



UNIVERSITÀ DI PISA



Sant'Anna
Scuola Universitaria Superiore Pisa



Consiglio Nazionale delle Ricerche

Book of Short Papers

SIS 2020



Società
Italiana di
Statistica

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Copyright © 2020

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891910776

Contents

Specialized sessions

Accounting for record linkage errors in inference (S2G-SIS).....	2
Probabilistic record linkage with less than three matching variables.	3
<i>Tiziana Tuoto and Marco Fortini</i>	
Advanced methods for measuring and communicating uncertainty in official statistics	9
A model for measuring the accuracy in spatial price statistics using scanner data.	10
<i>Ilaria Benedetti and Federico Crescenzi</i>	
Communication of Uncertainty of Official Statistics.	16
<i>Edwin de Jonge and Gian Luigi Mazzi</i>	
Measuring uncertainty for infra-annual macroeconomic statistics.	22
<i>George Kapetanios, Massimiliano Marcellino and Gian Luigi Mazzi</i>	
Bayesian methods in biostatistics	27
Network Estimation of Compositional Data.	28
<i>Nathan Osborne, Christine B. Peterson and Marina Vannucci</i>	
Using co-data to empower genomics-based prediction and variable selection.	34
<i>Magnus M. Münch, Mirrelijin M. van Nee and Mark A. van de Wiel</i>	
Data integration versus privacy protection: a methodological challenge?	40
Statistical Disclosure Control for Integrated Data.	41
<i>Natalie Shlomo</i>	
The Integrated System of Statistic Registers: first steps towards facing privacy issues.	47
<i>Mauro Bruno and Roberta Radini</i>	
Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics.	53
<i>Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann and Martina Hahn</i>	
Designing adaptive clinical trials	59
Optimal designs for multi-arm exponential trials.	60
<i>Rosamarie Frieri and Marco Novelli</i>	
Education: students' mobility and labour market.....	66
From measurement to explanatory approaches: an assessment of the attractiveness of the curricula programs supplied by Italian universities.	67
<i>Isabella Sullis, Silvia Columbu and Mariano Porcu</i>	
Pull factors for university students' mobility: a gravity model approach.	73
<i>Giovanni Boscaino and Vincenzo Giuseppe Genova</i>	
Spatial autoregressive gravity models to explain the university student mobility in Italy.	79
<i>Silvia Bacci, Bruno Bertaccini and Chiara Bocci</i>	

Environmental Statistics (GRASPA-SIS)	85
A Time Clustering Model for Spatio-Temporal Data. <i>Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona</i>	86
Reconstruction of sparsely sampled functional time series using frequency domain functional principal components. <i>Amira Elayouty, Marian Scott and Claire Miller</i>	93
Methods for High Dimensional Compositional Data Analysis	98
Algorithms for compositional tensors of third-order. <i>Violetta Simonacci</i>	99
High-dimensional regression with compositional covariates: a robust perspective. <i>Gianna Serafina Monti and Peter Filzmoser</i>	105
Three-way compositional analysis of energy intensity in manufacturing. <i>Valentin Todorov and Violetta Simonacci</i>	111
Modern Statistics for Physics Discoveries	117
Identification of high-energy λ -ray sources via nonparametric clustering. <i>Giovanna Menardi, Denise Costantin, and Federico Ferraccioli</i>	118
Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws. <i>Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio and Nicoletta D'Angelo</i>	124
Network Modelling in Biostatistics.....	130
Natural direct and indirect relative risk for mediation analysis. <i>Monia Lupparelli and Alessandra Mattei</i>	131
New issues on multivariate and univariate quantile regression	137
Mixtures of quantile regressions for longitudinal data: an R package. <i>Maria Francesca Marino, Maria Giovanna Ranalli and Marco Alfò</i>	138
Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles. <i>Luca Merlo, Lea Petrella and Nikos Tzavidis</i>	144
Recent methodological advances in finite mixture modeling with applications (CLADAG-SIS)	150
Aggregating Gaussian mixture components. <i>Roberto Rocci</i>	151
Local and overall coefficients of determination for mixtures of generalized linear models. <i>Roberto Di Mari, Salvatore Ingrassia and Antonio Punzo</i>	157
Statistical Analysis of Satellite Data (SDS-SIS)	163
Functional Data Analysis for Interferometric Synthetic Aperture Radar Data Post-Processing: The case of Santa Barbara mud volcano. <i>Matteo Fontana, Alessandra Menafoglio, Francesca Cigna and Deodato Tapete</i>	164
Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements. <i>Alessandro Fassò</i>	170
Statistical models and methods for Business and Industry	176
Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures. <i>Bianca Maria Colosimo, Marco Grasso and Federica Garghetti</i>	177
Open data powered territorial planning - Case study: The Turin historical center. <i>Silvia Casagrande, Gianmaria Origi, Alberto Pasanisi, Martina Tamburini, Pascal Terrien, Tania Cerquitelli and Alfonso Capozzoli</i>	183
Process optimization in Industry 4.0: Are all data analytics models useful? <i>Alberto Ferrer</i>	189

