. *J. Amer. Statist. Assoc.* **96**, 1077–87 (2001)
19. Parisi, G.: *Statistical Field Theory*. Addison Wesley, New York), (1988)
20. Pattison, P.E., Robins, G.L.: *Handbook of Probability Theory with Applications*, chap. Probabilistic Network Theory. Sage Publication (2007)
21. Picard, F., Daudin, J.J., Koskas, M., Schbath, S., Robin, S.: Assessing the exceptionality of network motifs,. *J. Comp. Biol.* **15**(1), 1–20 (2008)
22. Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic block-model. *Ann. Stat.* **39**(4), 1878–1915 (2011). DOI 10.1214/11-AOS887
23. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1**(1–2), 1–305 (2008). [Http://dx.doi.org/10.1561/22000000001](http://dx.doi.org/10.1561/22000000001)

Partial Possibilistic Regression Path Modeling

Rosaria Romano and Francesco Palumbo

Abstract In the Structural Equation Modeling framework, the proposal aims at introducing the possibilistic regression to model the net of relations linking the latent variables with their respective manifest variables and the latent variables among them. Possibilistic regression defines the relation between one dependent variable and a set of predictors through an interval-valued coefficients linear function. Possibilistic regression considers the error due to the vagueness of the human perception or knowledge of the model, and it reflects such ambiguity in the model interval parameters. Under the *component-based* paradigm, the *partial possibilistic regression path modeling* estimates model parameters through numerical optimization techniques. The optimization procedure aims to minimize the spread of all the interval coefficients in the model, under properly defined constraints that permit to consider the linear relations among the variables.

Key words: interval valued data, possibilistic regression, component-based SEM

1 Introduction

Structural Equation Models (SEM) [1] are reference techniques for measuring cause-effect relationships in complex systems. SEM consist of a network of causal relationships among latent variables (LV) defined by blocks of manifest variables (MV). Under the framework of *component-based* estimation methods [8], with an increasing popularity in several areas, Partial Least Squares Path Modeling (PLS-PM) represents a statistical approach to SEM [7]. PLS-PM formulates the causality

Rosaria Romano
Univeristy of Calabria, Cosenza, e-mail: rosaria.romano@unical.it
Francesco Palumbo
University of Naples Federico II, Napoli, e-mail: fpalumbo@unina.it

dependencies between LV in terms of linear conditional expectation, and estimates the LV through a system of interdependent equations based on simple and multiple regressions.

The present paper aims at introducing the Possibilistic Regression (PR) [6] to model the relations among LV and their related MV. In some previous contributions [4, 3] the use of PR in PLS-PM framework has been already proposed to estimate only some parameters of the SEM. The present work intends to be a further step ahead which extend the PR estimating to the entire model.

PR defines the relation between one dependent variable Y and a set of P predictors X_1, X_p, \dots, X_P through a linear function holding interval valued coefficients:

$$Y = \tilde{\omega}_1 X_1 + \dots + \tilde{\omega}_p X_p + \dots + \tilde{\omega}_P X_P \quad (1)$$

where $\tilde{\omega}_p$ denotes the generic interval coefficient in terms of midpoint and spread: $\tilde{\omega}_p = \{c_p; a_p\}$. There are no restrictive assumptions on the model. Differently from statistical regression, the deviations between data and linear models are assumed to depend on the imprecision of the parameters and not on measurement errors. This means that in PR there is no external error component but the spread of the coefficients embeds all uncertainty, such that PR minimizes the total spread of the interval coefficients:

$$\min_{a_p} \sum_{p=1}^P \left(\sum_{n=1}^N a_p |x_{np}| \right), \forall p = 1, \dots, P \quad (2)$$

under the following linear constraints:

$$\begin{aligned} \sum_{p=1}^P c_p x_{np} + \sum_{p=1}^P a_p |x_{np}| &\geq y_n, \forall n = 1, \dots, N, \\ \sum_{p=1}^P c_p x_{np} - \sum_{p=1}^P a_p |x_{np}| &\leq y_n, \forall n = 1, \dots, N. \end{aligned} \quad (3)$$

satisfying the following conditions: *i*) $a_p \geq 0$, *ii*) $c_p \in R$, *iii*) $x_{n1} = 1$. Constraints in (3) guarantee the inclusion of the whole given data set into the estimated boundaries.

In a geometric view, where statistical units are represented as points in the \mathfrak{R}^{P+1} space, the optimal solution ensures the inclusion of the whole given data set into the estimated boundaries with the minimum spread of parameters.

Kim *et al.* [2] have carried out that fuzzy regression, which is a special case of PR, is more useful when the data set is too small to support statistical regression analysis and/or the aptness of the model is poor due to vague relationships among variables or to a poor model specification. Romano and Palumbo [5] pointed out that fuzzy estimators are unbiased and not affected by *quasi* multi-collinearity.

2 Partial Possibilistic Regression Path Modeling

Let us assume predictors are collected into a partitioned table $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_h, \dots, \mathbf{X}_H]$, where \mathbf{X}_h is the generic block composed by J_h indicators. In the PLS-PM literature, it is used to distinguish the *structural model* (or inner model) linking the LV, and the *measurement model* (or outer model) linking the LV with their respective block of MV. The measurement model can be *reflective* or *formative* according to the linkage between the LV and the MV [7]. However, the present paper only focuses on the *reflective model*.

In PLS-PM, an iterative procedure permits to estimate the latent variable scores (ξ) and the outer weights (\mathbf{w}), while path coefficients (β) come afterward from a regular regressions between the estimated latent variables.

The algorithm computes the latent variables' scores alternating the *outer* and *inner* estimation till convergence. The procedure starts by choosing arbitrary weights w_{jh} . In the external estimation, the latent variable is estimated as a linear combination of its own MV:

$$\mathbf{v}_h \propto \sum_{j=1}^{J_h} w_{jh} \mathbf{x}_{jh} = \mathbf{x}_h \mathbf{w}_h \quad (4)$$

where \mathbf{v}_h is the standardized outer estimation of the latent variable ξ_h and the symbol \propto means that the left side of the equation corresponds to the standardized right side. In the internal estimation, the latent variable is estimated by considering its links with the other adjacent latent variables:

$$\mathbf{z}_h \propto \sum_{h' \neq h} e_{hh'} \mathbf{v}_{h'} \quad (5)$$

where the inner weights are defined according to different schemes [7] and the notation hh' indicates the adjacency condition of h to h' .

These first two steps allow us to update the outer weights w_{jh} . In the *reflective model* the weight is the regression coefficient of \mathbf{z}_h in the simple regression of \mathbf{x}_{jh} on the inner estimate \mathbf{z}_h , which corresponds to the covariance as \mathbf{z}_h is standardized:

$$w_{hj} = \text{cov}(\mathbf{x}_{jh}, \mathbf{z}_h) \quad (6)$$

The algorithm iterates till convergence, and it is demonstrated to be convergent for one and two-block models. After convergence, structural (or path) coefficients are estimated through OLS multiple regression among the estimated LV:

$$\xi_h = \beta_{h0} + \sum_{h'=1}^h \beta_{hh'} \xi_{h'} + \psi_h \quad (7)$$

In *Partial Possibilistic Regression Path Modeling* (PPR-PM), the estimation procedure starts by choosing arbitrary values for each latent variable. Each LV is then linked to its own block of MV by simple PR according to a *reflective model*:

$$x_{jh} = \tilde{\omega}_{1jh} + \tilde{\omega}_{2jh} v_h \quad (8)$$

The internal estimate z_h of each ξ_h is a linear combination of the indicators of the adjacent latent variables:

$$z_h = \sum_{j=1}^{J_h} c_{jh'} x_{jh'} + \sum_{j=1}^{J_h} c_{jh''} x_{jh''} + \dots \quad (9)$$

where h', h'', \dots refer to LV and MV of adjacent blocks. It is worth noticing that the LV are weighted linear combinations of the MV estimates midpoints, which are obtained from the PR in the outer model. The weights are the PR midpoints coefficients: the higher the midpoint coefficient the higher the contribution to the LV.

Finally the path coefficients are estimated by possibilistic regression between the estimated latent variables:

$$\xi_h = \tilde{\beta}_{h0} + \sum_{h'=1}^h \tilde{\beta}_{hh'} \xi_{h'} \quad (10)$$

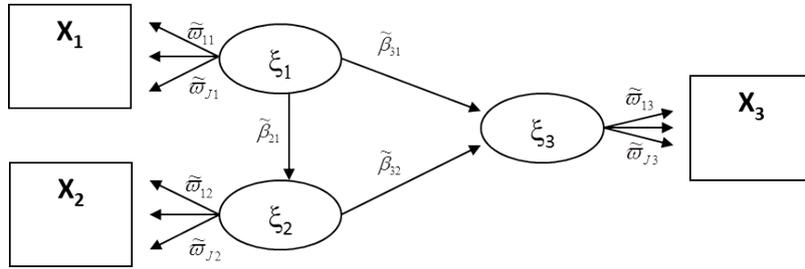


Fig. 1 Path Diagram

References

1. Bollen, K.A.: Structural equations with latent variables. Wiley, New York (1989)
2. Kim, K.J., Moskowitz, H., Koksalan, D.: Fuzzy versus statistical linear regression. *European J. Oper. Res.*, 92, 417-434 (1996)
3. Palumbo, F., Romano R.: Possibilistic PLS path modeling: A new approach to the multigroup comparison. In: Brito, P. (ed.) *Compstat 2008*, 303-314. Physica-Verl., Heidelberg (2008)
4. Palumbo, F., Romano R., Esposito Vinzi, V.: Fuzzy PLS path modeling: A new tool for handling sensory data. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) *Data Analysis, Machine Learning and Applications*, pp. 689-696. Springer, Heidelberg (2008)
5. Romano, R., Palumbo, F.: Fuzzy regression and least squares regression: the relationship between two different fitting criteria. *Atti del convegno SIS2006*, Torino, pp. 693-696, <http://old.sis-statistica.org/files/pdf/atti/Spontanee2006.693-696.pdf> (2006)
6. Tanaka, H., Guo, P.: *Possibilistic Data Analysis for Operations Research*. Physica-Verlag, Wurzburg (1999)
7. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y-M., Lauro, C.: PLS path modeling. *Computational Statistics and Data Analysis* **48**, 159-205 (2005).
8. Wold, H.: Modelling in complex situations with soft information. In: *Third World Congress of econometric Society*. Toronto, Canada (1975)

Classification of composite seismogenic sources through probabilistic score indices

Renata Rotondi

Abstract Six out of the ten natural disasters that caused the largest financial losses are earthquakes. If their forecasting is still, and will probably remain for a long time, an unsolved problem, however, it is time to start to draw up medium- and long-term protocols to be followed for the reduction of seismic risk. To do that we have to speak of prevention and of priorities of intervention, which imply the classification of the zones on the basis of their proneness to the seismic risk. In one of the possible probabilistic and geophysical frameworks supported by the present knowledge, some scoring procedures are presented which aim at different goals: comparing models, measuring their performance, identifying their limits, providing tools to the decision makers.

Key words: Gambling scores, scoring indices, renewal process, hazard function, Bayesian nonparametric inference

1 Introduction

In recent years there is a growing awareness that, on the one hand, the earthquake forecast is a complex, probabilistic-in-nature issue and, on the other hand, we have, on the basis of the current, albeit partial, knowledge, to start planning medium-term interventions for risk reduction. The simplest solution - retrofitting of all the buildings and infrastructures with high vulnerability to seismic code - is not feasible in a country like Italy with huge amount of historical buildings and remarkable fragility of the built heritage. In Italy, indeed, the ratio between damages caused by earthquakes and energy released during the events is much higher than in other countries with high seismicity like Japan, California and New Zealand.

Rotondi R,
CNR-IMATI, Via Bassini 15, 20133 Milano (I), e-mail: reni@mi.imati.cnr.it

The solution could be hence sought in the identification and classification of the areas most prone to seismic hazard, so as to focus on them the most urgent interventions for risk mitigation. To do that, it is necessary a partition of the national territory on tectonic bases, reliable estimations of the seismic hazard which take into account, as much as possible, any available information, methods for scoring the performance of forecasting probability models, an operational protocol which provides documented guidelines to political authorities for civil protection actions.

In this article we present an exercise implementing the points listed above in one of the possible probabilistic and geophysical frameworks.

2 Underlying probabilistic and geophysical assumptions

In the literature a variety of probability models have been proposed that assess the probability of occurrence of strong earthquakes in a seismic region at different time-space scales. Since we are interested in mitigation actions and prevention, we focus our attention on the medium-term models. One of the most shared assumption is that the occurrence probability depends on the time elapsed since the last event, which implies the choice of a renewal process. Consequently, the areas on which to apply this model must be identified so that the recurrence times between pairs of events localized in them satisfy the statistical conditions of independence and identical distribution. In the Italian tectonic environment it can be proved that these conditions are satisfied by the data that are associated with the composite seismogenic sources of the Database of Individual Seismogenic Sources (DISS) (version 3.0.2), and gathered according to the eight tectonically-coherent macro-regions which cover almost the whole Italy [1]. In addition to DISS, the other fundamental database is the historical earthquake catalogue; though the Italian CPTI04 catalogue is one of the best documented in the world, its inevitable incompleteness causes the poor quality and paucity of data in some of the datasets related to the abovementioned macro-regions. Bayesian nonparametric inference methods may reduce the negative consequences of these crucial features on the conclusions.

3 Measures for scoring

Many different performance measures for earthquake probability models have been proposed in the literature; the choice of the most suitable ones depends on the objectives of the study and on the methodologies applied. Possible aims are:

- to identify in which tectonic environment the model fits better the observed seismicity
- to compare the model in question with a reference model
- to establish probability threshold values, which, if exceeded, trigger an alert status or, at least, send a warning to the decision makers.

Considering probability forecasts the most natural scores are the information gain [3] and the gambling scores [4]. The former is defined as the logarithm of the likelihood ratio of the proposed model *versus* a reference model. It is essential that the models share the spatio-temporal scales and the objectives of the analysis. Moreover, also the statistical tools used as validation criteria must be consistent with these settings and with the methodological approach: so we will speak of hypothesis tests in the frequentist inference (paying attention to the hypotheses underlying the asymptotic properties) and of posterior predictive probabilities in the Bayesian approach.

The gambling scores evaluate the performance of probability forecasts in a betting framework: if a probability threshold has been assigned, i.e. as percentage of the hazard function, the forecast model bets against the bank (reference model) when its probability p_{occ} that at least an earthquake occurs in the forecasting horizon exceeds this threshold. On the contrary, without thresholds, the model bets p_{occ} on the occurrence and $(1 - p_{occ})$ on the non-occurrence, again against the bank. Of course the gambling schemes must be such as to be fair.

Many other measures have been proposed in disciplines with a long tradition in studies on prediction, like hydrology, medicine, economics, finance, engineering and, especially, meteorology. These disciplines differ from seismology in the possibility of recording large amounts of regular observations, and in the physical models that control the phenomena they study. On the contrary they have in common with seismology the study of rare events which require the development of specific validation scores in order to avoid misleading results, as with the well-known Finley tornado forecasts. The performance measures most used in these contexts are based on the Yes/No binary predictions, and therefore on the four counts obtained in a validation period: (a) the number of forecast hits; (b) the false alarms; (c) the correct rejections; and (d) the misses. Probability forecasts may also be reduced to binary predictions if, fixed a probability threshold, the exceedence of this threshold is considered as ‘Yes’ prediction and *viceversa*. Particular functions of (a)-(d) values generate indices which highlight the reliability and skill of probability forecasts; among them we consider the Hanssen-Kuiper score (S_{HK}), the equitable treat score (S_{ETS}), and the extreme dependence score (S_{EDS}). Other prediction characteristics are the fraction of failures to prediction $\nu = d/(a + d)$ and the fraction of time on prediction $\tau = (a + b)/(a + b + c + d)$; as well as being represented through Molchan error diagrams, the ν and τ measures can be used to evaluate γ -optimal thresholds, which minimise the loss functions $\gamma = \gamma(\nu, \tau)$, like $\gamma_1 = \nu + \tau$, $\gamma_2 = \max(\nu, \tau)$, $\gamma_3 = \tau/(1 - \nu)$, which increase in each argument.

Going back to the particular application of a renewal model to data from composite seismogenic sources, another score index, the rank-based score, can be obtained as a function of the rank of the sources hit by an earthquake in the time interval under validation, after ordering the sources with respect to decreasing occurrence probability.

The scoring measures that were used in the various combinations between different objectives and methodologies are summarized in Table 1.

Table 1 Scheme of scoring measures to use according to different objectives and methodologies.

		reference model	
		Yes	No
threshold probability	Yes	<ul style="list-style-type: none"> • gambling score (simple bet) 	<ul style="list-style-type: none"> • indices for binary predictions
	No	<ul style="list-style-type: none"> • gambling score (double bet) • information gain 	<ul style="list-style-type: none"> • rank-based score

4 Some comments on validation results

First of all it must be emphasized that forecast quality is an inherently multifaceted quantity, and only the examination of many different indices allows to capture the achievements of the forecaster with respect to each particular facet.

Since the rate of strong earthquakes ($M_w \geq 5.3$) in Italy is about one per year, you should wait a long time before having a significant number of events for prospective tests. A solution could be to plan retrospective studies in such a way that length of the time period under examination and quality of the corresponding database are balanced. In [2] the last century, divided in decades, has been considered by comparing the actually recorded earthquakes in each interval with the hazard function values $h(u, t) = \frac{F(t+u) - F(t)}{1 - F(t)}$, being $u = 10$ years and t the time elapsed since the last event in each source.

The application of the scoring measures presented in Section 3 has provided information on these topics:

- of which macro-regions the renewal model fits the seismicity better than others;
- which percentage of the hazard function is more appropriate to use in order to start an alert status;
- which features of the hazard function should be improved;
- whether the renewal model performs better than the Poisson model, chosen as reference model.

References

1. Rotondi, R.: Bayesian nonparametric inference for earthquake recurrence time distributions in different tectonic regimes. *J. Geophys. Res.*, **115**, B01302, doi:10.1029/2008JB006272 (2010)
2. Rotondi, R.: Retrospective validation of renewal-based, medium-term earthquake forecasts, submitted (2013)
3. Vere-Jones, D.: Probabilities and information gain for earthquake forecasting, *Comput. Seismol.*, **30**, 248-263 (1998)
4. Zhuang, J.: Gambling scores for earthquake predictions and forecasts, *Geophys. J. Int.*, **181**, 382-390 (2010)

On Wild Bootstrap and M Unit Root Test

Gabriella Schoier

Abstract In this paper we analyze the problem of testing for the presence of unit roots and cointegration in the case of macro-economic and financial time series. Traditional units test are not suitable for these series as they present permanent volatility shifts, so different tests have been proposed. We consider a wild bootstrap version of the M unit test. Instead of standard normal bootstrap residuals Student- t bootstrap residuals are used. Some simulations been performed comparing the results obtained using both type of bootstrap residuals. The application regards to the MSCI Equity Indices.

1 Unit root test and Wild bootstrap

The problem of testing for the presence of units roots and cointegration requires to use statistics which account for serial correlation in the error process. The traditional most used tests are the *Dickey-Fuller (DF)* test ([5]) and the *Augmented Dickey-Fuller (ADF)*. These tests are unreliable in the case of many macro-economic and financial variables which are characterized by permanent volatility shifts ([1],[2]).

To avoid this problem the *Phillips-Perron (PP)* ([14]) test, that takes into account the eventual presence of heteroschedasticity and autocorrelation of the errors, has been presented. Another solution has been proposed by Ng and Perron with *the M unit root test*, a modification of the Phillips-Perron test.([13]) and with the *efficient modified Phillips-Perron test* ([12]). Cavaliere and Taylor ([1],[2],[3]) consider a bootstrap version of the *M unit root* test applying the idea of the *local generalized least squares (GLS) detrending* proposed by ([6]).

Standard residual bootstrap methods consider the residuals as *i.i.d.*; this procedure is invalid in presence of conditional heteroschedasticity as is the case of many financial time series ([7],[8]). In order to avoid this problem Liu ([9]) developed the *wild*

¹ Gabriella Schoier, Department of Economics, Business, Mathematics and Statistics, University of Trieste, Italy, e-mail: gabriella.schoier@econ.units.it

I thank the Italian Ministry of Education, University and Research – MIUR grant “Multivariate statistical models for risk assessment” – for financial support”

bootstrap. The wild bootstrap is a *residual bootstrap* type methodology where the residual of the regression model are multiplied by a sequence of *i.i.d.* variables with standard Normal distribution, this allows to replicate the heteroskedasticity presents in the original shocks in the bootstrap shock ([10]). In the field of the unit root test this procedure is justified by the fact that under regularity conditions the asymptotic distribution of the test statistics is the same of the wild bootstrap test statistics.

Let us consider $T+1$ observation generated according to the follow reference data generating process (DGP) ([3]):

$$\begin{aligned} X_t &= d_t + Y_t \\ Y_t &= \alpha Y_{t-1} + u_t \\ u_t &= \sum_{j=0}^{\infty} c_j \varepsilon_{t-j} \quad t=0, \dots, T; \quad E(Y_0^2) < \infty; \end{aligned}$$

$d_t = \gamma' z_t$, $z_t = (1, t, \dots, t^p)'$; $\{u_t\}$ is a linear process in $\{\varepsilon_t\}$.

The bootstrap residuals u_t^b , are generating through the OLS $\hat{u}_{t,k}$ k is the lag obtained on the base of a regression based on the *GLS detrending* of \hat{X}_t according the relation: $u_t^b = \hat{u}_{t,k} w_t$ where $\{w_t\}_{t=1}^T$ is a sequence of independent $N(0,1)$.

After generating the bootstrap sample, the statistics of the M unit root test wild bootstrap corresponding to MZ_α , MZ_t , MSB , the p-values and their standard errors are generated ([3]).

