

Proceedings e report

114

SIS 2017
Statistics and Data Science:
new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference
of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

<http://digital.casalini.it/9788864535210>

ISBN 978-88-6453-521-0 (online)

Peer Review Process

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board

A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License
(CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/legalcode>)

CC 2017 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA

Sede: Salita de' Crescenzi 26 - 00186 Roma

Tel +39-06-6869845 - Fax +39-06-68806742

email: sis@caspur.it web:<http://www.sis-statistica.it>

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:

- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:

- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:

- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:

- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore

- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre

- Prof.ssa Francesca Bassi, Università di Padova

- Prof. Eugenio Brentari, Università di Brescia

- Dott. Stefano Falorsi, ISTAT

- Prof. Alessio Pollice, Università di Bari

- Prof.ssa Rosanna Verde, Seconda Università di Napoli

- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:

- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:

Rosanna Verde (chair), Università della Campania “Luigi Vanvitelli”
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara “G. D’Annunzio”
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:

Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:

Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Preface	XXV
Alexander Agapitov, Irina Lackman, Zoya Maksimenko <i>Determination of basis risk multiplier of a borrower default using survival analysis</i>	1
Tommaso Agasisti, Alex J. Bowers, Mara Soncin <i>School principals' leadership styles and students achievement: empirical results from a three-step Latent Class Analysis</i>	7
Tommaso Agasisti, Sergio Longobardi, Felice Russo <i>Poverty measures to analyse the educational inequality in the OECD Countries</i>	17
Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi <i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i>	23
Giacomo Aletti, Alessandra Micheletti <i>A clustering algorithm for multivariate big data with correlated components</i>	31
Emanuele Aliverti <i>A Bayesian semiparametric model for terrorist networks</i>	37

- Giorgio Alleva
Emerging challenges in official statistics: new sources, methods and skills 43
- Rémi André, Xavier Luciani and Eric Moreau
A fast algorithm for the canonical polyadic decomposition of large tensors 45
- Maria Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio
On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues 53
- Francesco Andreoli, Mauro Mussini
A spatial decomposition of the change in urban poverty concentration 59
- Margaret Antonicelli, Vito Flavio Covella
How green advertising can impact on gender different approach towards sustainability 65
- Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso
Stratified data: a permutation approach for hypotheses testing 71
- Marika Arena, Anna Calissano, Simone Vantini
Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015 79
- Maria Felice Arezzo, Giuseppina Guagnano
Using administrative data for statistical modeling: an application to tax evasion 83
- Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti
Are Numbers too Large for Kids? Possible Answers in Probable Stories 89

Index	IX
Simona Balbi, Michelangelo Misuraca, Germana Scepi <i>A polarity-based strategy for ranking social media reviews</i>	95
A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde <i>Monitoring the spatial correlation among functional data streams through Moran's Index</i>	103
Oumayma Banouar, Said Raghay <i>User query enrichment for personalized access to data through ontologies using matrix completion method</i>	109
Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, Andrea Di Lenarda <i>The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level</i>	115
Francesco Bartolucci, Stefano Peluso, Antonietta Mira <i>Marginal modeling of multilateral relational events</i>	123
Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini, Roberta Varriale <i>New Insights on Students Evaluation of Teaching in Italy</i>	129
Mauro Bernardi, Marco Bottone, Lea Petrella <i>Bayesian Quantile Regression using the Skew Exponential Power Distribution</i>	135
Mauro Bernardi <i>Bayesian Factor-Augmented Dynamic Quantile Vector Autoregression</i>	141