Technology and demographic behaviours (AISP-SIS)	195
Internet and the Timing of Births.	196
<i>Maria Sironi, Osea Giuntella and Francesco C. Billari</i>	
The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating.	202
<i>Francesco C. Billari, Osea Giuntella and Luca Stella</i>	

Solicited Sessions

Advanced Statistical Methods in Health Analytics	209
Assessing the impact of the intermediate event in a non-markovian illness-death model.	210
<i>Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi and Laura Antolini</i>	
Big data and AI: challenges and opportunities in healthcare.	216
<i>Vieri Emiliani, Gian Luca Cattani and Fabrizio Selmi</i>	
Statistical methodology for volume-outcome studies.	222
<i>Marta Fiocco and Floor van Oudenhoven</i>	
Advances in textual data mining	228
Distance measures for exploring pairs of novels in a large corpus of Italian literature.	229
<i>Matilde Trevisani and Arjuna Tuzzi</i>	
Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics.	235
<i>Mariangela Sciandra, Alessandro Albano and Irene Carola Spera</i>	
Advances in the interaction between artificial intelligence and official statistics	241
Automated Land Cover Maps from Satellite Imagery by Deep Learning.	242
<i>Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto</i>	
CROWD4SDG: Crowdsourcing for sustainable developments goals.	248
<i>Barbara Pernici</i>	
Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency. The direct experience of Tuscany.	253
<i>Linda Porciani, Luisa Francovich, Luca Faustini and Alessandro Valentini</i>	
Capture-recapture methods	259
Bayesian Model Averaging for Latent Class Models in Capture-Recapture.	260
<i>Davide Di Cecco</i>	
Combining "signs of life" and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts.	266
<i>Marco Fortini, Antonella Bernardini, Marco Caputi and Nicoletta Cibella</i>	
Population size estimation with interval censored counts and external information.	272
<i>Alessio Farcomeni</i>	
Changes in environment extremes and their impacts	278
FPCA Clustering of rainfall events.	279
<i>Gianluca Sottile, Antonio Francipane, Leonardo Noto and Giada Adelfio</i>	
Trends in rainfall extremes in the Venice lagoon catchment.	285
<i>Ilaria Prodocimi and Carlo Gaetan</i>	