We consider a modified version of the algorithm where the $\{w_t\}_{t=1}^T$ are distributed according to a *Student-t* with n degrees of freedom . To test this modification two series of 100 values are generated *i.e.* a *GARCH(1,1)* with parameters: $\alpha = 0.5, \beta = 0$ and v_t *i.i.d* $N(0,1)$ and a *GARCH(1,1)* with parameters: $\alpha = 0.3, \beta = 0.65$ and v_t *i.i.d* t_5 . The p-values and SE p-values estimates for the M statistics for 1000 bootstrap replications are reported in Tab. 1

Table 1: p-values and SE p-values estimates

gdl	AR	MA	M Statistics	p-value $w_t \sim N(0,1)$	SE p-value $w_t \sim N(0,1)$	p-value $w_t \sim t_n$	SE p-value $w_t \sim t_n$
<i>GARCH(1,1) with: $\alpha = 0.5, \beta = 0$ e v_t i.i.d $N(0,1)$</i>							
5	0.5	0.8	MZ_α	0.016	0.008	0.032	0.011
			MZ_t	0.008	0.006	0.024	0.010
			MSB	0.028	0.010	0.04	0.012
5	0.8	0.5	MZ_α	0.364	0.030	0.38	0.031
			MZ_t	0.360	0.030	0.352	0.030
			MSB	0.472	0.031	0.508	0.032
<i>GARCH(1,1) with: $\alpha = 0.3, \beta = 0.65$ e v_t i.i.d t_5</i>							
5	0.5	0.8	MZ_α	0.112	0.020	0.088	0.018
			MZ_t	0.18	0.024	0.16	0.023
			MSB	0.132	0.021	0.096	0.019
5	0.8	0.5	MZ_α	0.080	0.017	0.06	0.017
			MZ_t	0.112	0.020	0.097	0.019
			MSB	0.088	0.018	0.101	0.018

As one can see the bootstrap estimations of the p-values and SE p-values are similar.

MSCI's Daily Total Return methodology

The MSCI Equity Indices measure the performance of a set of equity securities over time. The MSCI Equity Indices are calculated using the Laspeyres' concept of a weighted arithmetic average together with the concept of chain-linking. MSCI country and regional equity Indices are calculated in "local currency" as well as in USD, with price, gross and net returns process. Price indices measure the market prices performance for a selection of securities. They are calculated daily and, for some of them, on a real time basis. Each index captures the market capitalization weighted return of all constituents included in the index.

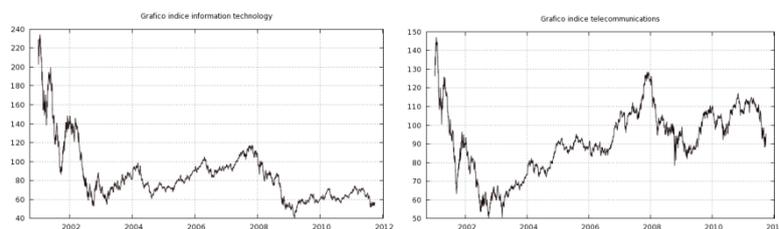
The MSCI indices considered follow the *Global Industry Classification Standard (GICS)*, are related to the period 29/11/2000 30/09/2011 for each index 2806 observation have been observed. The indices are referred to ten industrial sectors: MSCI EMU ENERGY (M1MUE1E(RI)), MSCI EMU MATERIALS (M1MUM1E(RI)), MSCI EMU INDUSTRIALS (M1MUIDE(RI)), MSCI EMU CONS DISCR (M1MUCDE(RI)), MSCI EMU CONSUMER STAPLES (M1MUCSE(RI)), MSCI EMU HEALTH CARE (M1MUHCE(RI)), MSCI EMU FINANCIALS (M1MUFNE(RI)), MSCI EMU INFORMATION TECHNOLOGY (M1MUIE(RI)), MSCI EMU TELECOMMUNICATIONS SERVICES (M1MUT1E(RI)), MSCI EMU UTILITIES (M1MUU1E(RI)) ([11]).

According to the traditional ADF test the MSCI EMU index for Information Technology (M1MUIE(RI)) and the MSCI for Telecommunications Services (M1MUT1E(RI)) are stationary as one can see from Tab. 3, while looking at Fig.1 this is not true.

Table 3: ADF test

MSCI EMU Index	Statistics	Model	retard	P-value
ENERGY	-1.776	With constant	10	0.397
MATERIALS	-0.096	no constant no trend	1	0.586
INDUSTRIALS	-0.251	no constant no trend	1	0.536
CONSUMER DIS.	-0.567	with constant and trend	19	0.338
CONSUMER STAP.	-1.985	with constant and trend	10	0.585
HEALTH CARE	-1.651	with constant	10	0.444
FINANCIALS	-1.188	no constant no trend	18	0.238
INFORM. TEC.	-4.597	with constant and trend	22	<0.01
TELECOM.S	-4.879	with constant and trend	17	<0.01
UTILITIES	-0.010	no constant no trend	13	0.613

Figure 1: the MSCI EMU indices for Information Technology and the for Telecommun. Services



Let us consider the M unit root test and the ADF test applying the wild bootstrap methodology with $N=1000$ replications. In Tab. 4 and Tab. 5 the values of the wild

bootstrap (WB), of the statistics evaluated on the GLS de-trended data, the p-value estimates and their standard errors are reported.

Table 4: the MSCI EMU index for Telecommunications Services

STATISTICS	WB	REAL	P-VALUE	SE P-VALUE
MZalfa	-4.640	-1.885	0.89	0.031
MSB	0.277	0.422	0.85	0.036
MZt	-1.284	-0.795	0.89	0.031
ADF	-1.318	-0.924	0.63	0.048

Table 5: the MSCI EMU index for Information Technology

STATISTICA	WB	REAL	P-VALUE	SE P-VALUE
MZalfa	-4.082	-3.190	0.81	0.039
MSB	0.326	0.378	0.85	0.036
MZt	-1.331	-1.207	0.78	0.041
ADF	-1.435	-1.203	0.48	0.050

As one can see according to the p-value of the M and ADF wild bootstrap the two MSCI EMU indices are not stationary.

References

1. Cavaliere Giuseppe, Taylor Robert A. M., (2004), Testing for unit roots in time series models with nonstationary volatility, Working paper 04-24, Department of Economics, University of Birmingham.
2. Cavaliere Giuseppe, Taylor Robert A. M., (2008), Bootstrap unit root tests for time series models with non-stationary volatility, *Econometric Theory*, Vol. 24, 43-71.
3. Cavaliere Giuseppe, Taylor Robert A. M., (2009), "Bootstrap M unit root test", *Econometric Reviews*, Vol. 28, No. 5, 393-421.
4. Davidson Russel, Flachaire Emmanuel, (2008), The wild bootstrap, tamed at last, *Journal of econometrics*, Vol. 146, 162-169.
5. Dickey D. A., Fuller W. A., (1979), Distribution of the estimators for autoregressive time series with unit root, *Journal of the american statistical association*, Vol.74, No. 366, 427-431.
6. Elliot Graham, Rothemberg Thomas J., Stock James H., (1996), Efficient test for an autoregressive unit root, *Econometrica*, Vol. 64, No. 4, 813-836.
7. Gonçalves Silvia, Kilian Lutz, (2004), Bootstrapping autoregressions with conditional heteroskedasticity of unknown form, *Journal of econometrics*, Vol. 123, 89-120.
8. Gonçalves Silvia, Kilian Lutz, (2007), Asymptotic and bootstrap inference for $AR(\infty)$ processes with conditional heteroskedasticity, *Econometric Reviews*, Vol. 26, No.6, 609-641.
9. Liu Regina Y., (1988), Bootstrap procedure under some non-i.i.d. models, *The annals of statistics*, Vol. 16, No.4, 1696-1708.
10. Mammen Enno, (1993), Bootstrap and wild bootstrap for high dimensional linear models, *The annals of statistics*, Vol. 21, No.1, 255-285.
11. MSCI Global Investable Market Indices Methodology May 2012, agosto 2012 http://www.msci.com/eqb/methodology/meth_docs/MSCI_May12_GIMIMethod.pdf
12. Ng Serena, Perron Pierre, (2001), Lag length selection and the construction of unit root tests with good size and power, *Econometrica*, Vol. 69, No. 6, 1519-1554.
13. Ng Serena, Perron Pierre, (1996), Useful Modifications to some unit root tests with dependent errors and their local asymptotic properties, *The review of economic studies*, Vol. 63, No. 3, 435-463.
14. Phillips Peter C.B., Perron P., (1998), Testing for Unit Roots in Time Series Regression, *Biometrika*, Vol. 75, No. 2, 335-346.

On the implementation of a parallel algorithm for variable selection in model-based clustering

Luca Scrucca

Abstract Variable selection in model-based clustering is often used to improve cluster identification. However, available algorithms need to operate on a large search space and, therefore, can be time consuming. Following the recent surge of interest in distributed processing, in this contribution we discuss the implementation of a parallel algorithm for variable selection in R. We conducted a simulation study to assess the performance of the proposed parallel variable selection algorithm. The results show that the increase of speed reached follows the well-known Amdahl's Law for the speedup achievable when using multiple processors.

Key words: Variable selection, Model-based clustering, Greedy search, Parallel computing, R.

1 Introduction

In the model-based approach to clustering each cluster is represented by a parametric distribution. A Gaussian finite mixture model is often used to model multivariate continuous data. Selecting a subset of relevant clustering variables allows to achieve parsimony of unknown parameters and improve cluster identification.

Raftery and Dean (2006) discussed the problem of variable selection for model-based clustering by recasting the problem as a model selection procedure. Their proposal is based on the use of BIC to approximate Bayes factors for comparing mixture models fitted on nested subsets of variables. Refinements to this approach have been proposed by Maugis et al. (2009).

The above mentioned approach to variable selection is implemented through a greedy search algorithm, which, depending on the number of features involved and the number of truly clustering variables, can be very time consuming. In this con-

Luca Scrucca
Department of Economics, Finance and Statistics
University of Perugia, Italy
Via A. Pascoli, 20 – 06123 Perugia, Italy
e-mail: luca@stat.unipg.it

tribution we consider the possibility to parallelize the search algorithm to speedup computing time, and we present some preliminary results.

2 A method for variable selection based on models comparison

Assume that the set of available variables can be partitioned into three disjoint parts: the set of selected variables, X_1 , the variable being considered for inclusion or exclusion from the active set, X_2 , and the set of irrelevant variables, X_3 . Raftery and Dean (2006) showed that the inclusion (or exclusion) of X_2 from the active set can be assessed by using the Bayes factor, which can be approximated by the following BIC difference:

$$\text{BIC}_{\text{diff}} = \text{BIC}_{\text{clust}}(X_1, X_2) - \text{BIC}_{\text{not clust}}(X_2|X_1), \quad (1)$$

where $\text{BIC}_{\text{clust}}(X_1, X_2)$ is the BIC value for the “best” clustering mixture model fitted using both X_1 and X_2 features, whereas $\text{BIC}_{\text{not clust}}(X_2|X_1)$ is the BIC value for no clustering for the same set of variables. This last term can be computed as

$$\text{BIC}_{\text{not clust}}(X_2|X_1) = \text{BIC}_{\text{clust}}(X_1) + \text{BIC}_{\text{reg}}(X_2|X_1), \quad (2)$$

i.e., as the sum of the BIC value for the “best” clustering model fitted using X_1 plus the BIC value for the regression of the candidate variable X_2 on the X_1 variables. In all cases the “best” clustering model is identified with respect to both the number of mixture components and model parametrization. A subset selection step is also needed in the regression part of the model to select the predictor variables. An important aspect of this formulation is that the set X_3 of remaining variables plays no role in the adopted criterion (1).

The method described above can be used to implement a stepwise greedy search. A backward-type algorithm starts with all the variables in the set X_1 , and, at each step, removes a candidate variable depending on the criterion in (1) being negative. When two or more variables have already been excluded, the inclusion of one of them can be considered at each step if the BIC difference becomes positive. The algorithm stops when no further variables could be dropped or added from the set of clustering variables X_1 . Of course, a forward-type algorithm could also be adopted.

3 Parallelism

Parallel computing is a form of computation in which many calculations are performed simultaneously, either on a single multi-core processors machine or on a cluster of multiple computers. By using a machine with P processors instead of just one, we would like to obtain an increase in calculation speed of P times. However, this is not the case, as in the implementation of a parallel algorithm there are some

inherent non-parallelizable parts and communication costs between tasks (Nakano, 2012).

Amdahl’s Law (Amdahl, 1967) is often used in parallel computing to predict the theoretical maximum speedup when using multiple processors. If f is the fraction of non-parallelizable task and P is the number of processors in use, then the maximum speedup achievable on a parallel computing platform is given by

$$S_P = \frac{1}{f + (1 - f)/P} \quad (3)$$

In the limit the above ratio converges to $S_{\max} = 1/f$, which represents the maximum increase of speed achievable in theory, i.e., by a machine with an infinite number of processors.

4 Implementing a parallel algorithm for variable selection in R

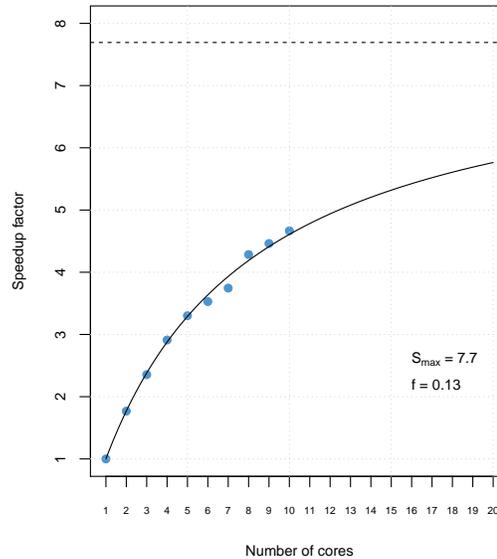
R (R Core Team, 2013) package `parallel` was first included in R 2.14.0. This is essentially a merger of the `multicore` package and the `snow` package. The `multicore` functionality supports parallelism via forking, which is a concept from POSIX operating systems, so it is available on all R platforms except Windows. On the contrary, `snow` supports different transport mechanisms (e.g., socket connections) to communicate between the master and workers, and it is available on all operating systems. Other approaches to parallel computing in R are available as described in McCallum and Weston (2011).

The backward search discussed in Section 2 constitutes an embarrassingly parallel problem, i.e., one for which little or no effort is required to separate the problem into a number of parallel tasks. Essentially, the sequential evaluation of candidate variables for inclusion or exclusion, which is the more time consuming task, can be done in parallel. For the actual implementation we used the `doParallel` package, a “parallel backend” which acts as an interface between the `foreach` package and the `parallel` package. Essentially, it provides a mechanism needed to execute for loops in parallel.

To investigate the performance of our parallel algorithm implementation, we conducted a small simulation study. We consider a ten dimensional problem where only the first two variables contain clustering information. Two equiprobable clusters have been generated from Gaussian distributions with, respectively, means $\mu_1 = (0, 0)$, $\mu_2 = (3, 3)$, and covariances $\Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1.5 & -0.7 \\ -0.7 & 1.5 \end{bmatrix}$. Two more features highly correlated with the clustering variables were generated, while the six remaining features are simply noise variables, some independent and some correlated among them. 200 observations were generated from the above simulation setting, and backward subset selection was applied for increasing number of cores. The study was performed on a 24 cores Intel® Xeon® CPU X5675 running at 3.07GHz and with 128GB of RAM.

Figure 1 shows the results averaged over 10 replications. The points represent the observed speedup factor (obtained as $s_P = t_P/t_1$, where t_P is the system time employed using P cores) for running the backward algorithm with up to 10 cores. The curve represents the Amdahl's Law (3) with f estimated by non-linear least squares. It turns out that the fraction of sequential part of the backward algorithm for variable selection is estimated as $\hat{f} = 0.13$, which yields a maximum speedup of $S_{\max} = 7.7$.

Fig. 1 Graph of speedup factor vs the number of cores employed in the parallel algorithm for backward subset selection in model-based clustering. The estimated fraction of non-parallelizable task (f) is estimated by fitting the Amdahl's Law equation with non-linear least squares. This allows us to compute the maximum speedup achievable by parallelization, which is around 7.7 times the sequential algorithm.



References

- Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *AFIPS Conference Proceedings*, volume 30, pages 483–485.
- Maugis, C., Celeux, G., and Martin-Magniette, M. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882.
- McCallum, E. and Weston, S. (2011). *Parallel R*. O'Reilly Media.
- Nakano, J. (2012). Parallel computing techniques. In Gentle, J. E., Härdle, W. K., and Mori, Y., editors, *Handbook of Computational Statistics*, pages 243–271. Springer, 2nd ed. edition.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

The Role of Learning Measurement in the Governance of an Education System: an Overview of the Issues

Paolo Sestito

Abstract This short note considers the potential advantages - and the relevant risks - stemming from the use of universal standardised students' tests in the governance of the education system. The reference is to a large extent to the Italian case, where student tests have been only recently introduced.

Key words: school evaluation systems, standardised tests, student assessment and school evaluation

1 Introduction

Schools' and teachers' accountability has been at the forefront of several policy initiatives and education reforms over the last 15 years all over the world. The US has experienced the No Child Left Behind (NCLB) initiative, while in several European countries there has been either the introduction of schools' league tables (coupled with the introduction of schools' competition in the recruitment of their students, as in UK and Sweden) or a strengthened standardisation of students' assessment (as in Germany).

The rationale for the accountability drive lies in the idea that fostering schools' autonomy - another policy fashion over the period - needs to be coupled with strengthening school's accountability. The latter is necessary in order to avoid schools becoming self-referencing and inflating the credentials by them attributed to their students. The evidence in favour of accountability is the more impressive as both the quantity and the apparent quality of many schools' inputs appear to be rather ineffective in fostering students' learning. The students-to-teacher ratio or even the number of teaching hours do not appear to matter a lot and even traditional measures of teachers' quality - for instance the credentials they acquired and

Paolo Sestito
Banca d'Italia and Invalsi, e-mail: paolo.sestito@bancaditalia.it

the amount of training received - are not much related to the learning outcomes of their students. So, huge differences appear to be present among schools and among teachers within a school - between high-and low-value added people and institutions - and they seem to be due to intrinsic quality and motivation.

The literature on these themes is too wide to be here summarised. On the limited relevance of schools' resources a classic paper is Hanushek (2006). On the huge and largely unexplained (by traditional factors like credentials, training and work experience) differences in teachers' quality, see Chetty, Friedman and Rockoff (2012).

2 A brief reminder about the Italian experience

What Italy has so far experienced has been a gradual and progressive introduction of standardised student tests in several grades, motivated precisely by the lack of comparability among locally attributed official marks. While in most grades the tests are low stake exercises, in the 8th grade the tests are part of a national exam attempting to provide a standardised benchmark for the official marks provided at the end of the first education cycle. Particularly in this grade they have had also a cultural orientation role, exemplifying what the otherwise rather vague nationally stated curricular indications are meant for; on practical grounds, their content has been very much derived from the experiences of the large international assessments (the ones conducted by IEA and the PISA-OECD), measuring relatively broad skills more than tests to whom one can be trained through simple mnemonic exercises.

The results of the testing exercises are given back to the individual school and are not made available to the general public, unless the school itself decides to advertise them, as some have started to do in order to advertise their own efficacy. No premia to either schools or individual teachers are attributed on the basis of those results and actually there is no centralised system linking students results (in level terms or in terms of value added longitudinally constructed) to their teachers. Possibly because of the fears that, maybe in the future, these tests might be used in order to evaluate individual teachers or in order to set up schools' races, their use has raised several concerns among teachers. A testimony of such opposition is in the widespread cheating to the tests which is identified, particularly in some Southern regions.

3 Potential and pitfalls of standardised tests as a measure of students' achievements

Teachers periodically test their own students in order to fine tune their future teaching activities. Having to pass high stakes exams provide students with strong incentives to put some effort in what they do. Not necessarily this testing has to be standardised. Yet the use of psychometric techniques and some standardisation may

help along both lines: firstly, it allows to better certify the results of those high stakes tests - with possible benefits in the labour market, where credible certificates may be more reliably used - preventing grade inflation and a downward drift in the standards concretely applied; furthermore it may boost the diagnostic quality of the tests themselves, as standardised tests usually derive from explicitly well specified constructs, are pre-tested and may provide for benchmarks otherwise unavailable to the single teacher setting up her own testing machine.

Precisely for the same reasons, standardised tests may become dangerous. Their contents may be too narrow - hindering a balanced overall assessment of the individual student. Whenever they acquire a high stake nature - either for the students or for the teachers of those students - such narrowness and their fore-seeability may distort incentives for both students and teachers: instead of stimulating learning as a global and balanced activity, they may distort classrooms' activity towards studying and teaching to-the-test. Furthermore, overt cheating to-the-test behaviour may arise.

Both technical and institutional arrangements may help preventing these risks. A first point relates to the need to combine different test exercises, balancing their respective pros and cons. So some tests may be high-stakes for the individual student, having an incentives' role, while others may be low-stakes and the latter may be used to study the possible biases arising in the former. Some tests may be conducted at the start of a given school segment - emphasizing the diagnostic identification of learning potential from the viewpoint of the teachers having to work with the students, so as to minimize the distortions possibly arising from their perception of being high-stakes exercises from either students or teachers (or both) - and still be used as an estimate of the arrival point of the previous school segment. Combining these different tests over time - properly modelling the differences in their nature - may still provide an estimate of the learning progress of a community of students, which is what any value added metric (and the accountability of either teachers or schools) is aiming at.

Whenever some tests, possibly because of their timing, are perceived as high stakes, in their construction one has to emphasize the fact of being unforeseeable, so as to downplay the risk of teaching to the test. In any case, it may be advisable that high stakes students' exams - and more generally the measures of either a school's or a teacher's achievements - are not uniquely driven by a single test exercise: students certifications are better based upon an array of different tests, some standardised and some more flexibly left to local educators and aiming at an overall (somehow olistic) assessment of each individual student, so that no single test becomes the exclusive focus of studying and teaching activities; similarly, the performance of teachers and schools have to be estimated combining different measures.