- Bruno Bertaccini, Giulia Biagi, Antonio Giusti, Laura Grassini
Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome 149
- Gaia Bertarelli and Franca Crippa, Fulvia Mecatti
A latent markov model approach for measuring national gender inequality 157
- Agne Bikauskaite, Dario Buono
Eurostat's methodological network: Skills mapping for a collaborative statistical office 161
- Francesco C. Billari, Emilio Zagheni
Big Data and Population Processes: A Revolution? 167
- Monica Billio, Roberto Casarin, Matteo Iacopini
Bayesian Tensor Regression models 179
- Monica Billio, Roberto Casarin, Luca Rossini
Bayesian nonparametric sparse Vector Autoregressive models 187
- Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni, Leonardo Piccini
Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area 193
- Michele Boreale, Fabio Corradi
Relative privacy risks and learning from anonymized data 199
- Giacomo Bormetti, Roberto Casarin, Fulvio Corsi, Giulia Livieri
A stochastic volatility framework with analytical filtering 205

Index	XI
Alessandro Brunetti, Stefania Fatello, Federico Polidoro <i>Estimating Italian inflation using scanner data: results and perspectives</i>	211
Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino <i>Clustering of histogram data : a topological learning approach</i>	219
Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi <i>Measuring Wellbeing by extracting Social Indicators from Big Data</i>	227
Maria Gabriella Campolo, Antonino Di Pino <i>Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples</i>	235
Stefania Capecchi, Rosaria Simone <i>Composite indicators for ordinal data: the impact of uncertainty</i>	241
Stefania Capecchi, Domenico Piccolo <i>The distribution of Net Promoter Score in socio-economic surveys</i>	247
Massimiliano Caporin, Francesco Poli <i>News, Volatility and Price Jumps</i>	253
Carmela Cappelli, Rosaria Simone, Francesca di Iorio <i>Growing happiness: a model-based tree</i>	261
Paolo Emilio Cardone <i>Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)</i>	267

- Alessandro Casa, Giovanna Menardi
Signal detection in high energy physics via a semisupervised nonparametric approach
273
- Claudio Ceccarelli, Silvia Montagna, Francesca Petrarca
Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys
279
- Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui
Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine
285
- Sana Chakri, Said Raghay, Salah El Hadaj
Contribution of extracting meaningful patterns from semantic trajectories
293
- Chieppa A., Ferrara R., Gallo G., Tomeo V.
Towards The Register-Based Statistical System: A New Valuable Source for Population Studies
301
- Shirley Coleman
Consulting, knowledge transfer and impact case studies of statistics in practice
305
- Michele Costa
The evaluation of the inequality between population subgroups
313
- Michele Costola
Bayesian Non-Negative l_1 -Regularised Regression
319
- Lisa Crosato, Caterina Liberati, Paolo Mariani, Biancamaria Zavanella
Industrial Production Index and the Web: an explorative cointegration analysis
327

Index	XIII
Francesca Romana Crucinio, Roberto Fontana <i>Comparison of conditional tests on Poisson data</i>	333
Riccardo D'Alberto, Meri Raggi <i>Non-parametric micro Statistical Matching techniques: some developments</i>	339
Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco <i>Measuring tourism from demand side</i>	345
Lucio De Capitani, Daniele De Martini <i>Optimal Ethical Balance for Phase III Trials Planning</i>	351
Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili <i>Sampling schemes using scanner data for the consumer price index</i>	357
Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone <i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i>	365
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor <i>Nonparametric classification for directional data</i>	371
Edwin Diday <i>Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce</i>	379
Carlo Drago <i>Identifying Meta Communities on Large Networks</i>	387

- Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, Mériem Jaidane
Random Forest–Based Approach for Physiological Functional Variable Selection for Drivers Stress Level Classification
393
- Silvia Facchinetti, Silvia A. Osmetti
A risk index to evaluate the criticality of a product defectiveness
399
- Federico Ferraccioli, Livio Finos
Exponential family graphical models and penalizations
405
- Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scodotto
Key–indicators for maternity hospitals and newborn readmission in Sicily
411
- Ferretti Camilla, Ganugi Piero, Zammori Francesco
Change of Variables theorem to fit Bimodal Distributions
417
- Francesco Finazzi, Lucia Paci
Space–time clustering for identifying population patterns from smartphone data
423
- Annunziata Fiore, Antonella Simone, Antonino Virgillito
IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI
429
- Michael Fop, Thomas Brendan Murphy, Luca Scrucca
Model–based Clustering with Sparse Covariance Matrices
437
- Maria Franco-Villoria, Marian Scott
Quantile Regression for Functional Data
441