Copulas: models and inference	291
Analysis of district heating demand through different copula-based approaches. <i>F. Marta L. Di Lascio and Andrea Menapace</i>	292
CoVaR and backtesting: a comparison between a copula approach and parametric models. <i>Michele Leonardo Bianchi, Giovanni De Luca and Giorgia Riveccio</i>	298
Estimating Asymmetric Dependence via Empirical Checkerboard Copulas. <i>Wolfgang Trutschnig and Florian Griessenberger</i>	304
Strong Convergence of Multivariate Maxima. <i>Michael Falk, Simone A. Padoan and Stefano Rizzelli</i>	310
Data Science: when different expertise meet	316
Bayesian stochastic modelling of the temporal evolution of seismicity. <i>Elisa Varini and Renata Rotondi</i>	317
Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field. <i>Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani</i>	323
Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics. <i>Andrea Serani, Matteo Diez and Frederick Stern</i>	328
Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method. <i>Lorenzo Tamellini and Joakim Beck</i>	334
Emerging challenges in official statistics: new data sources and methods	340
Small area poverty indicators adjusted using local spatial price indices. <i>Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi</i>	341
Smart solutions for trusted smart statistics: the European big data hackathon experience. <i>Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti and Manuela Michelini</i>	347
The ESSnet Project Smart Surveys: new data sources and tools for Surveys of Official Statistics	353
Factorial and dimensional reduction methods for the construction of indicators for evaluation (SVQS-SIS).....	359
A comparison of MBC with CLV and PCovR methods for dimensional reduction of the soccer players' performance attributes. <i>Maurizio Carpita, Enrico Ciavolino and Paola Pasca</i>	360
A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties. <i>Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra</i>	366
Exploring drug consumption via an ultrametric correlation matrix. <i>Giorgia Zaccaria and Maurizio Vichi</i>	372
Ranking extraction in ordinal multi-indicator systems. <i>Marco Fattore and Alberto Arcagni</i>	378
Gender statistics	384
Gender differences in Italian STEM degree courses: a discrete-time competing-risks model. <i>Marco Enea and Massimo Attanasio</i>	385
Some Challenges and Results in Measuring Gender Inequality. <i>Fabio Crescenzi and Francesco Di Pede</i>	391

How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)..	397
A modal approach for clustering matrices.	398
<i>Federico Ferraccioli and Giovanna Menardi</i>	
A Note on Detection of Perturbations in Biological Networks.	404
<i>Vera Djordjilović</i>	
Bayesian inference for DAG-probit models.	410
<i>Federico Castelletti</i>	
Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process.	416
<i>Sirio Legramanti</i>	
Measuring poverty and vulnerability	421
Choosing the vulnerability threshold using the ROC curve.	422
<i>Chiara Gigliarano and Conchita D'Ambrosio</i>	
New advances in applications, a Bayesian nonparametric perspective	428
Bayesian Mixture Models for Latent Class Analysis.	429
<i>Raffaele Argiento, Bruno Bodin and Maria De Iorio</i>	
Non-Parametric Inference and Forecasting of Functional and Object Data	435
An interpretable estimator for the function-on-function linear regression model with application to the Canadian weather data.	436
<i>Fabio Centofanti and Matteo Fontana</i>	
Statistical process monitoring of multivariate profiles from ship operating conditions.	440
<i>Christian Capezza</i>	
Prior choice in Bayesian Modelling (SISbayes)	446
Bayesian Learning of Multiple Essential Graphs.	447
<i>Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni</i>	
Bayesian post-processing of Gibbs sampling output for variable selection.	453
<i>Stefano Cabras</i>	
Priors on precision parameters of IGRMF models.	459
<i>Aldo Gardini, Fedele Greco and Carlo Trivisano</i>	
Sequence Analysis: methods and applications	465
Internal migration, family formation and social stratification in Europe. A life course approach.	466
<i>Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella</i>	
Socio economic integration of migrants	472
A study on the characteristics of spouses who intermarry in Italy.	473
<i>Agnese Vitali and Romina Fraboni</i>	
Statistical Analysis for mobility and transportation	479
A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree.	480
<i>Silvia Columbu and Ilaria Primerano</i>	
Analysis of mobility data through a novel Cheng and Church algorithm for functional data.	486
<i>Marta Galvani, Agostino Torti and Alessandra Menafoglio</i>	
Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy.	491
<i>Agostino Torti, Marika Arena, Giovanni Azzone, and Piercesare Secchi</i>	