Combining different test exercises may also take account of the tradeoffs existing in each test measurement properties. The precision at the level of each individual student - in terms of whether he or she has reached a given threshold, reachable with many questions focusing around the boundary of that threshold - may impede the depiction of the spectrum for the entire distribution of abilities over a wider population - in terms of how much the least and most able groups of students differ

from each other. Again this suggests to combine different testing exercises instead of focusing upon any single measure.

The details of any such combination may evolve over time, with new opportunities and new problems arising from technical progress. The computer delivery may allow for potentially richer individualised tests, as more complex questions may be posed, overcoming the strictness of closed end answers and introducing adaptive tests, the complexity of each item being driven by the ability shown by the individual student. This may add to the precision of the scoring at the individual level, while still preserving the use of items covering different levels of ability so as to describe the overall population of students. The computer delivery, even without any proper shift to adaptive testing, may also help in preventing cheating, as it allows for machine corrected individualised sequences of tests, so that students may not copy from each other and their own teachers and vigilators are not asked to grade their responses.

4 Reliability and (in)sufficiency of measures derived from standardises tests as indexes of schools' and teachers' performance

The performance of individual schools and teachers may not be immediately equated to the performance in the testing exercises of their students. Quite simply, students' achievements depend upon a host of factors, many of which are besides the direct influence of the individual school or the individual (group of) teachers. One should so focus upon value added estimates of what a school (and a teacher) has contributed to students' learning progresses over time. The derivation of value added measures however raises a lot of technical issues.

Properly considered value added measures require longitudinal data, linking students results over time. Alternatively, but somehow less precisely, the results of students at the sectional level may be purged from compositional effects by controlling for as many as possible covariates affecting students' achievements (firstly those related to the familiar background). Notice that test results may be plagued by measurement error, including into it the impact of any temporary factor impinging upon a student performance in a test; a peculiar measurement error may also arise when aggregating data at the school level, as the overall performance of a school, particularly when of small size, may be driven up or down by the presence of few abnormal individuals. While this second aspect is dealt with by the use of longitudinal data - as the permanent idiosyncratic effect of a given individual is removed by focusing upon changes over time - the first problem may be actually exacerbated by the use of longitudinal data. The general message is that, whichever approach is used - the properly longitudinal one or the less precise cross sectional pruning from compositional effects - value added measures quite often are reliable only insofar as the rather extreme categories of excellent and very bad performers are concerned. The ranking of most of the intermediate performers - be them individual schools or

individual teachers - is not sufficiently precise to say with enough certainty whether, and by how much, x is better than y .

On top of these purely statistical aspects, at least 3 interpretive issues arise when one tries to implement a value added approach in order to measure schools' and teachers' performance. A first one relates to the reliability of short term value added measures as proxies of the longer term measures one would like to know. A second point relates to the identification problem connected to the endogenous sorting of students across schools and teachers. Teachers and schools who apparently are the best performers might simply be the ones who happened to mate with the students with the best learning potential. A last issue relates to the policy relevance of either the school or the (individual) teacher effects. Students achievements, besides being affected by their previous history and the familiar and general context where they live, are a complex function of the group of students and the group of teachers with whom they interact; these classrooms effects are further shaped by the overall school environment. In this multi-level and multi-agent nature of schools' and teachers' impact upon students learning is difficult to identify the contribution of each single element (because of the sorting mechanism above described) and acting upon one of them, neglecting the picture of the entire network, may be problematic.

5 Conclusions

Centrally engineered performance indicators based upon students' assessment may help the governance of an education system. Their use has to take into account the specificities of those governance rules and the methodological aspects of students' assessment, which do not provide for simple and immediately to be used measures. A multiplicity of indicators needs to be used, so as to avoid making one indicator an easily to be corrupted benchmark and to take into account the fuzziness of learning processes. One has to recognize the lack of precision in the ranking for most performers, as only the two tails of the distribution are likely to be identified with sufficient precision. More generally, with differences among systems, very often the performance indicators may act more fruitfully as trigger (and constrain) to further decisions and processes, with no automatism.

References

1. Hanushek, E.: School Resources. In: Hanushek, E., Welch, F. (eds.) Handbook of Education Economics, North Holland (2006)
2. Chetty, R., Friedman, J.N., Rockoff, J.E.: The Long-Term Impacts of Teachers: Teacher Value-Added and Students Outcomes in Adulthood. NBER Working Paper No. 17699 (2012) Available at <http://www.nber.org/papers/w17699>

Novelty Detection with Support Vector Machines

John Shawe-Taylor and Blaž Žličar

The paper reviews algorithms and analysis of the application of Support Vector Machines to the problem of novelty detection. These include the relation to level set estimation and generalisation bounds for the false rejection of non-novel inputs. We will also describe applications of the approach to brain scan analysis for the detection of depression in mental health patients. Other applications will also be discussed.

Multidimensional scaling with incomplete distance matrices: an insight into the problem

Nadia Solaro

Abstract Given the substantial lack of contributions in the literature, the problem of handling incompleteness in proximity matrices is faced from a theoretical point of view with specific reference to Euclidean distance. In our intentions, the work provides a first contribution on which a more general approach could be grounded.

Key words: classical MDS, Euclidean distance, missing distance

1 Introduction

In spite of the extremely vast literature on missing data, the specific problem of proximity matrices with missing entries inexplicably seems to have not attracted as much attention. Proximity matrices, containing measures of pairwise similarity/dissimilarity within a set of units, are basic structures of multidimensional scaling (MDS) techniques [1, 2]. Within the MDS framework, missing proximities are substantially handled according to two approaches. If proximities can be computed from a data matrix, then a zero weight can be included into the proximity measure formula to zero out proximity scores pertaining to pairs of units with at least a missing datum. Otherwise, if a proximity matrix is already given, missing entries can be handled by choosing an appropriate MDS technique that allows a zero weight to be assigned to them. This option has not however been theoretically developed in every MDS method. In any way, to our knowledge, very few efforts have been made to propose a method for imputing missing proximities, rather than discard them from analyses. This work can then be seen as a contribution in this direction. Given the complexity of the matter, attention will be here confined to Euclidean distance matrices, which can be considered as a reference case.

Nadia Solaro
Department of Statistics and Quantitative Methods, University of Milano-Bicocca,
e-mail: nadia.solaro@unimib.it

2 Handling incompleteness through Euclidean properties

Our proposal for handling incompleteness of a Euclidean distance matrix has as main premises the set of theoretical results established by Gower with reference to: (1) the relationship between Euclidean space properties and the existence of different configurations of coordinates reproducing the same distances for n points [4], and: (2) a method for adding a new point in a given multidimensional space [3].

(1) *Euclidean space properties and configurations of points.* In [4], a fundamental result establishes the existence of different configurations that can perfectly reproduce the same Euclidean distance matrix in a multidimensional Euclidean space. Let $\mathbf{\Delta} = [\delta_{ij}]_{i,j=1,\dots,n}$ be a $n \times n$ symmetric Euclidean distance matrix, with zero diagonal elements, referred to a set of n units. Set: $\mathbf{A} = [-\frac{1}{2}\delta_{ij}^2]_{i,j=1,\dots,n}$, and $\mathbf{B} = \mathbf{H}_s \mathbf{A} \mathbf{H}_s^t$ where: $\mathbf{H}_s = \mathbf{I}_{(n)} - \mathbf{1}\mathbf{1}^t$, with: $\mathbf{I}_{(n)}$ the n -order identity matrix, $\mathbf{1}$ a vector of n ones, and \mathbf{s} a n -dimensional vector such that: $\mathbf{s}^t \mathbf{1} = 1$. In particular, being $\mathbf{\Delta}$ Euclidean, \mathbf{B} is positive semi-definite (p.s.d.) (the converse also holds) [4]. Then, it is always possible to find a $n \times r$ configuration of points \mathbf{X} , with: $r = \text{rank}(\mathbf{B}) \leq n - 1$, i.e. the maximum possible number of dimensions, such that: $d_{ij}(\mathbf{X}) \equiv \delta_{ij}$ for all couples i, j , where: $d_{ij}(\mathbf{X}) = d_{ij}$ expresses the Euclidean distance of units u_i and u_j computed with the coordinates of the r -dimensional space (i.e. the rows \mathbf{x}_i^t and \mathbf{x}_j^t of \mathbf{X}). Next, denoting by $\mathbf{B}_x = \mathbf{X}\mathbf{X}^t$ the inner-products matrix of \mathbf{X} (also called the Gram matrix), we have: $\mathbf{B}_x \equiv \mathbf{B}$, and: $\delta_{ij}^2 \equiv d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$, for all i, j , being, respectively, b_{ii} and b_{jj} diagonal, and b_{ij} off-diagonal, elements of \mathbf{B}_x . The configuration \mathbf{X} is not unique. In fact, other solutions in r dimensions exist with different translations of origin and different rotations of axis such that they perfectly reproduce $\mathbf{\Delta}$. Gower's fundamental result [4] concerns the form that the Gram matrix of such configurations must satisfy to reproduce the same matrix $\mathbf{\Delta}$. Assume that there exists another $n \times r$ configuration \mathbf{Y} with the property: $\mathbf{D}(\mathbf{Y}) \equiv \mathbf{\Delta}$, where henceforth \mathbf{D} will denote the Euclidean distance matrix computed from a configuration in r dimensions. Then, the Gram matrix of \mathbf{Y} , i.e. $\mathbf{B}_y = \mathbf{Y}\mathbf{Y}^t$, has the form: $\mathbf{B}_y = \mathbf{A} + \mathbf{G}$, with matrix \mathbf{G} , square symmetric, given by: $\mathbf{G} = \mathbf{g}\mathbf{1}^t + \mathbf{1}\mathbf{g}^t$, where the n elements of vector \mathbf{g} are one and a half the diagonal elements of \mathbf{G} : $\mathbf{g} = \frac{1}{2} \text{vecdiag}(\mathbf{G})$. Vector \mathbf{g} has to be chosen in a way that \mathbf{B}_y is p.s.d., and it is always possible by the Euclideanarity of $\mathbf{\Delta}$. Moreover, for any given \mathbf{g} it is possible to find a vector \mathbf{s} such that \mathbf{B}_y can be factorized into the product: $\mathbf{H}_s \mathbf{A} \mathbf{H}_s^t$. The choice of a vector \mathbf{g} fulfilling such characteristics is not unique, and thus it is not unique the associated configuration \mathbf{Y} , but whichever \mathbf{g} and \mathbf{Y} are, the reproduced distances will be still the same.

A well-known method for recovering the coordinates of a $n \times r$ configuration \mathbf{X} is provided by Classical Multidimensional Scaling (CMDS) [1, 2], which works on the factorization $\mathbf{H}_s \mathbf{A} \mathbf{H}_s^t$ specified with: $\mathbf{s} = \mathbf{1}/n$. By the spectral decomposition theorem, we have: $\mathbf{B}_x = \mathbf{X}\mathbf{X}^t = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^t$, where $\mathbf{\Lambda}$ is a diagonal matrix of the r positive eigenvalues of \mathbf{B}_x , taken in non-increasing order, and $\mathbf{\Gamma}$ is the corresponding $n \times r$ column-orthonormal matrix of eigenvectors (i.e. $\mathbf{\Gamma}^t \mathbf{\Gamma} = \mathbf{I}$). It descends that: $\mathbf{X} = \mathbf{\Gamma}\mathbf{\Lambda}^{1/2}$, where: $\mathbf{\Lambda}^{1/2} = \text{diag}[\sqrt{\lambda_k}]_{k=1,\dots,r}$. In addition, $\mathbf{X}^t \mathbf{1} = \mathbf{0}$ and $\mathbf{X}^t \mathbf{X} = \mathbf{\Lambda}$, from which it is apparent that the r dimensions are zero-mean and uncorrelated.

(2) *Adding a new point in a Euclidean space.* Given a Euclidean distance matrix Δ and a CMDS $n \times r$ configuration \mathbf{X} , Gower proposed a method for adding a new unit u_{n+1} to the set of the n units with coordinates \mathbf{X} [3]. The method assumes that the distances $\delta_{n+1,i}$ of unit u_{n+1} and each u_i are known. Moreover, it was shown that, in general, embedding a new unit in the r -dimensional Euclidean space requires an extra $(r+1)$ -th dimension, ($r \leq n-1$). Let \mathbf{x}_{n+1} be the vector of the $r+1$ coordinates of u_{n+1} to be determined, and $\boldsymbol{\delta}$ the vector with elements $d_i^2 - \delta_{n+1,i}^2$, with: $d_i^2 = b_{ii}$, $i = 1, \dots, n$. Then, the r coordinates $\mathbf{x}_{n+1,1:r} = \mathbf{x}_*$ are given by: $\mathbf{x}_* = \frac{1}{2} \mathbf{\Lambda}^{-1} \mathbf{X}' \boldsymbol{\delta}$, while for the $(r+1)$ -th coordinate $x_{n+1,r+1}$ we have: $x_{n+1,r+1}^2 = -\frac{1}{n} \boldsymbol{\delta}' \mathbf{1} - \mathbf{x}_*^t \mathbf{x}_*$, from which it is clear that $x_{n+1,r+1}$ is determined in value, but not in sign.

Handling incompleteness in a Euclidean distance matrix The previous results serve as theoretical reference for our proposal, here traced in its germinal idea. Assume to have a set of N units with pairwise Euclidean distances collected in the matrix $\Delta^* = [\delta_{ij}]_{i,j=1,\dots,N}$. In addition, assume that Δ^* contains a subset of missing distances such that Δ^* can be partitioned as follows:

$$\Delta^* = \begin{bmatrix} \Delta & \tilde{\Delta} \\ \tilde{\Delta}' & \Delta_{\text{NA}} \end{bmatrix}, \quad (1)$$

where Δ is a $n \times n$ complete distance matrix, Δ_{NA} is a $t \times t$ matrix with entirely Not Available (NA) distances ($t = N - n$), and $\tilde{\Delta}$ is a $n \times t$ matrix with known distances between each of the n and t units. Define: $\mathbf{A}^* = [-\frac{1}{2} \delta_{ij}^2]_{i,j=1,\dots,N}$, and: $\mathbf{B}^* = \mathbf{H}^* \mathbf{A}^* \mathbf{H}^*$, with: $\mathbf{H}^* = \mathbf{I}_{(N)} - \frac{1}{N} \mathbf{1} \mathbf{1}'$, and $\text{rank}(\mathbf{B}^*) = r \leq N - 1$. If Δ^* were completely known, we could obtain the CMDS solution: $\mathbf{X}^* = \mathbf{\Gamma}^* \mathbf{\Lambda}^{*1/2}$ in r dimension ($\mathbf{\Gamma}^*$ and $\mathbf{\Lambda}^*$ preserve the same meaning as before), thus perfectly reproducing Δ^* , and then the submatrix Δ in (1) also. In particular, by partitioning matrix \mathbf{X}^* as:

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{X}_1^* \\ (n \times r) \\ \mathbf{X}_2^* \\ (t \times r) \end{bmatrix}, \quad (2)$$

it is clearer that \mathbf{X}_1^* represents a set of coordinates for the n units such that: $\mathbf{D}(\mathbf{X}_1^*) = \Delta$, for which the Gram matrix $\mathbf{B}_1^* = \mathbf{X}_1^* \mathbf{X}_1^{*t}$ can be defined. We observe however that in general: $\mathbf{X}_1^{*t} \mathbf{1}_n \neq 0$, along with the fact that there do not exist, to our knowledge, explicit relations between the eigenvalues and eigenvectors of \mathbf{B}^* and \mathbf{B}_1^* , respectively. Alternatively, if $\text{rank}(\mathbf{B}^*) = r \leq n - 1$ we could work directly on Δ in (1) and find another $n \times r$ configuration \mathbf{X} such that $\mathbf{D}(\mathbf{X}) = \Delta$. Under these conditions, we would then have two different configurations, \mathbf{X}_1^* and \mathbf{X} , perfectly fitting Δ . In particular, it can be proved that such configurations fulfil the aforementioned fundamental result by Gower [4], (proof here omitted).

Now, turning back to the problem of the incomplete Δ^* with structure given in (1), if it is reasonable to assume that $\text{rank}(\mathbf{B}^*) = r \leq n - 1$, to recover the NA distances in Δ_{NA} we can then refer to the following procedure:

1. Set up a CMDS $n \times r$ configuration \mathbf{X} by working on the complete Δ only;
2. use Gower's method for adding t "new" units in the space generated by \mathbf{X} by involving known distances in matrix $\tilde{\Delta}$ between the n and t units. In such a way, a $t \times r$ configuration \mathbf{X}_{add} of projected coordinates into the space of \mathbf{X} is found;
3. the NA distances in Δ_{NA} can be estimated by computing Euclidean distances of the t units with the coordinates in \mathbf{X}_{add} . If in particular $\text{rank}(\mathbf{B}^*) < n - 1$, and the NA distances are *really* Euclidean, then it can be shown that the computed distances are exact, (proof here omitted).

3 A case study: Swiss data

The proposed procedure has been tested on various distance matrices, both Euclidean and non-Euclidean, with very different characteristics. In line with the present exposition, however, it is worth considering the simplest situation only, which concerns a distance matrix computed by the Euclidean distance formula. In this way, the artificially obtained NA distances are known to be really Euclidean. For the illustration, we have considered the dataset *Swiss*, freely distributed within the R environment [5]. It contains six socio-economic quantitative indicators, collected on the $N = 47$ French-speaking provinces of Switzerland in 1888. The distance matrix Δ^* , complete for now, is computed with the Euclidean formula after standardization of variables. The Gram matrix \mathbf{B}^* is p.s.d. with rank equal to $r = 6$, (i.e., the number of variables, as expected). The CMDS configuration \mathbf{X}^* in six dimensions then perfectly reproduce Δ^* . To simulate situations of strong incompleteness of Δ^* , we have set up the complete part Δ in two different ways: (1) $n = r + 1 = 7$ units, randomly drawn from the total set, so that: $t = 40$ units have reciprocally NA distances; (2) $n = r - 1 = 5$ units, randomly drawn, and $t = 42$ units with reciprocal NA distances. Then, summing up, the results obtained with the proposed procedure are in line with what expected, that is: (1) Δ_{NA} is perfectly recovered, in that the space generated by the 7×6 configuration \mathbf{X} is large enough; (2) Δ_{NA} is no more perfectly recovered. The obtained distance matrix is an approximation because the 5×4 configuration \mathbf{X} does not suffice to generate a space large enough.

Acknowledgements We thank the financial support of the project MIUR PRIN 2010-2011 MISURA - Multivariate models for risk assessment.

References

1. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling. Theory and Applications. 2nd edn. Springer, New York (2005)
2. Cox, T.F., Cox, M.A.A.: Multidimensional scaling. 2nd edn. Chapman & Hall/CRC, Boca Raton, FL (2001)
3. Gower, J.C.: Adding a Point to Vector Diagrams in Multivariate Analysis. *Biometrika*, **55**, 582–585 (1968)
4. Gower, J.C.: Euclidean distance geometry. *Mathematical Scientist*, **7**, 1-14 (1982)
5. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). <http://www.R-project.org>

Markov switching models for high-frequency time series: flapper skate's depth profile as a case study

Luigi Spezia and Cecilia Pinto

Abstract A Bayesian hierarchical model is proposed to analyse high-frequency time series. The model belongs to the class of Markov switching autoregressive models with autoregressive conditional heteroskedastic noises. Markov chain Monte Carlo inference is applied to estimate the unknown parameters of the model, which are all state-dependent. The model is used to analyse a time series of depths, obtained by a Data Storage Tag applied to a flapper skate

Introduction

Autoregressive processes can be extended to model non-linear and non-Normal time series by assuming that different autoregressions switch according to the state of an unobserved, or hidden, Markov chain. These models are known as Markov switching autoregressive models (MSARMs). The hidden Markov chain allows also modelling an eventual long memory of the observed process ([1]), making MSARMs a suitable tool in the analysis of high-frequency time series. MSARMs are widely studied and applied in econometrics, but are not common in other disciplines, such as environmental sciences and ecology.

Motivated by a research on animal movement ([2]), a new MSARM is proposed here to analyse a high-frequency time series. The series exhibit high autocorrelations at the higher lags, with a slow decay. This shape is due to both the non-linearity of the series (the series presents peaks which increase at a rate different from that they decrease) and the automatic recording of the data at a high frequency of time. Both issues are tackled by a MSARM in which the state-dependent noises are assumed of the

¹Luigi Spezia, Biomathematics & Statistics Scotland, Aberdeen, UK; luigi@bioss.ac.uk

²Cecilia Pinto, School of Biological Sciences, University of Aberdeen, Aberdeen, UK; r03cp11@abdn.ac.uk

autoregressive conditionally heteroskedastic (ARCH) type. Inference is developed under the Bayesian paradigm by Markov chain Monte Carlo (MCMC) algorithms.

The contribution of the paper is four-fold: (1) all parameters of the Markov switching ARCH (MSARCH) noises are state-dependent; (2) the identifiability constraints of any MSARCH noises are automatically satisfied by a suitable reparameterization of the coefficients; (3) the label switching problem is tackled by an automatic post-processing of the MCMC sample; (4) a better understanding of the flapper skate (*Dipturus intermedia*) behaviour is obtained, providing new information about one of the most precarious marine species at present.

The model

MSARMs are discrete-time stochastic processes $\{y_t; x_t\}$, so that $\{x_t\}$ is a latent, or hidden, finite-state Markov chain and $\{y_t\}$, given $\{x_t\}$, satisfies the order- p dependence and the contemporary dependence conditions: we have a sequence of observed random variables $\{y_t\}$ depending on the p previous observations, whose conditional distributions depend on $\{x_t\}$ only through the contemporary state of the Markov chain.