Index	XV
Gallo M., Simonacci V., Di Palma M.A. <i>Three-way compositional data: a multi-stage trilinear decomposition algorithm</i>	445
Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples <i>Nonparametric shared frailty model for classification of survival data</i>	451
Stefano A. Gattone, Angela De Sanctis <i>Clustering landmark-based shapes using Information Geometry tools</i>	457
Alan E. Gelfand, Shinichiro Shirota <i>Space and circular time log Gaussian Cox processes with application to crime event data</i>	461
Abdelghani Ghazdali <i>Blind source separation</i>	469
Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarcangelo <i>An innovative approach for Opinion Mining : the Plutchick analysis</i>	479
Massimiliano Giacalone, Demetrio Panarello <i>A G.E.D. method for market risk evaluation using a modified Gaussian Copula</i>	485
Chiara Gigliarano, Francesco Maria Chelli <i>Labour market dynamics and recent economic changes: the case of Italy</i>	491
Giuseppe Giordano, Giancarlo Ragozini, Maria Prosperina Vitale <i>On the use of DISTATIS to handle multiplex networks</i>	499

- Michela Gnaldi, Silvia Bacci, Samuel Greiff, Thiemo Kunze
Profiles of students on account of complex problem solving (CPS) strategies exploited via log–data
505
- Michela Gnaldi, Simone Del Sarto
Characterising Italian municipalities according to the annual report of the prevention–of–corruption supervisor: a Latent Class approach
513
- Silvia Golia
A proposal of a discretization method applicable to Rasch measures
519
- Anna Gottard
Tree–based Non–linear Graphical Models
525
- Sara Hbali, Youssef Hbali, Mohamed Sadgal, Abdelaziz El Fazziki
Sentiment Analysis for micro–blogging using LSTM Recurrent Neural Networks
531
- Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, Elena Siletti
How to Exploit Big Data from Social Networks: a Subjective Well–being Indicator via Twitter
537
- Francesca Ieva
Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data
543
- Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde
Automatic variable and components weighting systems for Fuzzy cmeans of distributional data
549
- Michael Jauch, Paolo Giordani, David Dunson
A Bayesian oblique factor model with extension to tensor data
553

Index	XVII
Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins <i>Statistical analysis for partially observed multilayered networks</i>	561
Francesco Lagona <i>Copula-based segmentation of environmental time series with linear and circular components</i>	569
Alessandro Lanteri, Mauro Maggioni <i>A Multiscale Approach to Manifold Estimation</i>	575
Tiziana Laureti, Carlo Ferrante, Barbara Dramis <i>Using scanner and CPI data to estimate Italian sub-national PPPs</i>	581
Antonio Lepore <i>Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters</i>	589
Antonio Lepore, Biagio Palumbo, Christian Capezza <i>Monitoring ship performance via multi-way partial least-squares analysis of functional data</i>	595
Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella <i>Dynamic profiling of banking customers: a pseudo-panel study</i>	601
Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis <i>A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns</i>	607
Rosaria Lombardo, Eric J Beh <i>Three-way Correspondence Analysis for Ordinal-Nominal Variables</i>	613

- Monia Lupparelli, Alessandra Mattei
Log–mean linear models for causal inference
621
- Badiaa Lyoussi, Zineb Selihi, Mohamed Berraho, Karima El Rhazi, Youness El Achhab, Adiba El Marrakchi , Chakib Nejjari
Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study
627
- Valentina Mamei, Debora Slanzi, Irene Poli
Bootstrap group penalty for high-dimensional regression models
633
- Stefano Marchetti, Monica Pratesi, Caterina Giusti
Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data
639
- Paolo Mariani, Andrea Marletta, Mariangela Zenga
Gross Annual Salary of a new graduate: is it a question of profile?
647
- Maria Francesca Marino, Marco Alfò
Dynamic random coefficient based drop-out models for longitudinal responses
653
- Antonello Maruotti, Jan Bulla
Hidden Markov models: dimensionality reduction, atypical observations and algorithms
659
- Chiara Masci, Geraint Johnes, Tommaso Agasisti
A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting
667