Statistical Methods and Applications in Social Network Analysis	496
A clustering procedure for ego-networks data: an application to Italian elders living in couple. <i>Elvira Pelle and Roberta Pappadà</i>	497
Analysing the mediating role of a network: a Bayesian latent space approach. <i>Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison</i>	503
Network-time autoregressive models for valued network panel. <i>Viviana Amati</i>	509
University student mobility flows and network data structures. <i>Maria Prosperina Vitale, Giuseppe Giordano and Giancarlo Ragozini</i>	515
Statistical Methods in Psychometrics	521
A simple probabilistic model to evaluate questionable interim analysis strategies. <i>Francesca Freuli and Luigi Lombardi</i>	522
Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research. <i>Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore</i>	528
Predicting social media addiction from Instagram profiles: A data mining approach. <i>Antonio Calcagni, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale</i>	534
Structural entropy based modeling for psychological measurement. <i>Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo</i>	540
Statistical modelling in environmental epidemiology	546
A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health. <i>Pasquale Valentini, Luigi Ippoliti and Clara Grazian</i>	547
Joint Analysis of Short and Long-Term Effects of Air Pollution. <i>Annibale Biggeri, Dolores Catelan, Giorgia Stoppa and Corrado Lagazio</i>	551
Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation	557
DNA mixtures with related contributors. <i>Peter J. Green and Julia Mortera</i>	558
Forensic Statistics: How to estimate life expectancy after injury. <i>Jane L Hutton</i>	564
The additional contribution of combining genetic evidence from multiple samples in a complex case. <i>Giampietro Lago</i>	570
The history of forensic inference and statistics: a thematic perspective. <i>Franco Taroni and Colin Aitken</i>	576
Topological learning: interpretable representations of complex data.....	581
Comparing Neural Networks via Generalized Persistence. <i>Mattia G. Bergomi and Pietro Vertechi</i>	582
On the topological complexity of decision boundaries. <i>António Leitão and Giovanni Petri</i>	588
Persistence-based Kernels for Data Classification. <i>Ulderico Fugacci</i>	594
Topological and Mixed-type learning of Brain Activity. <i>Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei</i>	600

Contributed papers and Posters

Bayesian Statistics	607
A Bayesian approach for modelling dependence among mixture densities. <i>Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi</i>	608
A change of glasses strategy to solve the rare type match problem. <i>Giulia Cereda and Fabio Corradi</i>	614
A new prior distribution on the simplex: the extended flexible Dirichlet. <i>Roberto Ascari, Sonia Migliorati and Andrea Ongaro</i>	620
ABC model choice via mixture weight estimation. <i>Gianmarco Caruso, Luca Tardella and Christian P. Robert</i>	626
An ABC algorithm for random partitions arising from the Dirichlet process. <i>Mario Beraha and Riccardo Corradin</i>	632
Bayesian Inference of Undirected Graphical Models from Count Data. <i>Pier Giovanni Bissiri, Monica Chiogna and Nguyen Thi Kim Hue</i>	638
Bayesian IRT models in NIMBLE. <i>Sally Paganin, Chris Paciorek and Perry de Valpine</i>	644
Bayesian modelling of Facebook communities via latent factor models. <i>Emanuele Aliverti</i>	650
Bayesian nonparametric adaptive classification with robust prior information. <i>Francesco Denti, Andrea Cappozzo and Francesca Greselin</i>	655
Choosing the right tool for the job: a systematic analysis of general purpose MCMC software. <i>Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi</i>	661
Empirical Bayes estimation for mixture models. <i>Catia Scricciolo</i>	667
Improving ABC via Large Deviations Theory. <i>Cecilia Viscardi, Michele Boreale and Fabio Corradi</i>	673
Learning Bayesian Networks for Nonparanormal Data. <i>Flaminia Musella and Vincenzina Vitale</i>	679
Measuring well-being combining different data sources: a Bayesian networks approach. <i>Federica Cugnata, Silvia Salini and Elena Siletti</i>	685
Penalising the complexity of extensions of the Gaussian distribution. <i>Diego Battagliese and Brunero Liseo</i>	691
Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution. <i>Fulvio De Santis and Stefania Gubbiotti</i>	697
Small-area statistical estimation of claim risk. <i>Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo</i>	702
Subject-specific Bayesian Hierarchical model for compositional data analysis. <i>Matteo Pedone and Francesco C. Stingo</i>	708
Wasserstein consensus for Bayesian sample size determination. <i>Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti</i>	714
Biostatistics	720
A comparison of the CAR and DAGAR spatial random effects models with an application to diabetes rate estimation in Belgium. <i>Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti</i>	721
A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma. <i>Marta Spreafico, Francesca Ieva and Marta Fiocco</i>	727