Let $\{x_t\}$ be a finite-state Markov chain on the state-space S_X whose cardinality is m . $\Gamma=[\gamma_{j,i}]$ is the $(m \times m)$ transition matrix, where $\gamma_{j,i} = P(x_t=j | x_{t-1}=i)$, with $0 < \gamma_{j,i} < 1$, for any $i, j \in S_X$, and any $t=2, \dots, T$; $\mathbf{x}^T=(x_1, \dots, x_T)'$ is the sequence of the states of the Markov chain and, for any $t=1, \dots, T$, x_t has values in S_X . Let $\mathbf{y}^T=(y_1, \dots, y_T)'$ be the sequence of observations; given the order- p dependence, the contemporary dependence, and the order- q ARCH conditions, the equations describing the MSARMs, when $x_t=i$, are:

$$y_t = \mu_i + \sum_{\tau=1}^p \varphi_{\tau(i)} y_{t-\tau} + e_t;$$

$$e_t = \sqrt{h_t} u_t, \quad u_t \sim N(0, 1);$$

$$h_t = \eta_i + \sum_{k=1}^q \alpha_{k(i)} e_{t-k}^2;$$

Any signal μ_i , any autoregressive coefficient $\varphi_{\tau(i)}$, for any $\tau = 1, \dots, p$, any ARCH coefficient η_i and $\alpha_{j(i)}$, for any $j = 1, \dots, q$, depend on the current state i of the Markov chain, for any $i \in S_X$. A sufficient condition for the stationarity of any state-dependent ARCH process, with $\eta_i > 0$ and $\alpha_{1(i)}, \dots, \alpha_{q(i)} \geq 0$, for any $i \in S_X$ is that the roots of the associated characteristic equations are all outwith the unit circle. To automatically satisfy this constraint, we reparametrize $\alpha_{1(i)}, \dots, \alpha_{q(i)}$ in terms of partial autocorrelations $\mathbf{r}_i=(r_{1(i)}, \dots, r_{q(i)})'$ of any sub-process $\{e_t^2 | x_t=i\}$, for any $i \in S_X$.

The unknown parameters of the MSARM are: $\boldsymbol{\mu}$ (the vector of the m signals μ_i), $\boldsymbol{\varphi}$ (the matrix of the $m \times q$ autoregressive coefficients $\varphi_{\tau(i)}$), $\boldsymbol{\eta}$ (the vector of the m ARCH intercepts η_i), \mathbf{R} (the matrix of the $m \times p$ Fisher-transformed partial autocorrelations $R_{q(i)}$), $\boldsymbol{\Gamma}$ (the matrix of the $m \times m$ transition probabilities $\gamma_{j,i}$). For our Bayesian inference, we place independent Dirichlet priors on each row of matrix $\boldsymbol{\Gamma}$; independent Normal priors on each entry of vector $\boldsymbol{\mu}$; independent Normal priors on each entry of matrix $\boldsymbol{\varphi}$; independent Gamma priors on each entry of vector $\boldsymbol{\eta}$; independent Normal priors on each entry of matrix \mathbf{R} . The joint distribution of all variables is

$$p(\mathbf{y}^T, \mathbf{x}^T, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \mathbf{R}, \mathbf{y}^0) = p(\mathbf{y}^T | \mathbf{x}^T, \boldsymbol{\mu}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \mathbf{R}, \mathbf{y}^0) \times p(\mathbf{x}^T | \boldsymbol{\Gamma}) p(\boldsymbol{\Gamma}) p(\boldsymbol{\mu}) p(\boldsymbol{\varphi}) p(\boldsymbol{\eta}) p(\mathbf{R}),$$

where $\mathbf{y}^0=(y_{-s+1}, \dots, y_0)'$, with $s=\max\{p; q\}$, are the initial fixed values.

Study species, study area, and tagging methodology

The study area chosen for the data collection was the Sound of Jura, on the West Coast of Scotland. The Sound of Jura has been identified as a potential Marine Protected Area (MPA) as being part of the West MPA region identified by the OSPAR convention (www.ospar.org), which identifies priority areas aimed at the conservation of the Scottish population of *Dipturus intermedia* (flapper skate). The flapper skate is listed in the International Union for Conservation of Nature Red List of Threatened Species as critically endangered (www.iucnredlist.org) due to the strong decline (90%) that catches and landings reported for this species in the last 30 years.

In the Sound of Jura eight individuals of flapper skate were tagged with Data Storage Tags (DST) during October-November 2011 and nine additional individuals were tagged during October-November 2012. The tags were applied by scientists of the Marine Scotland Laboratory following the Home Office regulations. DSTs collect pressure levels every two minutes and sea temperatures every ten minutes. Pressure data are then converted in depth values. In order to obtain the data back it is necessary to catch the tagged individual and get hold of the tag. Four skates have been successively caught, and so four time series are now available for our analysis.

Skate movement analysis

We started our analysis by considering the longest of the four available series (272,440 observations). At the current early stage of our study, we are analysing some shorter representative sub-series, in order to understand, in a reasonable amount of computing time, the main features of the observed process, describing the skate behaviour. The results from a sub-series of 2,000 points of the natural logarithm of the absolute values of the depths are proposed here.

Model comparison suggested the best model is that with two hidden states ($m=2$), autoregressions of the second order ($p=2$), and ARCH noises of the first order ($q=1$). The two hidden states (Figure 1) have the following interpretation: in state 1 (low variability) we can observe the skate whilst is resting, moving horizontally, slowly ascending and descending; in state 2 (high variability) the skate is swimming fast up and down, ascending and descending quickly. The fitting ability of the model is very good as shown by the graph in Figure 2.

Concluding remarks

The development of MSARMs with MSARCH noises sounds promising in analysing high-frequency time series recorded by DSTs. It will be informative to analyse the

whole time series, by assuming a non-homogeneous hidden Markov chain, in which the transition probabilities are time-varying, depending on the dynamics of some covariates, such as daylight duration, lunar cycles (a proxy for tidal cycles), daily rainfalls and sea surface water temperatures.

Figure 1: Observations classified in states 1 (skate resting, moving horizontally, slowly ascending and descending) and 2 (skate swimming fast up and down, ascending and descending quickly)

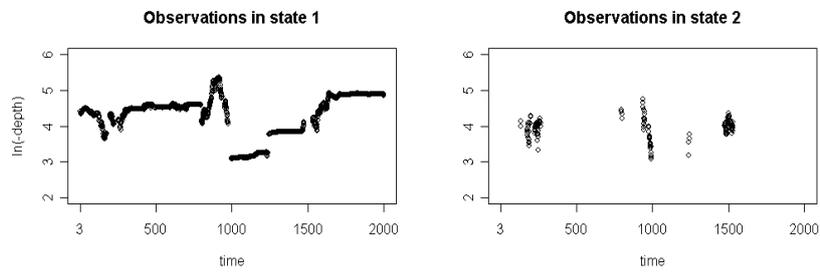
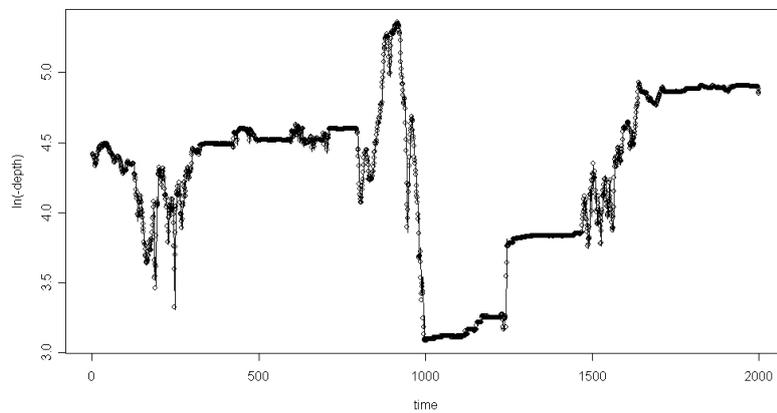


Figure 2: Actual (dots) and fitted values (continuous line) of the series



References

1. Diebolt, F.X., Inoue, A.: Long memory and regime switching. *J. of Econ.* 105, 131-159 (2001).
2. Pinto, C.: Estimating the probability of recolonization of endangered marine species integrating demographic and movement parameters. PhD Thesis, University of Aberdeen, in preparation.

Data Privacy in Credit Scoring: Evaluating SVM Approaches Based on Microaggregated Data

Ralf Stecking and Klaus B. Schebesch

Abstract In recent time privacy concerns affect many real world data classification tasks. For example, credit scoring applications usually have to deal with such data security limitations. Subsequently, privacy preserving data mining methods have become of growing interest. In our paper we propose a microaggregation procedure in order to anonymize sensitive credit client information. Moreover, we examine the performance of support vector machines (SVM) on anonymized data. SVM are powerful and robust machine learning methods, that have been proven to be superior in many credit scoring classification tasks with original data. We show, how to partition the original credit scoring data set and how to construct appropriate cluster representatives, that can be used to replace the original values in the data set. Finally, different SVM classification models with anonymized data input are evaluated and compared to models that are trained on the original data.

1 Introduction

According to [5] statistical procedures in order to preserve data privacy may follow two directions. The first one is exchanging data between different parties without disclosing private information to each others. Such shared data may be vertically, horizontally or arbitrarily partitioned, depending on whether the parties collect different information about the same set of objects, the same information about different objects or a combination of both [9]. Secondly, if the data model building process should be completely externalized, the whole data set itself must be

Ralf Stecking

Department of Economics, Carl von Ossietzky University Oldenburg, D-26111 Oldenburg, Germany, e-mail: ralf.w.stecking@uni-oldenburg.de

Klaus B. Schebesch

Faculty of Economics, Vasile Goldiș Western University Arad, Romania, e-mail: kbschebesch@uvvg.ro

anonymized. This can be accomplished in various ways: perturbing the data by adding random noise [2], generalising data by suppressing detailed information of identifier or quasi-identifier attributes [4] or clustering the data and replacing individual object information with condensed cluster representations [6]. In our work we divide the whole data set into *regions* or *clusters* and replace individual data with aggregated information. *Microaggregation procedures* ensure k -anonymity, if every partition contains at least k objects [7]. Moreover, our approach is not restricted to a special classification method like e.g. [5]. We introduce symbolic descriptions for microaggregated data, which allow for coding categorical, quantitative or mixed variable types, most common in credit scoring problems and we present an appropriate distance function for this type of data. The outline of this paper is as follows: In section 2 we describe microaggregation steps that are used for our data anonymization procedure. In section 3 our empirical results are presented. We compare the out-of-sample classification results of two different SVM for two different credit scoring data sets using original and anonymized data.

2 Microaggregation

Microaggregation usually proceeds in two steps: (i) a *partition step* will cluster the original data into groups with at least k similar records, and (ii) an *aggregation step* will represent each cluster with a prototype that is used to replace the original record. Therefore, microaggregation requires a *clustering method* and an *aggregation procedure* [8]. The *partition step* proceeds as follows: The first data set is divided into “good” and “bad” credit clients. Subsequently, unsupervised clustering is used, partitioning the data set into numbers of clusters from “good” and “bad” classes respectively, while preserving the class labels. In order to ensure a minimum of similar records in each cluster, cluster representatives with less than k records are omitted from the cluster solution and a classification step with the now reduced number of cluster representatives is performed. For the second data set, regional information are available. Each subregion consists of at least k credit client records, that are replaced by its representative. According to the *aggregation step*, what kind of information should be given to the classification method? We use representation techniques from symbolic data analysis, where a special data description is needed to represent the variable outcomes of a symbolic object. A complete overview of symbolic variable types can be found in [1]. In the present work the partitions serve as symbolic objects and are described by *modal variables* where categories or intervals appear with a given probability. Technically, the *aggregation step* is organized in the following way: For each partition the relative frequencies of all outcomes per *categorical variable* are recorded. *Quantitative variables* are divided into four equally sized intervals, with quartiles of the full variable range as interval borders. Relative frequencies for these intervals are also recorded, resulting in an input vector consisting of values between zero and one. *Distances* between symbolic coded

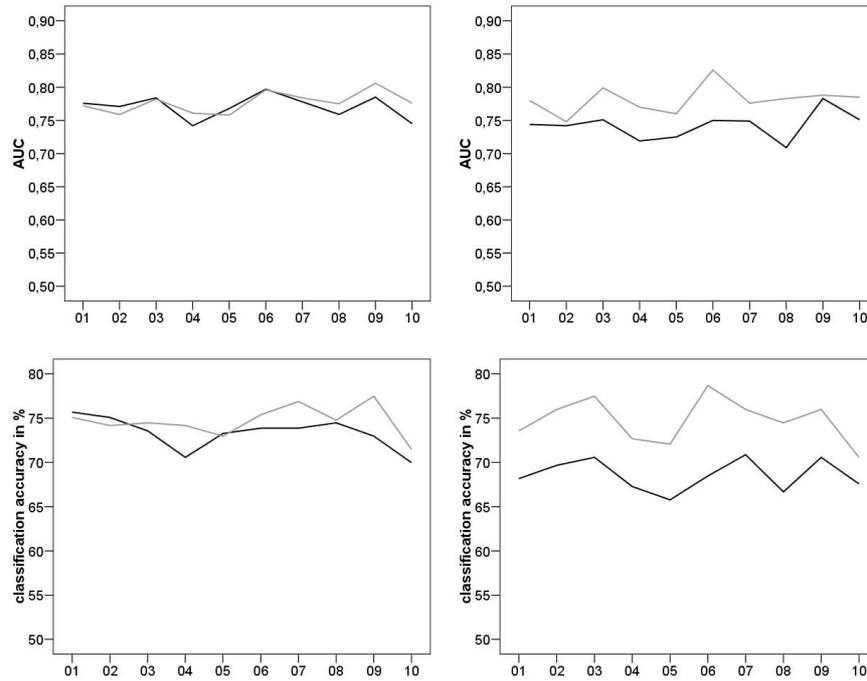


Fig. 1 *German Credit Data*: AUC (upper row) and classification accuracy (lower row) for RBF SVM (left) and Linear SVM (right) trained on anonymized (black line) and original data (grey line) with tenfold out of sample validation using original data ($N_V = 333$).

partitions are computed by an appropriate derivation of the Euclidian distance function [1].

3 Empirical results

There are two credit scoring data sets used in this paper: the *German credit data set* is publicly available at the UCI repository at <http://archive.ics.uci.edu/ml> [3]. This data set consists of information about 1000 credit clients. Initially there are 20 input variables per client, which are separated into 700 “good” and 300 “bad” cases. The second data set used consists of information about 139951 clients for a *building and loan credit* with twelve input variables per client and a default rate of 2.6% (3692 out of 139951). Both data sets are randomly divided into ten training and validation sets in a 2:1 relation. The training sets are anonymized through microaggregation as lined out in section 2. We built 80 SVM models, half of them were trained on the original data, the others on microaggregated data. Figures 1 and 2 show classification results for each data set and two different SVM model variations (Linear and

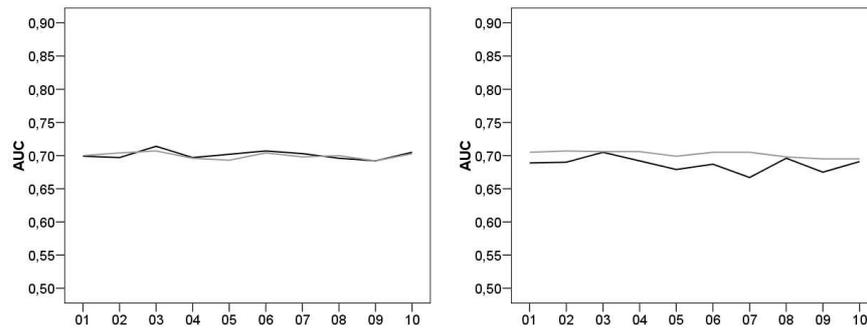


Fig. 2 *Building and Loan Credit Data*: Area under curve (AUC) for RBF SVM (left) and Linear SVM (right) trained on anonymized (black line) and original data (grey line) with tenfold out of sample validation using original data ($N_V = 46650$).

RBF), comparing area under curve (AUC) statistics (both figures) and classification accuracy (just figure 1). Models are evaluated with ten times cross validation using original data on the holdout set. Both figures show, that the classification performance of models trained on microaggregated data is quite comparable to these that were trained on the original data, revealing some more obvious deviations for Linear models (right side), whereas SVM with RBF kernel (left side) are not affected considerably by data anonymization.

References

1. Billard L, Diday E (2006) *Symbolic Data Analysis*. Wiley, New York.
2. Brand R (2002) Microdata protection through noise addition. In: Domingo-Ferrer J (ed) *Inference Control in Statistical Databases. From Theory to Practice. Lecture Notes in Computer Science*, vol. 2316, Springer:97-116.
3. Frank A, Asuncion A (2010) UCI Machine Learning Repository.
4. Inan A, Kantarcioglu M, Bertino E (2009) Using Anonymized Data for Classification. In: Ioannidis Y E, Lee D L, Ng R T (eds) *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, Shanghai, China*:429-440.
5. Lin K-P, Chen M-S (2010) Privacy-preserving outsourcing support vector machines with random transformation. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*. ACM, New York, NY, USA:363-372.
6. Marés J, Torra V (2012) Clustering-Based Categorical Data Protection. In: Domingo-Ferrer J, Tinnirello I (eds) *Privacy in Statistical Databases. Lecture Notes in Computer Science*, vol. 7556, Springer:78-89.
7. Navarro-Arribas G, Torra V (2012) Information fusion in data privacy: A survey. *Information Fusion* 13:235-244.
8. Torra V (2004) Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer J, Torra V (eds) *Privacy in Statistical Databases. Lecture Notes in Computer Science*, vol. 3050, Springer:162-174.
9. Vaidya J, Yu H, Jiang X (2008) Privacy-preserving SVM classification. *Knowledge and Information Systems* 14 (2):161-178.

Analyzing university students' careers using Multi-State Models

Isabella Sulis, Francesca Giambona and Nicola Tedesco

Abstract The methodologies adopted in the last decades to analyze students' university careers using cohort studies mainly focus on the risk to observe one of the possible competing status, specifically dropout or graduation, after several years of follow-up. In this perspective all the other event types that may prevent the occurrence of the target event are treated as censored observations. A broader analysis of students' university careers from undergraduate to postgraduate status reveals that several competing and non competing events may occur, some of which can be denoted as absorbing while others as intermediate. In this study we propose to use Multi-State Models to analyze students' careers. This class of models allows us to take into account: i) the sequence of events experienced by students during their careers (first level degree, dropout, second level degree and others postgraduate studies etc.); ii) how the risk to experience the different states varies along the time; iii) the paths of transition between intermediate events.

Key words: Cohort studies, Multi-States models, Risk factors, Students' careers, performance indicators

1 Introduction

The Italian system of higher education has been widely criticized for its ineffectiveness; it is characterized by lower university graduation rates comparing to the other European countries. In the last ten years the number of studies which pay attention to the inefficiency of the Italian university system has considerably increased, with the main aims to: i) highlight outstanding institutions in term of the effectiveness of their courses/ degree programmes /faculties; ii) to advance indicators related to the regularity of students' careers (e.g. dropout rates, longer time of staying in the

Isabella Sulis, Francesca Giambona, and Nicola Tedesco
Università degli Studi di Cagliari, e-mail: isulis@unica.it, francesca.giambona@unica.it

university system before graduating and the delay in the accumulation of formative credits); ii) to advance suitable methodologies to make comparative evaluations across institutions [2] [1].

The peculiarities of the Italian University system have arisen the interest of many social researches on the analysis of the main determinants of students' careers. The researches carried out in the last decades mainly focus on two main directions: i) the refinement of suitable methodological approaches to identify risk factors of non regularities of university students' careers; ii) the following up of students' careers using cohort studies. The main aim of this work is to explore the potential of Multi-State Models as methodological approach for analyzing university students' carriers in their overall complexity.

2 Data

Longitudinal cohort data on students' careers have been provided from the administrative archive of the University of Cagliari. For the sake of this application we consider just the 4336 students who enrolled for the first time in the 2006/07 academic year (a.y) in the first level degree programmes. The last information on students' status recorded in the data set refers to March 2013. After 7 academic years the possible states in which students can be observed are: formally dropout during the first level degree programme (S_1); graduated in the first level degree programme (S_2); implicitly dropout during the first level degree programme (S_3); enrolled at the second level degree programme (S_4); graduated in the second level degree programme (S_5); formally dropout during the second level degree programme (S_6); enrolled at one of the postgraduate programmes (S_7). We define implicitly dropouts those students who have not being payed their academic fees by more than two academic years (last year of observation in the archive 2010-11 a.y). Thus, implicitly dropouts can be observed just in the first level degree programmes. The possible careers' paths are depicted in Figure 1. We have restricted the analysis to the main states observable in students' careers by eliminating from the analysis all events which rarely occur (e.g. enrollment to a single exam after S_2 or S_3 , enrollment to the first level master degree after the first level graduation, formally dropout more than one time during the first level degree program). The data set contains the information on the exact time (days/month/year) on which the students enter in any of the states experienced during their careers. In the case of implicitly dropout, the expected time at which the event occurred has been imputed equal to the 30 of October of the year after the last one in which the student payed the academic fees. Two kinds of censored units are observable in the analysis of students' careers: (i) students who did not experience any transition in at least one of the listed states ($S_1 : S_7$) and are still enrolled at the first-level degree programme (First-Type censored units); ii) students who after the enrolment at one of the second level degree programmes did not experience any other transition to the states linked to S_4 ($S_5 : S_7$) and are still enrolled (Second-Type censored units). The information about students, degree programmes/

faculties available in the archive have been used to single-out profiles of students or degree programmes more at risk to experience specific paths. Specifically the following covariates have been adopted to analyze the risk of transition from a state to another: i) students details - sex, age, residence -; ii) students' curricula - type of secondary school attended, delay in school graduation (it has been expressed in years from 19), delay in enrollment at the university (it has been expressed in years from the school graduation), school final examination mark -; iii) characteristics of the degree programmes or faculties.