Index	XIX
Lucio Masserini, Matilde Bini <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i>	673
Angelo Mazza, Antonio Punzo, Salvatore Ingrassia <i>An R Package for Cluster-Weighted Models</i>	681
Antonino Mazzeo, Flora Amato <i>Methods and applications for the treatment of Big Data in strategic fields</i>	687
Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis <i>Happy parents' tweets</i>	693
Rodolfo Metulini, Marica Manisera, Paola Zuccolotto <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i>	701
Giorgio E. Montanari, Marco Doretto, Francesco Bartolucci <i>An ordinal Latent Markov model for the evaluation of health care services</i>	707
Isabella Morlini, Maristella Scorza <i>New fuzzy composite indicators for dyslexia</i>	713
Fionn Murtagh <i>Big Textual Data: Lessons and Challenges for Statistics</i>	719
Gaetano Musella, Gennaro Punzo <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i>	731

XX	Index
Marta Nai Ruscone <i>Exploratory factor analysis of ordinal variables: a copula approach</i>	737
Fausta Ongaro, Silvana Salvini <i>IPUMS Data for describing family and household structures in the world</i>	743
Tullia Padellini, Pierpaolo Brutti <i>Topological Summaries for Time-Varying Data</i>	747
Sally Paganin <i>Modeling of Complex Network Data for Targeted Marketing</i>	753
Francesco Palumbo, Giancarlo Ragozini <i>Statistical categorization through archetypal analysis</i>	759
Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier <i>Inference with the Unscented Kalman Filter and optimization of sigma points</i>	767
Xanthi Pedeli, Cristiano Varin <i>Pairwise Likelihood Inference for Parameter-Driven Models</i>	773
Felicia Pelagalli, Francesca Greco, Enrico De Santis <i>Social emotional data analysis. The map of Europe</i>	779
Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti <i>Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles</i>	785
Alessia Pini, Aymeric Stamm, Simone Vantini <i>Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces</i>	791

Index	XXI
Silvia Poletti, Serena Arima <i>Accounting for measurement error in small area models: a study on generosity</i>	795
Gennaro Punzo, Mariateresa Ciommi <i>Structural changes in the employment composition and wage inequality: A comparison across European countries</i>	801
Walter J. Radermacher <i>Official Statistics 4.0 – learning from history for the challenges of the future</i>	809
Fabio Rapallo <i>Comparison of contingency tables under quasi-symmetry</i>	821
Valentina Raponi, Cesare Robotti, Paolo Zaffaroni <i>Testing Beta-Pricing Models Using Large Cross-Sections</i>	827
Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza <i>On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data</i>	833
Alessandra Righi, Mauro Mario Gentile <i>Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics</i>	841
Paolo Righi, Giulio Barcaroli, Natalia Golini <i>Quality issues when using Big Data in Official Statistics</i>	847
Emilia Rocco <i>Indicators for the representativeness of survey response as well as convenience samples</i>	855

- Emilia Rocco, Bruno Bertaccini, Giulia Biagi, Andrea Giommi
A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence
861
- Elvira Romano, Jorge Mateu
A local regression technique for spatially dependent functional data: an heteroskedastic GWR model
867
- Eduardo Rossi, Paolo Santucci de Magistris
Models for jumps in trading volume
873
- Renata Rotondi, Elisa Varini
On a failure process driven by a self-correcting model in seismic hazard assessment
879
- M. Ruggieri, F. Di Salvo and A. Plaia
Functional principal component analysis of quantile curves
887
- Massimiliano Russo
Detecting group differences in multivariate categorical data
893
- Michele Scagliarini
A Sequential Test for the C_{pk} Index
899
- Steven L. Scott
Industrial Applications of Bayesian Structural Time Series
905
- Catia Scricciolo
Asymptotically Efficient Estimation in Measurement Error Models
913