A Statistical Approach to the Alignment of fMRI Data. <i>Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby and Livio Finos</i>	733
Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis. <i>Martin Forster and Marco Novelli</i>	739
Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients. <i>Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora</i>	745
Combining multiple sources to overcome misclassification bias in epidemiological database studies. <i>Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini</i>	751
Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics. <i>Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Nicola Rares Franco, Tiziana Rancati and Catharine West</i>	756
Graphical models for count data: an application to single-cell RNA sequencing. <i>Nguyen Thi Kim Hue, Monica Chiogna and Davide Rizzo</i>	762
Interregional mobility, socio-economic inequality and mortality among cancer patients. <i>Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci and Salvatore Scondotto</i>	768
PET radiomics-based lesions representation in Hodgkin lymphoma patients. <i>Lara Cavinato, Martina Sollini, Margarita Kirienko, Matteo Biroli, Francesca Ricci, Letizia Calderoni, Elena Tabacchi, Cristina Nanni, Pier Luigi Zinzani, Stefano Fanti, Anna Guidetti, Alessandra Alessi, Paolo Corradini, Ettore Seregni, Carmelo Carlo-Stella, Arturo Chiti and Francesca Ieva</i>	774
Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences. <i>Nicola Rares Franco, Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Tiziana Rancati and Catharine West</i>	780
Predictive versus posterior probabilities for phase II trial monitoring. <i>Valeria Sambucini</i>	785
Profile networks for precision medicine. <i>Andrea Lazerini, Monia Lupporelli and Francesco C. Stingo</i>	791
Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression. <i>Dario Delle Vedove, Francesca Ieva and Anna Maria Paganoni</i>	797
Selecting optimal thresholds in ROC analysis with clustered data. <i>Duc Khanh To, Gianfranco Adimari and Monica Chiogna</i>	803
Environment, Physics and Engineering	809
A hidden semi-Markov model for segmenting environmental toroidal data. <i>Francesco Lagona and Antonello Maruotti</i>	810
An experimental analysis on quality and security about green communication. <i>Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce and Massimiliano Giacalone</i>	816
An improved sensitivity-data based method for probabilistic ecological risk assessment. <i>Sonia Migliorati and Gianna Serafina Monti</i>	822
Comparing predictive distributions in EMOS. <i>Giummolè Federica and Mameli Valentina</i>	828
Compositional analysis of fish communities in a fast changing marine ecosystem. <i>Pierfrancesco Alaimo Di Loro, Marco Mingione, Giovanna Jona Lasinio, Sara Martino and Francesco Colloca</i>	834
FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy. <i>Francesca Di Salvo, Elena Piacenza and Delia Francesca Chillura Martino</i>	840
Functional Data Analysis for Spectroscopy Data. <i>Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis and Simone Vantini</i>	846
Functional graphical model for spectrometric data analysis. <i>Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini</i>	852
Local LGCP estimation for spatial seismic processes. <i>Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro and Giada Adelfio</i>	857