Standard survival analysis methodologies have been widely adopted in the analysis of students' careers. However, classical approaches drastically simplify the complex structure of university students' careers i) bounding the analysis to a single level of the university studies (first level/second level/ postgraduate) and ii) mainly focusing on the risk to observe one of the possible competing states (dropout/ graduation) during the follow-up and treating all the other event types that may prevent the occurrence of the target event as censored observations. A broader analysis of students' university careers from undergraduate to postgraduate reveals that in the time span from the first enrolment of a student to the university to his definitive sort-out several competing and non competing events may occur, some of which can be denoted as absorbing events (since are reachable just after that students have experienced some previous one) while others as intermediate. Multi-state models are systems of multivariate survival equations which allow us to assess the risk of experiencing several types of competing and non competing events and to move through a series of concatenate states following certain paths of possible transitions; furthermore, they allow researchers of jointly dealing with several absorbing (end-points) events, several intermediate events and several types of censored units [3][4].

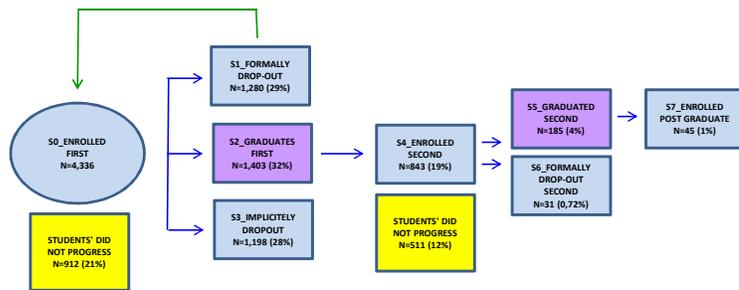


Fig. 1 Path diagram of students' careers

2.1 Methodological approach to analyze students' careers

To simplify the diagram of all possible paths (see Figure 1¹), in this first explorative analysis of the data a multi-state model without recurrent events has been adopted, allowing students to move just in one direction. Thus, we consider states S_1 , S_3 , S_6 and S_7 as absorbing and S_2 , S_4 and S_5 as intermediate. In this way a restriction has been advanced on the possible events that a unit can experience². To sum up a model with 7 transitions has been set up: 1) from enrolled to the first level degree programme to formally dropout $S_o \rightarrow S_1$; 2) from enrolled to the first level degree programme to implicitly dropout $S_o \rightarrow S_3$; 3) from enrolled to the first level degree programme to graduated $S_o \rightarrow S_2$; 4) from graduated in the first level degree programme to enrolled at the second level degree $S_2 \rightarrow S_4$; 5) from enrolled to the second level degree programme to graduated $S_4 \rightarrow S_5$; 6) from enrolled to the second level degree programme to formally dropout $S_4 \rightarrow S_6$; 7) from graduated to the second level degree programme to postgraduate studies (Master or PhD) $S_5 \rightarrow S_7$.

By denoting with T the time of reaching state j from state i , the hazard rate of the transition $i \rightarrow j$ is defined by $\lambda_{ij}(t) = \lim_{\Delta t \downarrow 0} \frac{Prob(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$. The risk of transition across the states of students' careers can be described by a set of 7 hazard ratios (HRs), which vary as function of students/degree programmes covariates (Z). In this first explorative analysis we adopted a Cox's proportional-hazards model:

$$\lambda_{ij}(t|Z) = \lambda_{ij,0}(t) \exp(\beta_{ij}^T Z)$$

where $\lambda_{ij,0}(t)$ is the baseline hazard of transition $i \rightarrow j$ for students with baseline covariates pattern. For details on how to relax the proportional-hazards assumption and on the aspects that play a determinant role in specifying more sophisticated multi-states models we remained to Putter et al. (2007) and to de Wreede et al. (2011) [4][3].

References

1. Enea, M., Attanasio, M.: Bivariate logistic models for the analysis of the Students University Success. Proceedings of the XLVI Scientific Meeting of the Italian Statistical Society, Sapienza University of Rome, 20-22 June (2012)
2. Attanasio, M., Boscaino, G., Capursi, V., Plaia, A.: Indicators and measures for the assessment of university students careers. Proceedings of the 8th Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society. University of Pavia, September 7-9, 1-4 (2011).
3. de Wreede, L. C., Fiocco, M., Putter, H.: mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software, 38(7), 1-30 (2011).
4. Putter, H., Fiocco, M., Geskus, R. B.. Tutorial in Biostatistics: Competing Risks and Multi-State Models. Statistics in Medicine, 26, 2389-2340, (2007).

¹ In the path is allowed to students enrolled to first degree to dropout and re-enrolled. For this reason the sum of students in S_1, S_2 and S_3 is not equal to the number of students in S_0 .

² This hypothesis can be relaxed in further analysis.

BBRecap for Bayesian Behavioural Capture-Recapture Modeling

Luca Tardella and Danilo Alunni Fegatelli

Abstract Motivated by the lack of a comprehensive software tool for Bayesian inference of capture-recapture data we introduce an R package which helps inferring on a very general class of capture-recapture models with behavioural response to capture. A brief explanation of the behavioural model building rationale and an example of alternative model fitting of *Great Copper Butterfly* data is detailed.

Key words: Capture-recapture, Bayesian inference, behavioural models, likelihood failure

1 Introduction

In the context of capture-recapture analyses, there are many software tools developed in different statistical environments and languages. However, there is no statistical environment or package especially dedicated to the Bayesian inference in this context. For a general recent overview on softwares for capture-recapture analysis see [3]. BBRecap is an R package which offers a core of functions useful for fitting a large collection of behavioural capture-recapture models within the Bayesian approach. Inference on the population size can be also carried out in terms of unconditional likelihood. However the user must be aware that the Bayesian analysis systematically yields an inferential improvement of point and interval estimates as shown in [2]. In fact a likelihood failure pathology consisting of infinite estimates of the population size for both conditional and unconditional maximum likelihood approaches has been pointed out and characterized for a large class of models where the probability of never being captured during the trapping stages depends only on

Luca Tardella
Sapienza Università di Roma, e-mail: luca.tardella@uniroma1.it

Danilo Alunni Fegatelli
Sapienza Università di Roma e-mail: danilo.alunnifegatelli@uniroma1.it

one parameter . The Bayesian approach totally overcomes the failure problem and this makes the availability of Bayesian software tools in this context extremely urgent.

2 General capture-recapture behavioural model framework

Consider a discrete-time capture-recapture experiment with t capture occasions. Let N be the population size which is the most important parameter of interest in closed population models. Data are expressed as a $N \times t$ binary matrix $\mathbf{X} = [x_{i,j}]$ where $x_{i,j} = 1$ if the i -th unit is captured in the j -th occasion and $x_{i,j} = 0$ otherwise. Assume that we have captured at least once M different units. These units are conventionally labelled from 1 to M and those $N - M$ never captured are labelled from $M + 1$ to N . Hence, the underlying binary matrix \mathbf{X} can be partitioned as follows $\mathbf{X}^T = [\mathbf{X}_{obs}^T, \mathbf{X}_{mis}^T]$ where \mathbf{X}_{obs} is an $M \times t$ binary matrix representing the observed data and \mathbf{X}_{mis} is an $(N - M) \times t$ matrix of unobserved zeros. The factorization of the joint probability of the binary representation of the individual complete capture history $(x_{i,1}, \dots, x_{i,t})$ is the basis for a natural understanding and modelling of the behavioural effect to capture. In fact, using the sequential product of conditional probabilities $p(\mathbf{h}) = Pr\{X_{i,l_h+1} = 1 | \mathbf{h}\}$ one has an easier probabilistic interpretation of the change of behaviour due to a particular previous partial capture history $\mathbf{h} = (x_{i,1}, \dots, x_{i,l_h})$.

We refer to [4] as a recent attempt to formalize this general model framework in terms of a saturated reparameterization consisting of all $2^t - 1$ probabilities conditioned on each possible partial capture history. Parsimonious nested models can be either specified through equality constraints on the vector of all conditional probabilities. In [2] the same reparameterization is alternatively represented in terms of partitions of the set of all partial capture histories $H = \{(), (0), (1), (00), (10), (01), (11), \dots\} = \cup_{j=0}^{t-1} \{0, 1\}^j$ in equivalence classes so that if one denotes with \mathcal{H}_B one of the possible partitions of H in B disjoint subsets $\mathcal{H}_B = \{H_1, \dots, H_b, \dots, H_B\}$ the corresponding parameter vector of probabilities representing the nested model associated to \mathcal{H}_B is denoted with $\mathbf{p}_{\mathcal{H}_B} = (p_{H_1}, \dots, p_{H_B})$ and it is such that

$$\forall \mathbf{h}, \mathbf{h}' \in H_b \Rightarrow p(\mathbf{h}) = p(\mathbf{h}') = p_{H_b} \quad \forall b = 1, \dots, B$$

Once specified the partition \mathcal{H}_B , under individual independence assumption, the likelihood function can be factorized as follows

$$L(N, \mathbf{p}_{\mathcal{H}_B}) \propto \binom{N}{M} \prod_{b=1}^B p_{H_b}^{n_{(H_b,1)}} (1 - p_{H_b})^{n_{(H_b,0)} + m_{(H_b,0)}}$$

where $n_{(H_b,0)} = \sum_{i=1}^n \sum_{\mathbf{h} \in H_b} I[(x_{i1}, \dots, x_{il_h}) = \mathbf{h}, x_{i(l_h+1)} = 0]$, $n_{(H_b,1)} = \sum_{i=1}^n \sum_{\mathbf{h} \in H_b} I[(x_{i1}, \dots, x_{il_h}) = \mathbf{h}, x_{i(l_h+1)} = 1]$, $m_{(H_b,0)} = \sum_{i=M+1}^N \sum_{\mathbf{h} \in H_b} I[(x_{i1}, \dots, x_{il_h}) = \mathbf{h}]$. In this way the factorized likelihood allows for an easy conjugate Bayesian analysis as

proposed in [2] and the `BBRecap` package allows for an easy end-user implementation as detailed in the next section.

3 Example: Great Copper data

We show how one can perform alternative model fitting of Great Copper data originally studied in [5] and reviewed in [4] and [2]. In the study a total of 45 different butterflies are observed during $t = 8$ capture occasions. Consider the classical model M_b introduced in [6] which is the simplest behavioural model where the capture probabilities vary only once when first capture occurs hence representing an enduring effect to capture.

```
> library(BBRecap)
> data(greatcopper)
> mod.Mb=BBRecap(greatcopper,mod="Mb")
> str(mod.Mb)
List of 5
 $ Model          : chr "mod.Mb"
 $ prior.N        : chr "Rissanen"
 $ N.hat.RMSE     : num 63
 $ HPD.interval   : num [1:2] 46 134
 $ log.marginal.lik : num -176
```

From the structure of the returned list object one can easily get the most relevant posterior output for point and interval estimates and model evaluation. If we want to get a better fit we could consider as an alternative ephemeral behavioural model such as a second order Markov model extending [7]

```
> mod.Mc2=BBRecap(greatcopper,mod="Mc",markov.ord=2)
> str(mod.Mc2)
List of 5
 $ Model          : chr "mod.Mc2"
 $ prior.N        : chr "Rissanen"
 $ N.hat.RMSE     : num 117
 $ HPD.interval   : num [1:2] 59 368
 $ log.marginal.lik : num -169
```

The behavioral model building can be customized in terms of specifying an alternative meaningful partition of the partial capture histories which corresponds to the collection of equivalence classes for which one has the same conditional probability of being captured. The construction of the partition can be sometimes driven by the definition of a suitable quantification $q(\mathbf{h})$ of the partial capture history in terms of suitable intervals of quantification as illustrated in [1]. For instance one can define $q(\mathbf{h})$ as the number of captures in \mathbf{h} . Here we limit ourselves to illustrate the implementation of the former idea in terms of the quantification qualified in [1] as *memory effect* which has been coded in the function `mem.eff.quantify.ch(...)` and which has been shown to be related to variable order Markovian models. In this

way one can recover the same model M_{L_2} originally proposed in [4] as a particular instance of the general equivalence classes approach as follows

```
> ML2.part=partition.ch(quantify.ch.fun="mem.eff.quant",t=8,
+ breaks=c(0,0.625,1),include.lowest=T)
> mod.ML2=BBRecap.custom.part(data=greatcopper,
+ partition=partition.ML2)
> str(mod.ML2)
List of 7
 $ Model          : chr "custom.partition"
 $ quantify.ch.fun : chr "mem.eff.quant"
 $ breaks         : num [1:3] 0 0.625 1
 $ prior.N       : chr "Rissanen"
 $ N.hat.RMSE    : num 84
 $ HPD.interval  : int [1:2] 58 133
 $ log.marginal.lik : num -167
```

`ML2.part` contains the partition of H induced by the memory effect quantification `mem.eff.quant` and the two disjoint intervals $[0, 0.625]$, $(0.625, 1]$ and it is provided as customized argument `partition` in the `BBRecap.custom.part`. More examples, functionalities and options are detailed in the vignette included in the package. Among them we mention the ability of obtaining unconditional maximum likelihood estimation and generalized linear model extensions which exploit the quantification of partial capture histories as time-dependent covariates.

References

1. Alunni Fegatelli, D. (2013) , New methods for capture-recapture modelling with behavioural response and individual heterogeneity.
2. Alunni Fegatelli, D., Tardella, L. (2013) , Improved inference on capture recapture models with behavioural effects, *Statistical Methods & Applications* **22**(1), 45–66.
3. Bunge, J. A. (2013) , A survey of software for fitting capturerecapture models, *Wiley Interdisciplinary Reviews: Computational Statistics* **5**(2), 114–120.
4. Farcomeni, A. (2011) , Recapture models under equality constraints for the conditional capture probabilities, *Biometrika* .
5. Ramsey, F., Severns, P. (2010) , Persistence models for mark-recapture, *Environmental and Ecological Statistics* **17**, 97–109.
6. White, G. C., Anderson, D. R., Burnham, K. P., Otis, D. L. (1982) , *Capture-recapture and Removal Methods for Sampling Closed Populations*, Los Alamos National Laboratory.
7. Yang, H.-C., Chao, A. (2005), Modeling animals' behavioral response by Markov chain models for capture-recapture experiments, *Biometrics* **61**(4), 1010–1017.

Mixtures of generalized hyperbolic factor analyzers

Cristina Tortora, Paul D. McNicholas and Ryan P. Browne

Abstract Model-based clustering assumes that the population is a convex combination of a finite number of densities, the distribution is a basic assumption of the model. Among all the possible distributions, the generalized hyperbolic distribution has the advantage to be a generalization of several other methods; the Gaussian distribution, skewed t-distribution, etc.. With specific parameters, it can represent a symmetric or a skewed distribution. Thanks to its flexibility it can be a valid instrument; however, it requires the estimation of a wide number of parameters that increases as the number of variables increases. The aim of this work is to propose a mixture of generalized hyperbolic factor analyzers to extend the method to high dimensional data. Mixtures of factor analyzers look for a latent, lower dimensional, subspace for each cluster, where the number of parameters to be estimated is much lower.

Key words: model-based clustering, subspace clustering, mixture of factor analyzers, generalized hyperbolic distribution.

Cristina Tortora
University of Guelph, Department of Mathematics & Statistics, Guelph, Ontario, Canada, e-mail:
ctortora@uoguelph.ca

Paul D. McNicholas
University of Guelph, Department of Mathematics & Statistics, Guelph, Ontario, Canada, e-mail:
pmcnicho@uoguelph.ca

Ryan P. Browne
University of Guelph, Department of Mathematics & Statistics, Guelph, Ontario, Canada, e-mail:
rbrowne@uoguelph.ca

1 Introduction

Cluster analysis aims at finding homogeneous group of units, according to a geometrical or probabilistic criterion. With high dimensional data sets classical clustering methods are usually unsatisfactory. Distance-based and model-based methods' performance can deteriorate as the number of variables increases. This work focuses the attention on model-based clustering, and, specifically, on the mixture of generalized hyperbolic factor analyzers. The generalized hyperbolic distribution has the advantage to be a general distribution that, with specific values of the parameters, can lead to other well known distribution like the Gaussian distribution. Moreover, it can detect clusters with non-elliptical form, because it contains a skewness parameter. However, the generalized hyperbolic distribution requires the estimation of a great number of parameters and the number of parameters increases as the number of variables increases. To extend the use of this distribution to high-dimensional data, we propose the mixture of generalized hyperbolic factor analyzers. The method uses the factorial analysis to project data in local low-dimensional subspaces, where the number of parameters to be estimated is much lower.

2 Mixture of generalized hyperbolic distributions

Model-based clustering considers the overall population as a mixture of groups and each component of this mixture is modelled through its conditional probability distribution [4]. In this context, the observations x_1, \dots, x_n are assumed to be independent realizations of a random vector \mathbf{X} . The matrix of labels \mathbf{z} , of generic element z_{ik} with $i = 1, \dots, n$, and $k = 1, \dots, K$, indicates cluster membership, and the unobserved labels z_{1k}, \dots, z_{nk} are assumed to be independent realizations of a random variable $z_{ik} = 1$ if $x_{ik} \in$ group k , $z_{ik} = 0$ otherwise. Let us define with g the probabilistic density function of \mathbf{X} , the finite mixture model can be defined as follow:

$$g(x, \theta) = \sum_{k=1}^K \pi_k f(x, \theta_k)$$

where π_k , such that $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \in (0, 1]$, indicates the mixture proportion, f the conditional density function and θ_k the parameter vector for the k^{th} mixture component. Commonly, the density functions are multivariate Gaussian. To adapt the model to a wider range of situations other distributions were considered, e.g. t-distribution or skew-normal distribution. Among them [1] propose to use the generalized hyperbolic (GH) distribution. It has the advantage to be the generalization of many special cases: normal inverse Gaussian distribution, Student t-distribution, variance-gamma distribution, Laplace distribution, hyperbolic distribution, and the multivariate Gaussian distribution [7]. The density of a generalized hyperbolic distribution can be expressed as in the following formula:

$$f(\mathbf{x}, \boldsymbol{\vartheta}) = \left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Delta})}{\psi + \boldsymbol{\alpha}' \boldsymbol{\Delta}^{-1} \boldsymbol{\alpha}} \right]^{\frac{\lambda - p}{2}} \frac{\left(\frac{\psi}{\chi} \right)^{\frac{\lambda}{2}} K_{\lambda - \frac{p}{2}} \left(\sqrt{[\psi + \boldsymbol{\alpha}' \boldsymbol{\Delta}^{-1} \boldsymbol{\alpha}][\chi + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Delta})]} \right)}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Delta}|^{\frac{1}{2}} K_{\lambda}(\sqrt{\chi\psi}) \exp\{(\boldsymbol{\mu} - \mathbf{x})' \boldsymbol{\Delta}^{-1} \boldsymbol{\alpha}\}},$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Delta}) = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Delta}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}$, and $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\alpha})$ is the vector of parameters. To ensure the identifiability, the constraint $|\boldsymbol{\Delta}| = 1$ was added. Given a random variable $Y \sim GIG(\psi, \chi, \lambda)$, where GIG indicates the generalized inverse Gaussian distribution, and $\mathbf{U} \sim N(\mathbf{0}, \boldsymbol{\Delta})$, a generalized hyperbolic random variable \mathbf{X} can be generated as follow:

$$\mathbf{X} = \boldsymbol{\mu} + Y\boldsymbol{\alpha} + \sqrt{Y}\mathbf{U}, \quad (1)$$

it follows that $\mathbf{X} | Y \sim N(\boldsymbol{\mu} + y\boldsymbol{\alpha}, y\boldsymbol{\Delta})$.

Dealing with mixture of generalized hyperbolic distribution the unknown parameters are: the latent variable Y , the membership label z_{ik} , and the vectors $\boldsymbol{\vartheta}_k = (\lambda_k, \chi_k, \psi_k, \boldsymbol{\mu}_k, \boldsymbol{\Delta}_k, \boldsymbol{\alpha}_k)$, with $i = 1, \dots, n$ and $k = 1, \dots, K$.

The expectation maximization (EM) algorithm can be used for the parameter estimation [2]. For detail refers to [1].

3 Mixtures of generalized hyperbolic factor analyzers

The estimation of mixture of generalized hyperbolic distributions requires an high number of parameters; the number increases as the number of variables increases. To use the model dealing with high number of variables different strategies can be applied; one suitable strategy can be the use of mixture of factor analyzers [3, 5]. Mixture of factor analyzers provides a local dimensionality reduction, it assumes that the observation \mathbf{X} , given k , can be expressed using the following model:

$$\mathbf{X} = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \mathbf{W} + \mathbf{e}_k, \text{ with probability } \pi_k,$$

with $k = 1, \dots, K$. The factors are independently distributed $\boldsymbol{\Lambda}_k \sim N(\mathbf{0}, \mathbf{I}_q)$ with $q \ll p$, independently from $\mathbf{e}_k \sim N(\mathbf{0}, \boldsymbol{\Psi}_k)$ with $\boldsymbol{\Psi}_k$ diagonal matrix. It follows that the marginal distribution of \mathbf{X} is $N(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k' \boldsymbol{\Lambda}_k + \boldsymbol{\Psi}_k)$.

Dealing with mixture of generalized hyperbolic distributions, the K vectors of parameters are: $\boldsymbol{\vartheta}_k = (\lambda_k, \chi_k, \psi_k, \boldsymbol{\mu}_k, \boldsymbol{\Delta}_k, \boldsymbol{\alpha}_k)$. The number of variables affects the dimension of $\boldsymbol{\Delta}_k$, to reduce the number of parameters the dimensionality of each cluster can be locally reduced by applying factorial transformation of variables for each cluster. This leads to the construction of a factor space for each cluster.