Index	XXIII
Angela Serra, Pietro Coretto, Roberto Tagliaferri <i>On the noisy high-dimensional gene expression data analysis</i>	919
Mirko Signorelli <i>Variable selection for (realistic) stochastic blockmodels</i>	927
Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio <i>Detection of spatio-temporal local structure on seismic data</i>	935
A. Sottosanti, D. Bastieri, A. R. Brazzale <i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i>	943
Federico Mattia Stefanini <i>Causal analysis of Cell Transformation Assays</i>	949
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i>	955
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Sparse Indirect Inference</i>	961
Peter Struijs, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, Nigel Swier <i>The ESSnet Big Data: Experimental Results</i>	969
Jérémie Sublime <i>Smart view selection in multi-view clustering</i>	977

- Emilio Sulis
Social Sensing and Official Statistics: call data records and social media sentiment analysis
985
- Matilde Trevisani, Arjuna Tuzzi
Knowledge mapping by a functional data analysis of scientific articles databases
993
- Amalia Vanacore, Maria Sole Pellegrino
Characterizing the extent of rater agreement via a non-parametric benchmarking procedure
999
- Maarten Vanhoof, Stephanie Combes, Marie-Pierre de Bellefon
Mining Mobile Phone Data to Detect Urban Areas
1005
- Viktoriya Voytsekhovska, Olivier Butzbach
Statistical methods in assessing the equality of income distribution, case study of Poland
1013
- Ernst C. Wit
Network inference in Genomics
1019
- Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland
Using Twitter data for Population Estimates
1025
- Marco Seabra dos Rei
Structured Approaches for High-Dimensional Predictive Modeling
1033

Exploratory factor analysis of ordinal variables: a copula approach

Analisi fattoriale esplorativa di variabili ordinali: un approccio via copula

Marta Nai Ruscone

Abstract Exploratory factor analysis attempts to identify the underlying factors that explain the pattern of correlations within a set of observed variables. The analysis is almost always performed with Pearson's correlations even when the data are ordinal, but this is not appropriate since they are not quantitative data. The use of Likert scales is increasingly common in the field of social research, so it is necessary to determine which methodology is the most suitable for analysing the data obtained as non quantitative measures. In this context, also by means of simulation studies, we aim to illustrate the advantages of using Spearman's grade correlation coefficient on a transformation operated by the copula function in order to perform exploratory factor analysis of ordinal variables. Moreover, by using the copula, we consider the general dependence structure, providing a more robust reproduction of the measurement model.

Abstract *L'analisi fattoriale esplorativa vuole identificare i fattori latenti che spiegano un insieme di variabili osservate. L'analisi quasi sempre utilizza la correlazione di Pearson, anche quando i dati sono di natura ordinale, ma questo non è appropriato in quanto questi dati non sono quantitativi. L'uso di scale Likert è sempre più comune nel campo della ricerca sociale, risulta quindi necessario determinare quale metodo risulta essere più idoneo per l'analisi di tali dati tenendo presente che spesso vengono analizzati utilizzando tecniche idonee solo per misure quantitative. In questo contesto, e mediante studi di simulazione, si illustrano i vantaggi nell'utilizzo dello Spearman grade correlation ottenuto mediante l'utilizzo dalla funzione copula anziché della correlazione di Pearson. Con l'utilizzo della copula, si considera così la struttura di dipendenza generale, fornendo così una misurazione più accurata*

Key words: Factor analysis, copula, ordinal variables, Likert scales, correlation

Marta Nai Ruscone

School of Economics and Management - LIUC - University Cattaneo, C.so Matteotti 22 - 21053 Castellanza (VA), Italy, e-mail: mnairuscone@liuc.it