Observation-driven models for storm counts. <i>Mirko Armillotta, Alessandra Luati and Monia Lupparelli</i>	863
Statistical control of complex geometries, with application to Additive Manufacturing. <i>Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi</i>	869
Tree attributes map by 3P sampling in a design-based framework. <i>Lorenzo Fattorini and Sara Franceschi</i>	875
Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms. <i>Roberto Baragona and Laura Bocci</i>	880
Finance, business and official statistics	886
A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy. <i>Stefano Ciliberti, Simone Del Sarto, Giulia Pastorelli, Angelo Frascarelli and Gaetano Martino</i>	887
A fuzzy approach to the measurement of the employment rate. <i>Bruno Cheli, Alessandra Coli and Andrea Regoli</i>	893
A proposal to model credit risk contagion using network count-based models. <i>Arianna Agosto and Daniel Felix Ahelegbey</i>	898
A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience. <i>Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini</i>	904
Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework. <i>Vincenzo Candila and Lea Petrella</i>	910
Company requirements in Italian tourism sector: an analysis for profiles. <i>Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga</i>	916
Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach. <i>Lucio Masserini, Matilde Bini and Alessandro Zeli</i>	921
Double Asymmetric GARCH-MIDAS model - new insights and results. <i>Alessandra Amendola, Vincenzo Candila and Giampiero M. Gallo</i>	927
European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects. <i>Luca Secondi</i>	933
Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins. <i>Paolo Pagnottoni</i>	939
Forecasting Banknote Flows in Bdl Branches: Speed-up with Machine Learning. <i>Marco Brandi, Monica Fusaro, Tiziana Laureti and Giorgia Rocco</i>	945
Fully reconciled GDP forecasts from Income and Expenditure sides. <i>Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto</i>	951
GLASSO Estimation of Commodity Risks. <i>Beatrice Foroni, Saverio Mazza, Giacomo Morelli and Lea Petrella</i>	957
Measuring the Effect of Unconventional Policies on Stock Market Volatility. <i>Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto</i>	963
Multidimensional versus unidimensional poverty measurement. <i>Michele Costa</i>	969
Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach. <i>Alessandro Magrini</i>	975
Nowcasting GDP using mixed-frequency based composite confidence indicators. <i>Maria Carannante, Raffaele Mattered, Michelangelo Misuraca, Germana Scepi and Maria Spano</i>	981
On the tangible and intangible assets of Initial Coin Offerings. <i>Paola Cerchiello and Anca Mirela Toma</i>	987

Seasonality variation of electricity demand: decompositions and tests. <i>Luigi Grossi and Mauro Mussini</i>	993
SMEs circular economy practices in the European Union: Implications for sustainability. <i>Nunzio Tritto, José G. Dias and Francesca Bassi</i>	999
Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises. <i>Alessandra Righi</i>	1005
The determinants of eco-innovation: a country comparison using the community innovation survey. <i>Ida D'Attoma and Silvia Pacei</i>	1011
World ranking of urban sustainability through composite indicators. <i>Elena Grimaccia, Alessia Naccarato and Silvia Terzi</i>	1017
Machine Learning and Data Science.....	1023
A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values. <i>Paolo Giudici and Emanuela Raffinetti</i>	1024
A warning signal for variable importance interpretation in tree-based algorithms. <i>Anna Gottard and Giulia Vannucci</i>	1030
Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking. <i>Gianpaolo Zammarchi, Claudio Conversano and Francesco Mola</i>	1036
At risk mental status analysis: a comparison of model selection methods for ordinal target variable. <i>Elena Ballante, Silvia Molteni, Martina Mensi and Silvia Figini</i>	1042
Categorical Encoding for Machine Learning. <i>Agostino Di Ciaccio</i>	1048
Dynamic Quantile Regression Forest. <i>Mila Andreani and Lea Petrella</i>	1054
Estimating the UK Sentiment Using Twitter. <i>Stephan Schlosser, Daniele Toninelli and Michela Cameletti</i>	1059
Forecasting local rice prices from crowdsourced data in Nigeria. <i>Ilaria Lucrezia Amerise and Gloria Solano Hermosilla</i>	1065
Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout? <i>Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni</i>	1071
HateViz: a textual dashboard Twitter data-driven. <i>Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Nicola Canestrari</i>	1077
How to perform cyber risk assessment via cumulative logit models. <i>Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola</i>	1083
Machine learning prediction for accounting system. <i>Chiara Bardelli and Silvia Figini</i>	1087
Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. <i>Cristina Davino, Rosa Fabbriatore, Carla Galluccio, Daniela Pacella, Domenico Vistocco, Francesco Palumbo</i>	1093
The weight of words: textual data versus sentiment analysis in stock returns prediction. <i>Riccardo Ferretti and Andrea Sciandra</i>	1099
Unsupervised Energy Trees: clustering with complex and mixed-type variables. <i>Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti</i>	1105
Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale. <i>Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella</i>	1111
Variable selection for robust model-based learning from contaminated data. <i>Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy</i>	1117