In detail, considering a single-factor analysis model, the matrix \mathbf{U} in (1), can be decomposed as: $\mathbf{U} = \boldsymbol{\Lambda} \mathbf{W} + \boldsymbol{\varepsilon}$, where $\mathbf{W} \sim N(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$. The covariance matrix become $y\boldsymbol{\Delta} = y(\boldsymbol{\Lambda}' \boldsymbol{\Lambda} + \boldsymbol{\Psi})$, and the distribution of $\mathbf{X} | Y$ is:

$$\mathbf{X} | Y \sim N(\boldsymbol{\mu} + y\boldsymbol{\alpha}, y(\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\Psi})) \quad (2)$$

Starting from these assumptions, indicating with $k = 1, \dots, K$ the clusters, the complete log-likelihood function of the mixture of generalized hyperbolic factor analyzers is given by:

$$\begin{aligned} \log(L) = & \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log(\pi_k) + \sum_{i=1}^n \sum_{k=1}^K z_{ik} \sum_{j=1}^P \log \left(\phi \left(\frac{x_i}{\boldsymbol{\mu}_k} + y_i \boldsymbol{\alpha}_k, y_i (\boldsymbol{\Lambda}'_k \boldsymbol{\Lambda}_k + \boldsymbol{\Psi}_k) \right) \right) + \\ & + \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log \left(h \left(\frac{y_i}{\boldsymbol{\omega}_k}, \lambda_k \right) \right), \end{aligned}$$

where $h(y_i/\boldsymbol{\omega}_k, \lambda_k)$ is the density function of the GIG.

The transformation, through a factor-analytic representation of the covariance matrices, allows us to reduce the number of parameters to be estimated and to use the generalize the hyperbolic distribution to cluster higher dimensional data.

The parameters can be estimated using the alternating expectation-conditional maximization (AECM) algorithm [6].

References

1. Ryan P. Browne and Paul D. McNicholas. A mixture of generalized hyperbolic distributions. *arXiv: 1305.1036v1*, 2013.
2. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.
3. Z. Ghahramani and G. E. Hinton. The em algorithm for mixtures of factor analyzers. Crg-tr-96-1, Univ. Toronto, Toronto, ON, Canada, 1997.
4. G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.
5. G. J. McLachlan and D. Peel. Mixtures of factor analyzers. In San Francisco Morgan Kaufman, editor, *Proceedings of the seventeenth International Conference on Machine Learning.*, pages 599–606, 2000.
6. Xiao-Li Meng and David Van Dyk. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.
7. Karsten Prause. *The generalized hyperbolic model: Estimation, financial derivatives, and risk measures*. 1999.

Testing for endogeneity and country heterogeneity

Giovanni Trovato

Abstract An empirical model is built to test for convergence in the Solowian sense. For this purpose, parametric and semi-parametric time-invariant random effect models are considered, by using the *minimal* and the *augmented* specification of the Solow model. An extension to time-varying random effect models is discussed based on the Hidden Markov Models framework.

Key words: human capital, economic development, convergence, parametric and semi-parametric mixtures, time-varying and time-invariant heterogeneity

1 Introduction

Starting from the framework of economic growth dynamics introduced by [24] and [25], a question arises whether countries are fatefully destined to be divided into richer and poorer. This represents an important issue that has been analyzed from both theoretical and empirical perspectives, with contrasting answers. First, a major question is related to the definition of *convergence* in Solow-type models. If Solow assumptions hold, countries' growth rates of income per capita should converge, in the long run, to a single equilibrium, regardless of their initial conditions (the *absolute convergence* hypothesis). Since in the augmented version of the Solow model physical and human capital grow at the same rate, the only explanation for differences across countries in the observed levels of income is that the countries are in different points with respect to the balanced growth path. This difference is temporary since poorer countries, moving from short to long period, grow faster than richer ones; as a side-effect, cross-country convergence in economic aggregates, such as income per capita, requires that structural characteristics (technology, pref-

Giovanni Trovato

Dipartimento di Economia e Finanza, Università di Roma Tor Vergata, e-mail: giovanni.trovato@uniroma2.it

erences, population growth, government policy, democracy, factor market structure, etc.) should converge as well. If differences in growth rates of income per capita are permanent, due eg to country-specific structural heterogeneity, conditional convergence may occur. While the hypothesis of absolute convergence is often rejected, see, among others, [3], [16], [6], recent developments in economic growth empirics suggest that the cross-country distribution of income per capita is multimodal; see [18], [19]. Durlauf and Johnson [10], see also [11], show that the Solow model may be consistent with convergence clubs; countries that are similar in structural characteristics, tend to converge to the same steady-state equilibrium, conditional on country-specific reactions to random shocks, if (and only if) their initial conditions (eg the initial levels of income per capita) are similar. The distinction between initial conditions and unobserved heterogeneity leads to substantial differences in terms of convergence (conditional vs club). According to [12], differences in country-specific growth rates could be due to geographical features, institutional, demographic and cultural factors, trade patterns, colonial status, and public policy. Does this mean that differences in initial conditions are related to structural heterogeneity only? Empirical literature usually represents structural heterogeneity through variables describing geographic, cultural, religious, political differences across countries, while initial levels (at the starting point of the observation window) of gdp per capita, stocks of human and physical capital are used to approximate initial conditions. This structure may lead to endogeneity bias due to dependence between structural variables and random terms representing unobserved heterogeneity. Alfó et al. [2] show that the classical growth equation augmented by proxies may suffer by endogeneity with potentially biased inferences on model parameter estimates. In most empirical studies, unobserved heterogeneity has been found to play a substantial role and influence the path towards convergence. Recent approaches to heterogeneous growth modeling use non- or semi-parametric specifications, see [19], [23], and [26], or threshold models, see, among others, [14], [15], [17]. The paper is structured as follows; in section 2 the empirical model is described from a purely theoretical perspective, while section 3 presents the results obtained by analyzing data drawn from the PWT 7.0, years 1960–2010.

2 Model Structure

We propose a parsimonious specification for the growth model equation via a mixed effect approach and a three-step strategy. First, we model growth using the *standard* long-term equation where only the intercept and the initial gdp per worker are included in the linear predictor specification; both terms are associated to a random coefficient. We adopt both a parametric Gaussian random coefficient approach, see [7], and a semi-parametric alternative based on finite mixtures. We also estimate the standard solowian augmented version of growth model by both parametric and semi-parametric mixtures, to allow for comparison with the “minimal regression model”.

We start by defining the output Y_{it} , the labour input, $L_{it} = L_{i0} \exp(n_{it})$, the state of technology, $A_{it} = A_{i0} \exp(g_{it})$, sometimes referred to as the efficiency of labour input, for the i -th country $i = 1, \dots, m$ at time $t = 1, \dots, T$. In this formulation, technology (A_{it}) and labor force (L_{it}) are assumed to grow at exogeneous, constant, rates identified, respectively, by the progress rate of labour-augmenting technology (g_i) and by the population growth rate (n_i). The theory describes the process of economic growth as a linear (on a log scale) function linking the unobservable output per efficient unit of labour ($y_{it}^E = \frac{Y_{it}}{A_{it}L_{it}}$) and the observable output per effective labor unit $y_{it} = \frac{Y_{it}}{L_{it}}$. In formulas:

$$\log(y_{it}) = [1 - \exp(-\lambda_i t)] \log(y_{i\infty}^E) - \exp(-\lambda_i t) \log(y_{i0}^E) \quad (1)$$

where $y_{i\infty}^E$ represents the steady-state value for the output per efficient labour unit, $\lambda_i > 0$ measures the convergence rate of the output per labor force to the corresponding steady-state value. While useful from a theoretical perspective, this formulation does not help building an estimable model, since the efficiency of labour is not observed. Using the observed output/labor ratio, we may derive from eq. (1):

$$\frac{1}{t} [\log(y_{it}) - \log(y_{i0})] = g_i + \beta_i [\log(y_{i0}) - \log(y_{i\infty}^E) - \log(A_{i0})] \quad (2)$$

where $\gamma_i = \frac{1}{t} [\log(y_{it}) - \log(y_{i0})]$ is the growth rate of the output per unit of labour input, while $\beta_i = -\frac{1}{t} [1 - \exp(-\lambda_i t)]$, $\lambda_i > 0$, measures the convergence rate of the output per unit of labor force to the corresponding steady state value. Equation (2) states that only g_i and $[\log(y_{i0}) - \log(y_{i\infty}^E) - \log(A_{i0})]$ influence the growth process. The first term specifies the technological progress, while the second one identifies the catching-up process, measured by the gap between initial conditions and the steady state values. With increasing time, the model predicts poorer countries catch-up the richer ones, conditional on the technological progress, while initial conditions play no substantial role in the long run. The previous equation may not be estimated, unless the unobservable steady state value is not estimated or approximated as well. Barro and Sala-i-Martin [6] showed that it is possible to relax the steady state assumption to obtain some feasible specification for equation (1). According to [16], we may describe growth dynamics starting from a standard Cobb-Douglas production function, where output Y_{it} is produced using labor force (L_{it}), physical capital (K_{it}), human capital (H_{it}) and technology (A_{it}) as inputs. This relationship may be elicited as follows:

$$Y_{it} = K_{it}^{\alpha_K} H_{it}^{\alpha_H} (A_{it} L_{it})^{(1-\alpha_K-\alpha_H)} \quad (3)$$

where $\alpha_K, \alpha_H \in (0, 1)$ represent the shares of physical and human capital, respectively. According to the deterministic version of the Solow model, country-specific laws of motion for capital inputs determine the corresponding accumulation process. Using lower-case letters to denote per-worker quantities, i.e. $y_{it} = Y_{it}/L_{it}$,

$k_{it} = K_{it}/L_{it}$ and $h_{it} = H_{it}/L_{it}$, we can re-write the production function as:

$$y_{it} = A_{it}^{(1-\alpha_K-\alpha_H)} k_{it}^{\alpha_K} h_{it}^{\alpha_H} \quad (4)$$

Assuming, for sake of simplicity, a constant population growth rate, $g_i = g$, and a constant convergence of the output per unit of labor force, $\lambda_i = \lambda$, $i = 1, \dots, n$, we may define the steady-state value for the output per unit of efficient labour force as

$$y_{i,t}^E = \left[\frac{s_{Ki}^{\alpha_K} s_{Hi}^{\alpha_H}}{(n_i + g + \delta)^{(\alpha_K + \alpha_H)}} \right]^{\left[\frac{1}{1-\alpha_K-\alpha_H} \right]} \quad (5)$$

where sk_i , sh_i represent the share of output invested in physical and human capital, respectively. According to equation (5), the steady state value of the gdp per unit of efficient labour depends on the shares of capital inputs, the growth of labor force and the depreciation rate. Following [16], Durlauf [9] shows that we can define an estimable equation for growth dynamics, by substituting y_{it}^E with the predicted value found through equation (5). Taking logs and subtracting from both side of eq. (1) the initial log-income per worker, $\log(y_{i0})$, we obtain:

$$\begin{aligned} \gamma_{it} = g_i + \beta \log(A_0) - \beta \left[\frac{\alpha_K}{1 - \alpha_K - \alpha_H} \log(sk_{it}) + \right. \\ \left. + \frac{\alpha_H}{1 - \alpha_K - \alpha_H} \log(sh_{it}) - \frac{\alpha_K + \alpha_H}{1 - \alpha_K - \alpha_H} \log(n_i + g + \delta) - \log(y_{i,0}) \right] + \varepsilon_{it} \end{aligned} \quad (6)$$

Following this approach, the equation for the augmented Solow model can be written in a compact form as:

$$\gamma_{it} = \beta \log(y_{i0}) + \mathbf{X}_{it} \boldsymbol{\psi} + \mathbf{Z}_{it} \boldsymbol{\phi} + \varepsilon_i \quad (7)$$

where $\mathbf{X}_{it} = [(g_i + \log(A_0)), \log(sh_{it}), \log(sk_{it}), \log(n_i + g + \delta)]$ denotes the matrix design, while the elements in the additional vector \mathbf{Z}_{it} represent the variables that are identified outside the Solow theory (eg in a Barro-type regression model). According to [9], if we use the previous formulation, we can not adequately control for the correlation between the variables in \mathbf{Z}_{it} and the individual-specific values of the random intercept since some information are yet included in g_i or in $\log(A_{i0})$; at the same time, saving rates can be correlated with initial conditions. For a similar reason, see [13], convergence results based on equation (7) may be biased since physical and human capital accumulations may depend on income per capita (or per worker). From the perspective of model formulation, [9] stress that Barro-type regressions may be difficult to estimate since uncertainty upon the adopted model structure is a substantial problem, which can be only partially solved by means of nonlinear model specifications, as in [14]. Empirical results suggest that cross-sectional tests on convergence may be inflated by collinearity and, since initial levels of per capita gdp are closely related to capital saving rates, additional covariates effects may be ill-estimated.

For these reasons, we proceeded to estimate a growth dynamics model looking at the minimal regression equation only, as described by equation (1), augmented by random terms to account for unobserved heterogeneity sources.

3 Results

By using a semiparametric finite mixture approach to model the random coefficient distribution, we identify 4 to 7 clusters of countries, depending on the model specifications, with homogeneous values for unobserved heterogeneity (random intercept) and convergence parameters (estimate for the effect of initial gdp per worker). The use of finite mixtures allows us to classify countries in groups characterized by homogeneous values of random parameters by looking at posterior probabilities of component membership; this posterior classification can represent an important tool to better understand countries' similarities in the growth process dynamics. Once clusters have been determined, by using a so-called *three-step* procedure, we fit a multinomial logit model where proxies for structural heterogeneity are used to model country-specific posteriors to belong to a given component. In this perspective, we do not propose a formal test of conditional *versus* club convergence; rather, we suggest a parsimonious empirical model to describe, as closely as we can, what is naturally hidden behind the process of economic growth.

References

1. Alfó, M., Trovato G., Semiparametric mixture models for multivariate count data, with application, *Econometrics Journal*, 7, 1–29, 2004
2. Alfó, M., Trovato G. and R.J. Waldmann, Testing for country heterogeneity in growth models using a finite mixture approach, *Journal of applied econometrics*, 23, 487–514, 2008.
3. Barro, R. J., Economic Growth in a Cross Section of Countries, *Quarterly Journal of Economics*, 106, 407–443, 1991.
4. Barro, R. J., Human Capital and Growth, *American Economic Review*, 91, 12–17, 2001.
5. Barro, R. J. and Lee, J.W., International Data on Educational Attainment Updates and Implications, *Oxford Economic Papers*, 53, 541–63, 2001.
6. Barro R. and Sala-I-Martin, X., Convergence, *Journal of Political Economy*, 100, 223–251, 1992.
7. Pineheiro, J.C., Bates, D.M., Approximations to the log-likelihood function in nonlinear mixed-effects models, *Journal of Computational and Graphical Statistics*, 4, 12–35, 1995.
8. Durlauf, S.N., The Local Solow Growth Model, *Journal of Econometrics*, 100, 1, 65–69, 2001.
9. Durlauf, S.N., Johnson P.A. and Jonathan R. W. Temple, *Growth Econometrics* in Handbook of Economic Growth, Volume A, Aghion P. and Durlauf S.N. ed., North-Holland, 2005.
10. Durlauf, S. and Johnson, P.A., Multiple regimes and cross-country growth behaviour, *Journal of Applied Econometrics*, 10, 365–384, 1995.
11. Galor, O., Convergence? Inferences from Theoretical Models, *Economic Journal*, 106, 1056–69, 1996.
12. Galor, O., Moav, O., From Physical to Human Capital Accumulation: Inequality and the Process of Development, *Review of Economic Studies*, 71, 1001–1026, 2004.

13. Goetz, S.J. and Hu, D., Economic Growth and Human Capital Accumulation: Simultaneity and Expanded Convergence Tests, *Economics Letters*, 51, 355–362, 1996.
14. Kalaitzidakis, T., Mamuneas, T.A., Savvides, A. and Stengos, T., Measures of Human Capital and Nonlinearities in Economic Growth, *Journal of Economic Growth*, 6, 229–254, 2001.
15. Liu, Z. and Stengos, T., Non-linearities in Cross-Country Growth Regressions: A Semiparametric Approach, *Journal of Applied Econometrics*, 14, 527–538, 1999.
16. Mankiw, N.G. and Romer, D. and Weil, D.N., A contribution to the empirics of economic growth, *Quarterly Journal of Economics*, 107, 407–37, 1992.
17. Masanjala, W.H. and Papageorgiou, C., The Solow model with CES Technology: nonlinearities and parameter heterogeneity, *Journal of Applied Econometrics*, 19, 171–201, 2004.
18. Paap, R., H.K. van Dijk, Distribution and mobility of wealth of nations, *European Economic Review*, 42, 1269–1293, 1998.
19. Paap, R., P.H. Franses and Dick van Dijk, Does Africa grow slower than Asia, Latin America and the Middle East? Evidence from a new data-based classification method, *Journal of Development Economics*, 77, 553–570, 2005.
20. Pittau, M.G., Zelli, R. and Johnson P.A., Mixture Models, Convergence Clubs and Polarization, *Review of Income and Wealth*, 56, 101–122, 2010.
21. Psacharopoulos G. Patrinos H.A., Returns to investment in education: a further update, *Education Economics*, 12, 111–134, 2004.
22. Sala-i-Martin X., Doppelhofer G. and Ronald I. Miller, Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach, *The American Economic Review*, 94, 813–835, 2004.
23. Savvides, A., Mamuneas, T.P., Stengos, T., Economic development and the return to human capital: a smooth coefficient semiparametric approach, *Journal of Applied Econometrics*, 21, pages 111–132, 2006.
24. Solow, Robert M., A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics*, 70, 65–94. doi:10.2307/1884513, 1956.
25. Swan, Trevor W., Economic Growth and Capital Accumulation. *Economic Record*, 32, 334–361. doi:10.1111/j.1475-4932, 1956.
26. Tsionas, E.G., Kumbhakar S.C., Markov switching stochastic frontier model, *Econometrics Journal*, 7, 398–425, 2004.
27. Wößmann, L., Specifying Human Capital”, *Journal of Economic Surveys*, 17, 3, pp. 239–270, 2003.

Open science in machine learning

Joaquin Vanschoren and Mikio L. Braun and Cheng Soon Ong

Abstract We present OpenML and mldata, open science platforms that provides easy access to machine learning data, software and results to encourage further study and application. They go beyond the more traditional repositories for data sets and software packages in that they allow researchers to also easily share the results they obtained in experiments and to compare their solutions with those of others.

Key words: machine learning, open science

1 Introduction

Research in machine learning and data mining can be speeded up tremendously by moving empirical research results “out of people’s heads and labs, onto the network and into tools that help us structure and alter the information” [3]. The massive streams of experiments that are being executed to benchmark new algorithms, test hypotheses or model new data sets have many more uses beyond their original intent, but are often discarded or their details are lost over time. In this paper, we present recently developed infrastructures that aim to make machine learning research more open. They go beyond the more traditional repositories¹ for data sets, implementations and workflows in that they allow researchers to also share detailed results obtained in experiments and to compare their solutions with those of others.

Joaquin Vanschoren

Leiden University, Leiden, Netherlands, e-mail: joaquin@liacs.nl

Mikio L. Braun

TU Berlin, Berlin, Germany, e-mail: mikio.braun@tu-berlin.de

Cheng Soon Ong

National ICT Australia, Melbourne, Australia, e-mail: chengsoon.ong@unimelb.edu.au

¹ Well-known examples are the UCI repository, (<http://archive.ics.uci.edu/ml>), my-Experiment (<http://myexperiment.org>) and MLOSS (<http://mloss.org>).

This *collaborative* approach to experimentation allows researchers to share all code and results that are possibly of interest to others, which may boost their visibility, speed up further research and applications, and engender new collaborations. Indeed, many questions about machine learning algorithms can be answered on the fly by querying the combined results of thousands of studies on all available data sets. This facilitates much larger-scale machine learning studies, yielding more generalizable results [1]. Last but not least, these infrastructures keep track of experiment details, ensuring that we can easily reproduce them later on, and confidently build upon earlier work [2].

2 OpenML

OpenML (<http://openml.org>) is a website where researchers can share their data sets, implementations and experiments in such a way that they can easily be found and reused by others. It offers a web API through which new resources and results can be submitted automatically, and is being integrated in a number of popular machine learning and data mining platforms, such as Weka, RapidMiner, KNIME, and data mining packages in R, so that new results can be submitted automatically. Vice versa, it enables researchers to easily search for certain results (e.g. evaluations of algorithms on a certain data set), to directly compare certain techniques against each other, and to combine all submitted data in advanced queries.

To make experiments from different researchers comparable, OpenML uses *tasks*, well-described problems to be solved by a machine learning algorithm or workflow. A typical task would be: *Predict (target) attribute X of data set Y with maximal predictive accuracy*. Similar to a data mining challenge, researchers are thus challenged to build algorithms or workflows that solve these tasks. Tasks can be searched online, and will be generated on demand for newly submitted data sets.

Tasks contain all necessary information to complete it, always including the input data and what results should be submitted to the server. Some tasks offers more structured input and output: predictive tasks, for instance, include train and test splits for cross-validation, and a submission format for all predictions. The server will evaluate the predictions and compute scores for various evaluation metrics.

An attempt to solve a task is called a *run*, and includes the task itself, the algorithm or workflow (i.e., *implementation*) used, and a file detailing the obtained results. These are all submitted to the server, where new implementations will be registered. For each implementation, an online overview page is generated summarising the results obtained over all tasks, over various parameter settings. For each data set, a similar page is created, containing a ranking of implementations that were run on tasks with that data set as input.

OpenML provides a REST API for downloading tasks and uploading data sets, implementations and results. This API is currently being integrated in various machine learning platforms such as Weka, R packages, RapidMiner and KNIME².