1 Introduction

Exploratory factor analysis is a widely used statistical technique in the social sciences where the main interest lies in measuring the unobserved construct, such as emotions, attitudes, beliefs and behaviors. The main idea behind the analysis is that the latent variables (also named factors) account for the dependencies among the observed variables (also named items or indicators) in the sense that if the factors are held fixed, the observed variables would be independent. In exploratory factor analysis the goal is the following: for a given set of observed variables x_1, \dots, x_p one wants to find a set of latent factors ξ_1, \dots, ξ_k , fewer in number than the observed variables ($k < p$), that contain essentially the same information. In its classical formulation [1], it concerns a set of continuous variables measured on a set of independent units. The data usually encountered in social sciences are of categorical nature (ordinal or nominal). The Likert Rating Scale [10], [11] is a simple procedure for generating measurement instruments which is widely used by social scientists to measure a variety of latent constructs, and meticulous statistical procedures have therefore been developed to design and validate these scales [3], [15]. However, most of these ignore the ordinal nature of observed responses and assume the presence of continuous observed variables measured at interval level. Evidence shows that, under relatively common circumstances, classical factor analysis (FA) yields inaccurate results characterizing the internal structure of the scale or selecting the most informative items within each factor [4], [7].

In the present work Spearman's grade correlation coefficient on a transformation operated by the copula function is employed, in order to take into account the ordinal nature of the data. The copula is a helpful tool for handling multivariate continuous distributions with given univariate marginals [14]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures, different from the linear one, that characterises the multivariate normal distribution.

So taking into account that the use of measurement instruments which require categorical responses from subjects is increasingly common in social research, and this implies the use of ordinal scales, the present work aims to point out a correct definition of dependence measure for ordinal variables rather than the Pearson correlation coefficient correctly applied to quantitative variables. Moreover, the use of several copulae with specific tail dependence allow us to obtain an index that weights the ordinal variables categories in several ways. In so doing we can address and recognize the ordinal nature of observed variables and estimate that weight directly from the data.

2 The copula function

The copula function is the key ingredient for handling multivariate continuous distributions with given univariate marginals. We will discuss this issue briefly below,

for further details and proofs, see for instance [14], [8] and [2]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures different from the linear one that characterises the multivariate normal distribution.

A bivariate copula $C : I^2 \rightarrow I$, with $I^2 = [0, 1] \times [0, 1]$ and $I = [0, 1]$, is the cumulative bivariate distribution function of a random variable (U_1, U_2) with uniform marginal random variables in $[0, 1]$

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_2 \leq u_2; \theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \quad (1)$$

where θ is a parameter measuring the dependence between U_1 and U_2 .

The following theorem by Sklar [14] explains the use of the copula in the characterization of a joint distribution. Let (X_1, X_2) be a bivariate random variable with marginal cdfs $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ and joint cdf $F_{X_1, X_2}(x_1, x_2; \theta)$, then there is always a **copula function** $C(\cdot, \cdot; \theta)$ with $C : I^2 \rightarrow I$ such that

$$F_{X_1, X_2}(x_1, x_2; \theta) = C(F_{X_1}(x_1), F_{X_2}(x_2); \theta), \quad x_1, x_2 \in \mathbb{R}. \quad (2)$$

Conversely, if $C(\cdot, \cdot; \theta)$ is a copula function and $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are marginal cdfs, then $F_{X_1, X_2}(x_1, x_2; \theta)$ is a joint cdf.

If $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are **continuous** functions then the copula $C(\cdot, \cdot; \theta)$ is **unique**. Moreover, if $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are continuous the copula can be found by the inverse of (2):

$$C(u_1, u_2) = F_{X_1, X_2}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)), \quad (3)$$

with $u_1 = F_{X_1}(x_1)$ and $u_2 = F_{X_2}(x_2)$. This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The **dependence structure** is explained by the copula function $C(\cdot, \cdot; \theta)$. Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the multivariate normal distribution.