To make the shared results maximally useful, OpenML links various bits of information together in a single database. All results are stored in such a way that implementations can directly be compared to each other (using various evaluation measures), and parameter settings are stored so that the impact of individual parameters can be tracked. Moreover, for all data sets, it calculates meta-data about the features and the data distribution[4], and for all implementations, meta-data is stored about their (hyper)parameters and properties such as what input data they can handle, what tasks they can solve and, if possible, advanced properties such as bias-variance profiles.

Finally, the OpenML website offers various search functionalities. data sets, algorithms and implementations can be found through simple keyword searches, linked to all results and meta-data. Runs can be aggregated to directly compare many implementations over many data sets (e.g. for benchmarking). Furthermore, the database can be queried directly through an SQL editor, or through pre-defined advanced queries.³ The results of such queries are displayed as data tables, scatter-plots or line plots, which can be downloaded directly.

3 mldata

mldata (<http://mldata.org>) is a community-based website for the exchange of machine learning data sets. Data sets can either be raw data files or collections of files, or use one of the supported file formats like HDF5 or ARFF in which case mldata looks at meta data contained in the files to display more information. Similar to OpenML, mldata can define learning tasks based on data sets, where mldata currently focuses on supervised learning data. Learning tasks identify which features are used for input and output and also which score is used to evaluate the functions. mldata also allows to create learning challenges by grouping learning tasks together, and lets users submit results in the form of predicted labels which are then automatically evaluated.

mldata.org supports four kinds of information: raw data sets, learning tasks, learning methods, and challenges. A raw data set is just some data, while the learning task also specifies the input and output variables and the cost function used in evaluation. A learning method is the description of a full learning workflow, including feature extraction and learner. One can upload predicted labels for a data set and a task to create a solution entry which automatically evaluates the error on the predicted labels. Finally, a number of learning tasks can be grouped to create a challenge.

² Beta versions of these integrations can be downloaded from the OpenML website.

³ See the Advanced tab on <http://openml.org/search>.

Most of this data is text. `mldata` defines a general file exchange format for supervised learning based on HDF5, a structured compressed file format. It is similar to an archive of files but has additional structure on the level of the files, such that users can directly store and access matrices, or numerical arrays. Using this specified file format is not mandatory, but using it unlocks a number of additional features like a summary of the data set, and automatic conversion into a number of other formats.

Currently, OpenML is being integrated with `mldata`, so that data sets and learning methods can be shared between both platforms.

4 Related work

There also exist platforms aimed at providing reproducible benchmarks. DELVE (<http://www.cs.utoronto.ca/~delve>) was the first, but is currently in abeyance. MLComp (<http://mlcomp.org>) allows users to upload their algorithms and evaluate them on known data sets (or vice versa) on MLComp servers. RunMyCode (<http://runmycode.org>) allows researchers to create *companion websites* for publications by uploading code and building an interface. Users can then fill in all inputs online and get the result of the algorithm.

Compared to these systems, OpenML and `mldata` allow users more flexibility in running experiments: new tasks can be introduced for novel types of experiments and experiments can be run in any environment. OpenML also offers clean integration in data mining platforms that researchers already use in daily research, and closer data integration so that researchers can reuse results in many ways beyond direct benchmark comparisons, such as meta-learning studies [5].

Acknowledgments

This work is supported by grant 600.065.120.12N150 from the Dutch Fund for Scientific Research (NWO), and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

1. Hand, D.: Classifier technology and the illusion of progress. *Statistical Science* (Jan 2006)
2. Hirsh, H.: Data mining research: Current status and future opportunities. *Statistical Analysis and Data Mining* 1(2), 104–107 (Jan 2008)
3. Nielsen, M.: The future of science: Building a better collective memory. *APS Physics* 17(10) (2008)
4. Peng, Y., Flach, P., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. *Lecture Notes in Computer Science* 2534, 141–152 (Jan 2002)
5. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases. A new way to share, organize and learn from experiments. *Machine Learning* 87(2), 127–158 (2012)

Logistic Regression and Decision Tree: Performance Comparisons in Estimating Customers' Risk of Churn

Valerio Veglio

Abstract Churn prediction is becoming a critical business issue for global organizations. Many previous studies have tried to identify accurate churn management models. Predictive modeling based on knowledge discovery through data mining is an approach being used to facilitate the estimation of the risk of churn probability. This research compares the effectiveness of both logistic regression and decision trees models in solving a real-world churn risk problem in order to identify which of them is more accurate in estimating the risk of customer churn. The receiver operating characteristic (ROC) curve proves that logistic regression is slightly more accurate than decision trees in the estimation of customers' churn.

1. Introduction

Customer churn is defined as the propensity of customers to cease doing business with a companies in a given time period. Estimating it has become a significant problem and is one of the prime challenges that many companies worldwide are having to face [2]. The churn of customers causes a huge loss for companies and it becomes a very serious problem. In order to survive in hyper competitive markets companies are turning to data mining techniques for churn analysis.

This research addresses the importance of data mining models in estimating the risk of churn. A comparison between logistic regressions and decision trees models is provided to test which model better performs in managing the customer churn. Section 2 provides a general conceptual background related to the customer churn problem. Section 3 explains both data and methodology developed in this study. Section 4 shows

the main empirical findings. Section 5 draws conclusions and suggests future research directions.

2. Conceptual Background

The most important predictive modeling techniques include decision trees and neural networks [4]. Neural networks and decision trees are typical classification technologies [1]. A number of studies using various algorithms, such as sequential patterns [3], genetic modeling [6], and neural networks [11], have been used to explore customers churn and to demonstrate the potential of data mining through experiments and case studies. Decision trees, neural networks, and logistic regression are well suited to study the customer churn management problem [8]. An investigation conducted by Datta, et al. (2001) [5] reveals that neural networks were only used by few companies in estimating the risk of churn probability. In fact, the current literature identifies four main reasons. First, neural networks are often cited as a methodology that builds a black box. Second, neural networks are known for the lack of clarity of outputs. Third, neural networks analysis provides internal weights, which are distributed throughout the network. But these weights do not provide insight into why the solution is valid [14]. Finally, neural networks are not readily understandable by the decision makers. On the other hand, decision trees and logistic regression are characterized by “simple” if then rules that can sometimes be understood by decision makers. Regression models provide a straightforward relationship between the independent variables and the predicted variable [8], and are one of the main data mining models used to predict the risk of churn within companies.

3. Data and Methodology

The dataset analysed was provided by a global marketing consulting firm that offers digital data driven marketing solutions across all interactive channels: digital, direct response, relationship based media and design.

The whole database contains 1.463.199 customers and 16 quantitative variables related to their purchase behaviour. The target variable (Purchases) is dichotomous and it assumes two values: 1 when the customer purchases an online service (good customer) and 0 otherwise (bad customers). Logistic regression models based on “Enter” method, and classification decision trees based on the “CART” algorithm are developed in this research. Logistic regression builds a model with a dichotomous outcome and is proven as a powerful algorithm [10]. It is well studied and is used in many applications, especially in the marketing area, to estimate the relationship among variables [13]. The “Enter” method tries to detect “step-by-step” small changes of the variables selected. This is the main reason why we use this method rather than methods such as “Backward” and “Forward”. A decision tree, based on the “CART” algorithm, is constructed by recursively splitting the instance space into smaller sub-groups until a specified criterion has been met. We choose the “CART” algorithm because it is one of the most popular algorithms for business problems in the literature. This algorithm prefers to stop the growth of the classification tree through a pruning mechanism rather than with a stopping criterion based on the significance of the chi-squared test such as the CHAID algorithms [9]. Finally, the ROC curve has been used in order to evaluate the correctness of the predictive models previously implemented. Remarks that the curve will always lie above the 45° line. The area between the curve and the line can

also be calculated, and gives the Gini index of performance. The higher the area, the better the model [7].

Other decision trees algorithm such as C4.5 and C5.0 are used in the current literature, but mainly in engineering rather than business contests [7].

4. Empirical Findings

Table 1 provides the main results of the logistic regression based on the “Enter” methods, while Table 2 identifies the main customer profiles generated by the decision tree based on “CART” algorithm.

Table 1: Logistic Regression (Final Model)

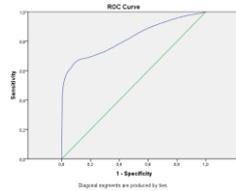
VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β)
Average Click Through Rate	4.194	.245	294.090	.000	66.267
Average Position Best Five	.972	.102	90.300	.000	2.643
Brand Search	1.887	.085	494.667	.000	6.601
Impression or Click at 4pm	.018	.002	80.482	.000	1.018
Match Type: Broad	1.760	.085	427.276	.000	5.811
Match Type: Exact	2.725	.095	819.180	.000	15.254
Search Engine on Google	.346	.096	13.086	.000	1.414

*Sig. < 0.05

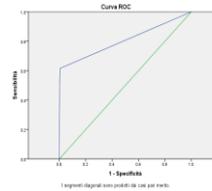
Table 2: Interpretation of the Decision Trees Model based on “CART” Algorithm

Node	Customer Category	Purchase probability	Number of potential customers
2	The potential customer has been exposed to a banner on an affiliate website at least one time.	91%	7.362
15	The potential customer has never been exposed to a banner on an affiliate website AND has been exposed to banners whose mean click through rate is lower than 0,5% AND has never digit a campaign keyword on Google AND has visited “Conion MCUK” website less than 7 times AND has visited “MCUKYahooQ2006” website less than 21 times.	1%	1.438.292

Graph 1: Evaluation of the Logistic Regression Model



Graph 2: Evaluation of the Decision Tree Model



The general correctness of both logistic regression and decision tree is moderate [12]. Graph 1 shows that the Area Under ROC curve (AUC) is equal to 82.10%, while Graph 2 has AUC is equal to 80.20%.

5. Conclusions

Logistic regression outperforms the decision trees. The logistic regression model identifies seven strategic drivers (see table 1) that lead customers to purchase the

service online. A clear connection between online marketing and customer conversion has been strongly confirmed by the data. The data analysis shows an association between purchases and almost all of the marketing trackers in the dataset. Customers that search the exact name of a keyword are seven time more likely to purchase the service than potential customers who have not. The decision tree model is consistent with the regression one. In fact, affiliate web site is a strategic driver for minimizing the probability of churn. In addition, the classification decision tree accurately discriminates a category of customers who are very unlikely to purchase the company's service. The ROC curve confirms that the predictive power of the logistic regression is slightly more accurate than the classification tree in terms of AUC. Neural networks models have been not developed in this research, according to the conceptual background previously described.

References

1. Baesens, B., Verstaeten, G., Van den Poel, D, Egmont-Peterson, M., Van Kenhove, P., and Vanthienen, J.: Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*,156:508-23 (2004).
2. Chandar, M., Laha, A., and Krishna, P.: Modelling Churn behaviour of bank customers using predictive data mining techniques. *National conference on soft computing techniques for engineering applications* (2006).
3. Chiang, D., Wang, Y., Lee, S., and Lin, C.: Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications*, 25(3), 293-302 (2003).
4. Crespo, F., Weber, R.: A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems*,150:267-84 (2004).
5. Datta, P., Masand, B., Mani, DR., and Li, B.: Automated cellular modeling and prediction on a large scale. *Issues on the Application of Data Mining*, 485-502 (2001).
6. Eiben, A.E., Koudijs, A.E., and Slisser, F.: Genetic modelling of customer retention. *Lecture Notes in Computer Science*, 1391, 178-186 (1998).
7. Giudici, P.: *Data Mining*, Milano, McGraw-Hill (2010).
8. Hadden, J., Tiwari, A., Roy, R., and Ruta, D.: Computer assisted customer churn management: State-of-the-art and future trends. In *Computers & Operations Research*, 34:2902-2917 (2005).
9. Kass, GV.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2):119-127 (1980).
10. Khakabi, S., and Mohammad, R.G.: *Data Mining Applications in Customer Churn Management*. International Conference on Intelligent Systems, Modelling and Simulation. Computer Society, IEEE (2010).
11. Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E., and Kaushansky, H.: Predictive subscriber dissatisfaction and improving retention in the wireless telecommunication industry. *IEEE Transactions on Neural Networks*, Special issue on Data Mining and Knowledge Representation,690-696 (2000).
12. Swets, J.A.: Measuring the accuracy of diagnostic system. *Science*, 240:1285-1293 (1988).
13. Tan, PN., Steinbach, M., and Kumar, V.: *Introduction to Data Mining*, Pearson Education Ltd, Boston (2011).
14. Yang, L., and Chiu, C.: Knowledge discovery on consumer churn prediction. *Proceedings of the 10th WSEAS International Conference on Applied Mathematics*, Dallas, Texas, USA, 523-528 (2006).

Robust Two-mode clustering

Maurizio Vichi

Department of Statistics, Sapienza University of Rome, Italy,
maurizio.vichi@uniroma1.it

Two-mode clustering is the activity of clustering modes (e.g., objects, variables) of an observed two-mode data matrix, simultaneously. This task is required because objects, frequently, are homogeneous only within subsets of variables, while variables may be strongly associated only on subsets of objects. For example, in microarray data analysis groups of genes are generally co-regulated within subsets of samples and groups of samples share a common gene expression pattern only for some subsets of genes. In market basket analysis customers have similar preference patterns only on subsets of products and, vice-versa, classes of products are more frequently consumed and preferred by subgroups of customers. In these situations a classical cluster analysis would cluster one mode (e.g., objects) on the basis of the complete set of the other mode (e.g., variables), thus producing weak results, while this is avoided with a more appropriate two-mode clustering. For *big data*, represented by matrices with a huge number of rows and columns, frequently the main analysis is a two-mode clustering, trying to mine and synthesize the relevant information by reducing the size of the data to a matrix of compact dimensions formed by prototype objects and variables. This is achieved by the simultaneous grouping rows and columns so that results are informative and easy to interpret, denoting compressed, but relevant representation of the big data, while trying to preserve most of the original information. The reduction is generally soft to obtain a light compression of the multivariate data in order to allow the successive application of other multivariate statistical methods that are computationally prohibitive for large data sets. The quality of big data is not always certifiable and frequently they are inflated by many outliers and influential data that have an high impact on the two-mode clustering and successive analyses. Therefore, robust multimode clustering techniques are needed for compressing large data set, while preserving the most relevant information.

A new robust asymmetrical two-mode clustering technique is proposed. A coordinate descent algorithm is developed. The applications on both, synthetic and real datasets, validate the performance and applicability of the new algorithm.

Key Words: Two-mode clustering, double k-means, disjoint principal component analysis, robustness.

Hierarchical Graphical Models and Item Response Theory

Vincenzina Vitale

Abstract The topic of this article is IRT modeling in the presence of nonignorable missing item responses. A Multilevel (or Hierarchical) Bayesian model is developed and represented by a Directed Acyclic Graph. In the context of Educational Assessment (PISA, TIMSS, NEAP), it makes sense to believe that pattern of missingness depends on the ability that is measured and hence data are missing not at random. Three different Graphical Models, in a Hierarchical framework, are considered and compared: 1) a between-item-multidimensional model, taking into account missing data mechanism; 2) a unidimensional Rasch model, ignoring missing data process; 3) a unidimensional Rasch model for which omitted responses are treated as incorrect.

Key words: Rasch Model, Missing Data, Graphical Model, Hierarchical Model

1 Introduction

In Educational measurement, many large-scale surveys typically suffer from a substantial amount of missing data. It often happens that item nonresponses are non-ignorable missing data: pattern of missingness depends on the ability that is measured. Practical approaches treat missing data either as missing observations either as wrong responses. Both cases appear problematic because the former assumes that nonresponse is ignorable while the latter assumes that an omission always indicates an incorrect response. Holman and Glas (2005) suggested the inclusion of an IRT model that governs the missing data to reduce the bias in the estimates of the model parameters in case of violation of Rubin's ignorability principle (Rubin 1976). As shown in Rose, von Davier, and Xu (2010), it is possible to define the model as a between multidimensional item response model (Adams, Wilson, and Wang 1997)

Phd Student, University of Roma Tre, Department of Economics, Via Silvio D'Amico, 77, 00145, Rome, Italy. e-mail: vincenzina.vitale@uniroma3.it

and, hence, to introduce two latent traits: the ability and the response propensity that, in case of nonignorable missing, are correlated. The authors (Rose, von Davier, and Xu 2010) also compared parameter estimates (difficulty and ability parameters of an IRT model) obtained from these three different approaches in a MML framework. In this article it will be shown the same comparison but from another point of view. In the next section, the three models, one of these bidimensional, will be defined by means of a Directed Acyclic Graph (DAG); this graphical representation makes clear the conditional independence structure between random variables, both observed and latent. Bayesian approach enriches these models because it allows to assign a prior distribution to all stochastic nodes. The structure of data, a sample of Italian students of PISA 2006 survey, suggests to consider items nested in students, that are nested in other hierarchical levels (in this study subnations); in other words, a multilevel structure is added.

2 Model Specification

To set the notation, let $i = 1, \dots, I$ denote the level-1 units (items), $j = 1, \dots, J$ denote the level-2 units (students) and $g = 1, \dots, G$ the level-3 units (Italian subnations). Let y_{ijg} be the dichotomous response with code 1 if student j in group g responds to item i correctly, 0 otherwise. Let d_{ijg} be the dichotomous indicator variable with code 1 if y_{ijg} is observed, 0 otherwise. For all i, j and g , y_{ijg} and d_{ijg} are assumed to be independent Bernoulli random variables with the probability of correct response $p_{ijg} = P(y_{ijg} = 1)$ and $k_{ijg} = P(d_{ijg} = 1)$. It is worthwhile to index individual responses as $l = 1, \dots, n$ with each response l associated with a group $g[l]$, a person $j[l]$ and an item $i[l]$. The bidimensional model (called NIM model) can be written as:

$$y_l \sim \text{Bernoulli}(p_l); \text{logit}(p_l) = \theta_{1,g[l]}^{(3)} + \theta_{1,j[l]}^{(2)} - \beta_{i[l]} \quad (1)$$

$$d_l \sim \text{Bernoulli}(k_l); \text{logit}(k_l) = \theta_{2,g[l]}^{(3)} + \theta_{2,j[l]}^{(2)} - \delta_{i[l]} \quad (2)$$

where:

- $\underline{\theta}_j^{(2)} \sim \text{BVN}(\underline{\gamma}, T_{\theta^{(2)}}^{-1})$ is a random effect, the ability of student j nested in group g ;
- $\underline{\theta}_g^{(3)} \sim \text{BVN}(\underline{0}, T_{\theta^{(3)}})$ is the random effect for group g ;
- $\beta_i \sim N(0, 0.0001^2)$ e $\delta_i \sim N(0, 0.0001)$ are fixed effects representing difficulty of item i .

The following prior distributions complete the model:

$$\underline{\gamma} \sim \text{BVN}(\underline{0}, I_{2 \times 2}); T_{\theta^{(3)}} \sim \text{Wishart}(R, 2); T_{\theta^{(2)}} \sim \text{Wishart}(S, 2), \text{ with } R, S = I_{2 \times 2}$$

All these priors are “non-informative”³. As shown in Chaimongkol, Huffer, and

¹ Precision matrix

² Precision parameter

³ For Wishart distribution (the conjugate prior for the inverse covariance matrix), the least informative, proper prior is given by setting, in this case, the degree of freedom equal to 2.

Kamata (2007), one way to identify the model is by allowing the parameters to float and then defining new (adjusted) quantities that are well-identified but preserve the logit of the model; in this case we obtain:

$$\text{logit}(p_i) = \theta_{1,g[l]}^{adj,(3)} + \theta_{1,j[l]}^{adj,(2)} - \beta_i^{adj} \quad (3)$$

$$\text{logit}(k_i) = \theta_{2,g[l]}^{adj,(3)} + \theta_{2,j[l]}^{adj,(2)} - \delta_i^{adj}. \quad (4)$$

$$\text{where: } \beta_i^{adj} = \beta_i - \bar{\beta}; \quad \theta_{1,g}^{adj,(3)} = \theta_{1,g}^{(3)} - \bar{\theta}_1^{(3)}; \quad \theta_{1,j}^{adj,(2)} = \theta_{1,j}^{(2)} - \bar{\beta} + \bar{\theta}_1^{(3)} \quad (5)$$

$$\delta_i^{adj} = \delta_i - \bar{\delta}; \quad \theta_{2,g}^{adj,(3)} = \theta_{2,g}^{(3)} - \bar{\theta}_2^{(3)}; \quad \theta_{2,j}^{adj,(2)} = \theta_{2,j}^{(2)} - \bar{\delta} + \bar{\theta}_2^{(3)} \quad (6)$$

Graphical model are represented in Figure 1. The unidimensional model, called

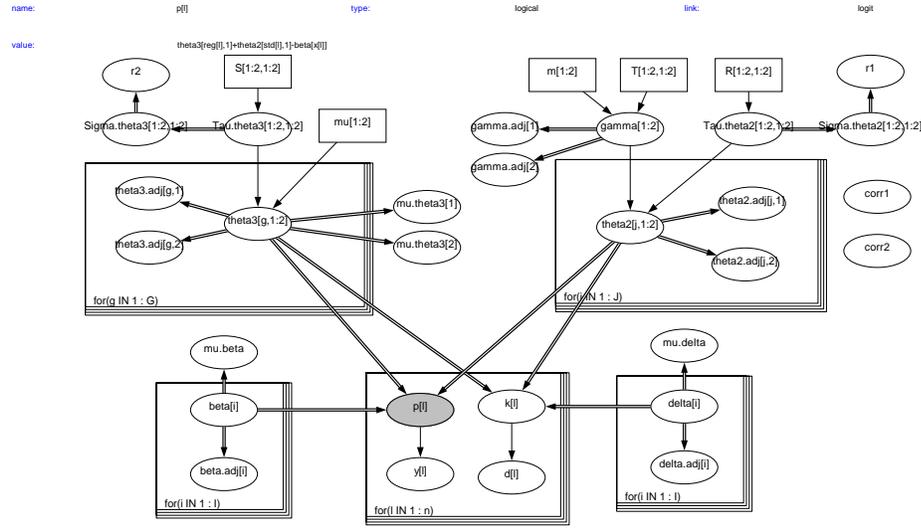


Fig. 1 DAG for NIM model

IM model, assumes that missing data mechanism is ignorable, so we do not need a missingness model as imputation of y_{miss} is unnecessary for valid inference about parameters. Values for y_{miss} can be generated from the posterior predictive distribution. The unidimensional model, called ZIM model, assumes that all missing data are wrong responses, so data matrix is complete. For these two approaches, hence, we have only the model as in Eq (3) with:

- $\theta_j^{(2)} \sim N(\gamma, \tau_{\theta^{(2)}})$; $\beta_i \sim N(0, 0.0001)$; $\tau_{\theta^{(2)}} \sim \text{Gamma}(0.001, 0.001)$
- $\theta_g^{(3)} \sim N(0, \tau_{\theta^{(3)}})$; $\gamma \sim N(0, 1)$; $\tau_{\theta^{(3)}} \sim \text{Gamma}(0.001, 0.001)$

3 Applications and Results

Using software Openbugs, posterior inference is based on the output of a Gibbs sampler. After monitoring convergence of two chains with different initial values, we obtain parameter estimates both for item difficulties and group abilities. First of all, it is important to evaluate correlation between ability and response propensity because its magnitude gives an indication of the extent to which the assumption of ignorability is violated. Referring to mathematical ability of our sample, correlation, at student and group level, is 0.68 and 0.67. Both are statistically significant so it is important to compare parameter estimates in order to evaluate bias produced by IM e ZIM models with respect to NIM model. Bias is proportional to the amount of missing data (only 17% for this sample) and the correlation between dimensions. More generally, it seems that NIM model produces estimates that are always included between those produced by the other two models. ZIM model leads to heavily biased parameter estimates; with respect to NIM model, it increases differences between opposite groups⁴ and increases estimates of difficulty parameters with high positive sign. IM model overestimates items with low and medium difficulty and groups with negative ability; it underestimates items with high difficulty and groups with positive ability.

More generally, NIM model weighs ability estimates with relation to amount of missing data.

Comparing ranking of NIM and IM parameter estimates, then ranking of NIM and ZIM parameter estimates, it is important to note some switches in the position of contiguous items (or groups). Covariates, added to NIM model, confirm that what affects ability process, it also affects response strategy: non immigrant, male students of North Italy, with a good Economic, Social and Cultural background, have the best performances with relation to Math ability and Response propensity.

References

- Adams, R. J., M. R. Wilson, and W. Wang (1997). "The multidimensional random coefficients multinomial logit model." In: *Applied Psychological Measurement*, 21, pp. 1–23.
- Chaimongkol, S., F. Huffer, and A. Kamata (2007). "An explanatory differential item functioning (DIF) model by the WinBUG 1.4". In: *Songklanakarin J. Sci. Technol.*, 29(2), pp. 449–458.
- Holman, R. and C. A. W. Glas (2005). "Modelling nonignorable missing data mechanisms with item response theory models." In: *British Journal of Mathematical and Statistical Psychology*. 58, pp 1–17.
- Rose, N., M. von Davier, and X. Xu (2010). *Modeling Nonignorable Missing Data with Item Response theory(IRT)*, tech. rep. RR-10-11. ETS.
- Rubin, D. B. (1976). "Inference and missing data." In: *Biometrika*, 63(3), pp 581–592.

⁴ Opposite with relation to ability and amount of missing data

Extending the JM libraRy

Sara Viviani

Abstract Joint Model (JM) is a useful modelling framework to account for non-ignorability of the dropout in longitudinal studies. This model assumes that the expected value of the response at the observation time may have an influence on the risk of dropout at the same time. JM in R is a useful and extended package to implement joint models for Gaussian longitudinal responses. In this paper, we aim at proposing an extension of this package to longitudinal non-Gaussian outcomes. We introduce the corresponding model framework, that we referred to as Generalized Linear Mixed Joint Model (GLMJM), and the estimation method. Numerical integration over the random effect posterior distribution of the log-likelihood is pursued by using Gauss-Hermite and Pseudo-Adaptive Gaussian quadrature rules. The model is implemented for the case of Poisson and Binomial longitudinal responses.

Key words: Discrete longitudinal responses, dropout, survival analysis, joint models, pseudo-adaptive Gaussian quadrature

1 The generalized linear mixed joint model

In this Section, we describe the class of generalized linear mixed joint models (GLMJMs), introduced in [6]. These models may be seen as an extension of the linear mixed joint model, see [7], to non-Gaussian responses in longitudinal studies with attrition. The basic idea is to assume that the expected value of the longitudinal outcome, measured at the dropout time, may influence the risk of dropout. In the proposed parametrization, the longitudinal response has distribution in the exponential family, while the time a subject spends in the study is assumed continuous and parametrized through a proportional hazard model, see [2].

Let us denote by $T_i = \min(T_i^*, C_i)$ the observed failure time for the i th individual,

Sara Viviani
Sapienza, University of Rome, e-mail: sara.viviani@uniroma1.it

$i = 1, \dots, n$, taken as the minimum between the true event time T_i^* and the censoring time C_i , which may correspond to the end of the follow-up. Further, let δ_i be the event indicator defined by $\delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function. The outcome $Y_i(t)$ is repeatedly observed before T_i at $t = 1, \dots, n_i$ occasions, and is missing for $t \geq T_i$. We assume that the longitudinal process is associated with T_i^* , i.e. the *true* event time, but it is independent of the censoring time C_i . Let \mathbf{Y} be a random variable with distribution in the exponential family and natural parameter $\mu_i(t)$, that is $\mathbf{Y} \sim \text{EF}(\mu_i(t))$.

The GLMJM is defined by the following equations:

$$\begin{cases} g(m_i(t)) = \boldsymbol{\beta}^\top \mathbf{X}_i(t) + \mathbf{b}_i^\top \mathbf{Z}_i(t) \\ h(T_i | M_i(T_i), \mathbf{W}_i) = h_0(T_i) \exp\{\boldsymbol{\gamma}^\top \mathbf{W}_i + \alpha m_i(T_i)\}. \end{cases} \quad (1)$$

In the first sub-equation in (1), $\mathbf{X}_i(t)$ is a vector of p predictors with fixed effects $\boldsymbol{\beta}$ and $\mathbf{Z}_i(t)$ is a vector of q predictors with random coefficients \mathbf{b}_i . The covariate matrices may contain the time at which the response is measured, as well as the interaction between time and other covariates. On the other hand, $m_i(t) = m(\mu_i(t)) = g^{-1}(\boldsymbol{\beta}^\top \mathbf{X}_i(t) + \mathbf{b}_i^\top \mathbf{Z}_i(t))$ is the expected value of the longitudinal response at time t , $g(\cdot)$ is a link function and $g(m_i(t)) = \mu_i(t)$.

For what concerns the second sub-equation in (1), $M_i(T_i) = \{m_i(u) : 0 \leq u \leq T_i\}$ denotes the history of the true, but unobserved, longitudinal process up to T_i and \mathbf{W}_i is a row vector of additional (time constant) covariates, with fixed parameter vector $\boldsymbol{\gamma}$. The baseline risk function $h_0(T_i)$ is parametrized through a Weibull baseline function, i.e. $h_0(T_i) = \xi T_i^{\xi-1}$, to avoid underestimation of the parameter standard errors, see [3].

2 Parameter Estimation

Maximum likelihood estimation for the complete parameter vector in the proposed parametrization is pursued by using the EM algorithm, as it is typical in mixed effect and joint modeling frameworks. A Bayesian estimation procedure for a similar JM has been proposed by [4].

In this Section, the expectation and the maximization steps are described. The estimation algorithm is based on the computation of the score vector and the Hessian matrix; in both cases, numerical integration with respect to the posterior random effect distribution is required.

The Hessian matrix at convergence is also considered, as standard errors for parameter estimates are based on inverting the observed information matrix.

The E-step requires the calculation of the observed-data score vector as follows:

$$\begin{aligned}
\mathcal{S}(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}^\top} \log \int p(T_i, \delta_i | \mathbf{b}_i; \boldsymbol{\Phi}) p(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta}) p(\mathbf{b}_i; \mathbf{D}) d\mathbf{b}_i \\
&= \sum_i \int \omega(\boldsymbol{\theta}, \mathbf{b}_i) p(\mathbf{b}_i | T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{b}_i,
\end{aligned} \tag{2}$$

where $\omega(\boldsymbol{\theta}, \mathbf{b}_i) = \partial \{\log p(T_i, \delta_i | \mathbf{b}_i; \boldsymbol{\Phi}) + \log p(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta}) + \log p(\mathbf{b}_i; \mathbf{D})\} / \partial \boldsymbol{\theta}^\top$ denotes the complete data score vector. At the q th iteration, given the observed data and the parameter estimates calculated at the previous iteration, $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(q-1)}$, the posterior random effect distribution $p(\mathbf{b}_i | T_i, \delta_i, \mathbf{y}_i; \hat{\boldsymbol{\theta}}^{(q-1)})$ can be written as follows:

$$p(\mathbf{b}_i | T_i, \delta_i, \mathbf{y}_i; \hat{\boldsymbol{\theta}}^{(q-1)}) = \frac{p(\mathbf{y}_i, T_i, \delta_i, \mathbf{b}_i; \hat{\boldsymbol{\theta}}^{(q-1)})}{\int p(\mathbf{y}_i, T_i, \delta_i, \mathbf{b}_i; \hat{\boldsymbol{\theta}}^{(q-1)}) d\mathbf{b}_i}. \tag{3}$$

For the score vector computation, numerical integration methods are needed. We consider Gauss-Hermite (GH) and Pseudo-Adaptive Gaussian (PA) rules, [5]. The GH rule approximates the integral in (2) through a weighted sum of the integrand calculated at pre-specified abscissas, taking into account the distribution of each involved process, i.e. $p(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta})$, $p(T_i, \delta_i | \mathbf{b}_i; \boldsymbol{\Phi})$ and $p(\mathbf{b}_i; \mathbf{D})$. However, there are some known limitations of this approach, that concern the poor approximation of the integral when the random effect distribution is not well approximated by a Gaussian distribution, the need for a high number of quadrature points and the corresponding high computational burden, especially when $\dim(\mathbf{b}_i) > 1$. The PA method is considered as an alternative to handle these problems. Effectively, it computes the integrand by scaling and centering it only once, at the beginning of the optimization algorithm, to approximate the random effect posterior distribution for faster fitting.

At the $q + 1$ -th iteration, the maximization step provides maximum likelihood parameter estimates as follows:

$$\begin{aligned}
\hat{\mathbf{D}}^{(q+1)} &= n^{-1} \sum_i \widehat{\text{Cov}}(\mathbf{b}_i | \mathbf{y}_i, T_i, \delta_i; \hat{\boldsymbol{\theta}}^{(q)}) = n^{-1} \sum_i \mathbf{b}_i^\top \mathbf{b}_i p(\mathbf{b}_i | \mathbf{y}_i, T_i, \delta_i; \hat{\boldsymbol{\theta}}^{(q)}) \\
\hat{\boldsymbol{\beta}}^{(q+1)} &= \hat{\boldsymbol{\beta}}^{(q)} - \left\{ \frac{\partial}{\partial \boldsymbol{\beta}^{(q)\top}} \mathcal{S}(\hat{\boldsymbol{\beta}}^{(q+1)}) \right\}^{-1} \mathcal{S}(\hat{\boldsymbol{\beta}}^{(q+1)}) \\
\hat{\boldsymbol{\Phi}}^{(q+1)} &= \hat{\boldsymbol{\Phi}}^{(q)} - \left\{ \frac{\partial}{\partial \boldsymbol{\Phi}^{(q)\top}} \mathcal{S}(\hat{\boldsymbol{\Phi}}^{(q)}) \right\}^{-1} \mathcal{S}(\hat{\boldsymbol{\Phi}}^{(q)})
\end{aligned}$$

3 The extended JM package

The proposed package has a single fitting function, `jointModel.Discr()`, with the following arguments:

`lmerObject`: a fitted longitudinal model with the function `lmer()` of the **lme4** package.
`survObject`: a fitted survival model with the functions `survreg()` or `flexsurv()` of the **survival** package.
`timeVar`: the name of the time variable in the longitudinal model.
`method`: it can be `'weibull-PH-GH'` or `'weibull-PH-aGH'`. The first specification fits a GLMJM with survival time parametrized through Weibull proportional hazard model, and Gauss-Hermite quadrature rule; the second specification uses Pseudo Adaptive quadrature rule as integration method.
`distribution`: the longitudinal outcome distribution. It can be `'poisson'` or `'binomial'`.
`control`: a list of control arguments.

Hence, `lmerObject` and `survObject` are the *ignorable* models, based on which the *non-ignorability* parameter α is estimated. Moreover, the `control` argument includes, among others, the maximum number of iterations for the EM algorithm and the number of quadrature points for the Gauss-Hermite and the Pseudo-Adaptive quadrature rule, fixed at 15 by default.

The resulting values from the `jointModel.Discr()` function are:

`coefficients`: a list of the estimated model parameters.
`Hessian`: the Hessian matrix at convergence.
`logLik`: the model log-likelihood.
`EB`: a list of random effect posterior quantities.
`iters`: the number of iterations at convergence.
`convergence`: has the estimation algorithm reached convergence?
 ...

References

1. Albert, P. S. and Follmann, D. A.: Modeling repeated count data subject to informative dropout. *Biometrics*. **56**, 667–677 (2000).
2. Cox, D. R.: Regression models and life tables (with discussion). *J. of the Royal Stat. Society, Series B*. **34**, 187–220 (1972).
3. Hsieh, F., Tseng, Y. K. and Wang, J. K.: Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*. **62**, 1037–1043 (2006).
4. Rizopoulos, D. and Ghosh, P.: A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. in Med.* **30**, 1366–1380 (2011).
5. Rizopoulos, D.: Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Comput. Stat. and Data Anal.* **56**, 491–501 (2012).
6. Viviani, S., Alfó, M. and Rizopoulos, D.: Generalized linear mixed joint model for longitudinal and survival outcomes. *Stat. and Comput.* (2013) doi: DOI 10.1007/s11222-013-9378-4.
7. Wulfsohn, M. and Tsiatis, A.: A joint model for survival and longitudinal data measured with error. *Biometrics*. **53**, 330–339 (1997).

Visualisations of Classification Tree Models: An Evaluative Comparison

Adalbert F.X. Wilhelm

Abstract Classification trees are a widely used supervised learning tool in data analysis and knowledge discovery in databases. A major reason for their popularity is their straightforward interpretability and the ease of turning the model into decision rules for automated application. The ease and quality of interpretation, however, depends strongly on visual representations of the tree. Based on a three-stage conceptual framework of data analysis, we evaluate visualisation techniques for classification trees and logistic regression models using a data set from political science. Guided by common goals of data exploration in this context, we compare different visualisation techniques, such as tree plots, tree maps and spine plot of leaves, with each other. At the same time we juxtapose these plots to visualisation tools used in logistic regression, such as scatter plot matrices and effects plots.

Key words: Classification Tree, Visualization, Logistic regression, Treemap, Linked Highlighting

1 Introduction, Framework and Data

Data analysis is an iterative process that comprises various steps. It is claimed repeatedly (e.g. Unwin et al. (2002)) that visualisation techniques can successfully enhance the knowledge discovery in data base process. Liu and Salvendy (2007) present a conceptual framework for the visualisation support in the KDD process. As Liu and Salvendy (2007, p. 97) point out, there are three main stages in the analysis process, namely algorithm selection, model construction, and model evaluation, that can be enhanced particularly by the use of appropriate visualisation techniques. A core aim of applying visual techniques in the data analysis process is to ease the interaction between algorithms and the user. In particular, context knowledge of the

Adalbert F.X. Wilhelm
Jacobs University, Bremen, Germany e-mail: a.wilhelm@jacobs-university.de

data owner is considered to be of vital importance and often provides a substantial input to improve modelling results.

For the purpose of this paper we have chosen a data set from political science about arms trade, military expenditure and their impact on the occurrence of armed conflicts in sub-Saharan Africa. The data set has been prepared and originally analysed by Craft and Smaldone (2002) and is available in the replication data archive of the *Journal of Peace Research* at <http://legacy.prio.no/Research-and-Publications/Journal-of-Peace-Research/Replication-Data/Detail?oid=301968>. A primary goal of the original research with this data was to determine whether arms trade is a predictor of political violence in sub-Saharan Africa (cf. Craft and Smaldone, 2002, p. 696). In case of a positive assertion to this first hypothesis, a second question is whether arms trade's influence is already covered by the effect of military expenditure or whether it adds additional information. Craft and Smaldone (2002) used various logistic regression models to answer these questions. Hence, we will use logistic regression models and their visualisations as kind of a benchmark for our evaluative comparison. Use of a data set from political science is motivated by the strongly theory-driven research in this field, in which data exploration prior to modelling is rather uncommon or at least rarely reported.

2 Visual support for classification

The general goal of decision trees is to provide a collection of rules that allow to classify new objects (cases) into one of a number of known classes or groups. The basic principle of tree-based methods is the hierarchical division of all observations into homogeneous subcategories. The splits that are performed while constructing the tree constitute simple criteria that can be straightforwardly used to classify new cases. Moreover, the robustness of the technique to incomplete data sets made it very popular. Common goals in generating classification trees are

- identifying variables that contribute most to the splitting and hence the characterisation of subgroups
- analyse the leaves of a tree for impurity to detect subgroups that are well and not so well modelled
- assess the overall quality of the tree model by looking at misclassification rates or other measures
- developing classification rules that can be applied to unclassified data

Visual support for algorithm selection

The first stage of visual exploration in multi-variate data analysis is typically performed using some univariate or bivariate plots of the raw data. While these plots may render a number of valuable insights to the data, little information is typically drawn to decide which modelling techniques to use or, for classification trees, which

algorithm to use. For growing a classification tree, there are mainly three families of algorithms:

- The CART family (CART, IND CART, Splus CART, etc.);
- The ML family (ID3, C4.5, C5 and other derivatives, etc.);
- The AID family (THAID, CHAID, XAID, TREEDISC, etc.).

While these families differ in the type of splitting criterion used, visualising the raw data will not be helpful to differentiate between these algorithms. The choice of algorithm will mostly depend on the context, the researcher's background and available implementations in the researcher's favourite software. Similarly, the decision whether to use a classification tree or a logistic regression model is not data-driven but a deliberate choice taken by the researcher.

Visual support for model construction

Having decided about the tree algorithm, selecting the most appropriate variables for splitting rules constitutes the prior task in the next stage of tree modelling. While many plots showing low-dimensional distributions might indicate potential subgroups in the data, the most effective way is to combine the low-dimensional displays of the predictor variables with a colouring according to the response variable. In the case of a binary response this can be easily achieved via linked highlighting (Wilhelm, 2005). While the same visualisation tools are available for guiding the early modelling stage for logistic regression, the regression framework will most likely favour plots that stress the relational structure of the data, e.g. scatter plot matrices, mosaic plots, and others.

A further aspect at this stage is to differentiate between similar models. In particular, tree based modelling more and more provides a number of competing models that are similarly efficient and valid. Differentiating between them or deciding to combine them as forests is of crucial importance.

Visual support for model evaluation

To visualise a decision tree model the standard tree plot is the most widely used representation. Their full potential requires however, an interactive environment in which the user can further manipulate the graphical representation. Such a system is, for example, provided by KLIMT (see Urbanek and Unwin, 2002). Choosing plot parameters in such a way that the terminal nodes are aligned offers an easy visual assessment of the purity of terminal nodes and hence the quality of the tree model based on misclassification rates. Treemaps (Shneiderman, 1992), spine plots of leaves and mosaic plots can be used effectfully to further study the structure of the classification tree, see for example Conversano (2011).

3 Conclusion

While classification trees and logistic regression models both perform classification tasks, each approach emphasises different aspects of analysis. Moreover, both algorithms are well supported by visualisation techniques that provide additional insight to the data analysis process. The following table summarises briefly the main findings of our evaluative comparison: In absence of general benchmarking data, the

Table 1 General analysis tasks and visual support for them.

Tasks	Exploratory Plots	Tree Plots	Tree Maps	SPLOM	Effects Plots
Algorithm selection	-	-	-	-	-
Variable Selection	+	+	0	+	
Model parameters	?	?	?	?	?
Leave impurity	-	+	+	-	-
Model Quality	-/+	+	+	-	0
Model understanding	+	+	0	0	+

evaluations are based on a specific data set. However, since the evaluation is merely based on general analysis tasks, the evaluative comparison will carry validity to a large population of classification data sets.

References

- C. Conversano. Interactive visualization in multiclass learning: integrating the SASSC algorithm with KLIMT. *Computational Statistics*, 26(4):711–731, 2011.
- C. Craft and J. P. Smaldone. The Arms Trade and the Incidence of Political Violence in sub-Saharan Africa, 1967-97. *Journal of Peace Research*, 39(6):pp. 693–710, 2002.
- Y. Liu and G. Salvendy. Design and evaluation of visualization support to facilitate decision trees classification. *Int. J. Hum.-Comput. Stud.*, 65(2):95–110, Feb. 2007.
- B. Shneiderman. Tree visualization with Treemaps: A 2D space-filling approach. *ACM Transactions on Graphics*, 11:92–99, 1992.
- A. R. Unwin, H. Hofmann, and A. F. X. Wilhelm. Direct manipulation graphics for data mining. *International Journal of Image and Graphics*, 02(01):49–65, 2002.
- S. Urbanek and A. Unwin. Making Trees Interactive with KLIMT - a COSADA Software Project. *Statistical Computing and Graphics Newsletter*, 13(1):13–16, 2002.
- A. F. Wilhelm. Interactive statistical graphics: The paradigm of linked views. In C. Rao, E. Wegman, and J. Solka, editors, *Handbook of Statistics*, volume 24, pages 437–537. Elsevier, 2005.