Each copula is related to the most important measures of dependence: the Pearson correlation coefficient, the Spearman grade correlation coefficient and tail dependence parameters. The Spearman grade correlation coefficient (see [14] pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measure the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ , then the Spearman grade correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3 = \frac{\text{Cov}(U_1, U_2)}{\sqrt{\text{Var}(U_1)} \sqrt{\text{Var}(U_2)}}. \quad (4)$$

For continuous random variables this is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [6]. Among all copulas $C: I^2 \rightarrow I$ such that for every $u, v \in I$, three especially noteworthy ones are $W(u, v) = \max(u + v - 1, 0)$, $\Pi(u, v) = uv$, and $M(u, v) = \min(u, v)$. These copulae correspond to perfect negative association ($\rho_S(C) = -1$), independence ($\rho_S(C) = 0$), and perfect positive association ($\rho_S(C) = +1$) between the two random variables, respectively. For all $(u, v) \in I^2$ it holds that $W(u, v) \leq \Pi(u, v) \leq M(u, v)$.

The tail dependence relationship can be measured by means of the upper and lower tail dependence parameters

$$\lambda_u = \lim_{u \rightarrow 1^-} P[X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)] = \lim_{u \rightarrow 1^-} \frac{C(u, u)}{u}, \quad (5)$$

$$\lambda_l = \lim_{u \rightarrow 0^+} P[X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)] = \lim_{u \rightarrow 0^+} \frac{1 - 2u + C(u, u)}{1 - u}. \quad (6)$$

If $\lambda_u \in (0, 1]$ or $\lambda_l \in (0, 1]$, the random variables X_1 and X_2 present upper or lower tail dependence. If $\lambda_u = 0$ or $\lambda_l = 0$, there is no upper or lower tail dependence. These parameters measure the dependence in the tails of the joint distribution, i.e. high/low values of one variable are associated with high/low values of the other one. They represent the probability that one variable is extreme given that the other is extreme. The Spearman grade correlation coefficient and both tail dependence parameters are directly associated with the parameters of some copula family [14].

3 Our proposal

Theory and methodology for exploratory factor analysis have been well developed for continuous variables, but in practice observed or measured variables are often ordinal.

Observations on an ordinal variable are assumed to have logical ordering categories. This logical ordering is typical when data are collected from questionnaires. A good example is the Likert Scale that is frequently used in survey research: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, and 5 = *Strongly agree*. Although a question is designed to measure a theoretical concept, the observed responses are only a discrete realization of a small number of categories and distances between categories are unknown. Following [13], [9] and others, it is assumed that there is a continuous variable x_{i*} underlying the ordinal variable x_i , $i = 1, \dots, p$. This continuous variable x_{i*} represents the attitude underlying the order responses to x_i and it is assumed to have a range from $-\infty$ to $+\infty$.

The underlying variable x_{i*} is unobservable. Only the ordinal variable x_i is observed. For an ordinal variable x_i with m_i categories, the connection between the ordinal variable x_i and the underlying variable x_{i*} is:

$$x_i \Leftrightarrow \tau_{i-1}^i < x_i^* < \tau_i^i, \quad i = 1, 2, \dots, m_i \tag{7}$$

where

$$-\infty = \tau_0^i < \tau_1^i < \tau_2^i < \dots < \tau_{m_i-1}^i < \tau_{m_i}^i = +\infty \tag{8}$$

are threshold parameters. For variable x_i with m_i categories, there are $m_i - 1$ strictly increasing threshold parameters $\tau_1^i < \tau_2^i < \dots < \tau_{m_i-1}^i$.

Let x_i and x_j be the two ordinal variables with m_i and m_j categories respectively. We define now Spearman's grade correlation via copula. We consider a copula C_θ associated with each pair (X_i^*, X_j^*) underlying the pair (X_i, X_j) in the set of ordinal items X_1, X_2, \dots, X_i , we thus assume that each pair (X_i, X_j) corresponds to a bivariate discrete random variable obtained by a discretisation of a bivariate continuous latent variable $U_i = F(X_i^*), U_j = F(X_j^*)$ with support on the unit interval.

Let $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$, $i = 1, 2, \dots, m_i$ $j = 1, 2, \dots, m_j$, be the rectangles defining the discretisation. Let $p_{11}, \dots, p_{m_i m_j}$ be the joint probabilities of the ordinal variables corresponding to the rectangles $A_{11}, \dots, A_{m_i m_j}$. Let $V_{C_\theta}(A_{11}, \dots, A_{m_i m_j})$ be the volumes of the rectangles under the copula C_θ , then

$$V_{C_\theta}(A_{11}, \dots, A_{m_i m_j}) = p_{11}, \dots, p_{m_i m_j} \tag{9}$$

There exists a unique element in the family of copula for which (9) holds true. We apply this to each pair (X_i, X_j) $i \neq j$ in the set of the items. θ can be estimated via maximum likelihood [5] [12]. The multivariate normality assumption pertaining to the underlying variables, assumed by polychoric correlation and Pearson correlation, is relaxed. To apply the index one needs only to specify the dependence structure of the variables by means of a copula family.

In this way the construct validity is analysed according to ordinal data obtained from Likert scales using the most suitable method. The factor results show a better fit to the theoretical model when the factorization is carried out using the Spearman's grade correlation via copula rather than Pearson correlation. Our focus here has been to identify the type of correlation that yields a factor solution more in keeping with the original measurement model, as we believe this to have great importances in terms of drawing correct substantive conclusions. When we conduct a FA our results can be summarized as follow:

- regardless of the number of dimensions and items with skewness, Pearson correlations are lower than Spearman's grade correlations. The results are more significant when all items are asymmetric.
- The model obtained is more consistent with the original measurement model when we factorize using the Spearman's grade correlation. This result does not depend on the number of dimensions and asymmetric items.

To summarize the factor results obtained when we use Spearman's grade correlation better reproduce the measurement model present in the data, regardless of the number of factors.

References

1. Anderson, T.W.: An introduction to multivariate statistical analysis. Wiley, New York (2003)
2. Cherubini, U., Luciano, E., Vecchiato, W.: Copula methods in finance. John Wiley & Sons (2004)
3. DeVellis, R.F.: Scale development, theory and applications. Sage, Newbury Park (1991)
4. DiStefano, C.: The impact of categorization with confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal* **9**, 327–346 (2002)
5. Ekstrom, J.: Contributions to the Theory of measures of association for ordinal variables. *Digital Comprehensive Summaries of Uppsala Dissertation from the Faculty of Social Sciences*, Uppsala (2003)
6. Embrechts, P., McNeil, A., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 176–223 (2002)
7. Holgado-Tello, F.P., Chacón-Moscoso, S., Barbero-García, I., Vila-Abad E.: Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity* **44**, 153–166 (2010)
8. Joe, H.: Multivariate models and multivariate dependence concepts. CRC Press (1997)
9. 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)
10. Likert, R.: A technique for the measurement of attitudes. *Archives of Psychology*, 44–45 (1932)
11. Likert, R., Sydney, R., Murphy, G.: A simple and reliable method of scoring Thurstone attitudes scales. *The Journal of Social Psychology*, 228–238 (1934)
12. Martinson, E.O., Hamdan, M.A.: Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *J. Stat. Comput. Simul.* **1**, 45–54 (1971)
13. Muthén, B. O.: Full maximum likelihood analysis of structural equation models with polytomous variables. *Psychometrika* **9(1)**, 91–97 (1984)
14. Nelsen, R.B.: An introduction to copulas. Springer Science & Business Media (2013)
15. Spector, P.E.: Summating rating scale construction: an introduction. Sage, Newbury Park (1992)