Variable Selection in Text Regressions: Back to Lasso? <i>Marzia Freo and Alessandra Luati</i>	1123
Web Usage Mining and Website Effectiveness. <i>Maria Francesca Cracolici and Furio Urso</i>	1129
Models and methods - Categorical, Ordinal, Rank Data	1135
Aberration for the analysis of two-way contingency tables. <i>Roberto Fontana and Fabio Rapallo</i>	1136
An investigation of the paradoxical behaviour of κ -type inter-rater agreement coefficients for nominal data. <i>Amalia Vanacore and Maria Sole Pellegrino</i>	1142
Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach. <i>Luigi Lombardi and Antonio Calcagni</i>	1148
BOD – min range: A Robustness Analysis Method for Composite Indicators. <i>Emiliano Seri, Leonardo Salvatore Alaimo and Vittoria Carolina Malpassuti</i>	1154
Comparing classifiers for ordinal variables. <i>Silvia Golia and Maurizio Carpita</i>	1160
Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data. <i>Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio, Mark De Rooij and Elise Dusseldorp</i>	1166
Hybrid random forests for ordinal data. <i>Rosaria Simone and Gerhard Tutz</i>	1171
Model-based approach to biclustering ordinal data. <i>Monia Ranalli and Francesca Martella</i>	1177
New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution. <i>Cristina Mollica and Luca Tardella</i>	1183
Non-metric unfolding on augmented data matrix: a copula-based approach. <i>Marta Nai Ruscone and Antonio D'Ambrosio</i>	1189
Ordinal probability effect measures for dyadic analysis in cumulative models. <i>Maria Iannario and Domenico Vistocco</i>	1194
Simulated annealing for maximum rater agreement. <i>Fabio Rapallo and Maria Piera Rogantin</i>	1200
Models and methods – Regression.....	1206
A Clusterwise regression method for Distributional-valued Data. <i>Rosanna Verde, Francisco de A. T. de Carvalho and Antonio Balzanella</i>	1207
A nonparametric approach for nonlinear variable screening in high-dimensions. <i>Francesco Giordano, Sara Milito and Lucia Maria Parrella</i>	1213
Adjusted scores for inference in negative binomial regression. <i>Euloge C. Kenne Pagui, Alessandra Salvan and Nicola Sartori</i>	1219
Estimation of the treatment effect variance in a difference-in-differences framework. <i>Marco Doretti and Giorgio E. Montanari</i>	1224
Exploring multicollinearity in quantile regression. <i>Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco</i>	1230
Generalized M-quantile random effects model. <i>Francesco Schirripa Spagnolo and Vincenzo Mauro</i>	1236
Goodness-of-fit assessment in linear quantile regression. <i>Ilaria Lucrezia Amerise and Agostino Tarsitano</i>	1242
Joint Redundancy Analysis by a multivariate linear predictor. <i>Laura Marcis and Renato Salvatore</i>	1248

M-quantile regression shrinkage and selection via the lasso. <i>M. Giovanna Ranalli, Lea Petrella and Francesco Pantalone</i>	1254
New insights into the Conditioning and Gain Score approaches in multilevel analysis. <i>Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini</i>	1260
Simultaneous confidence regions and curvature measures in nonlinear models. <i>Claudia Furlan and Cinzia Mortarino</i>	1265
Models and methods – Sampling	1271
Design-based consistency of the Horvitz-Thompson estimator for spatial populations. <i>Lorenzo Fattorini, Marzia Marcheselli, Caterina Pisani and Luca Pratelli</i>	1272
Empirical likelihood in the statistical matching for informative samples. <i>Daniela Marella and Danny Pfeffermann</i>	1278
Evaluating a Hybrid One-Stage Snowball Sampling through Bootstrap Method on a Simulated Population. <i>Venera Tomaselli and Giulio Giacomo Cantone</i>	1284
How optimal subsampling depends on guessed parameter values. <i>Laura Deldossi and Chiara Tommasi</i>	1290
Indicators for risk of selection bias in non-probability samples. <i>Emilia Rocco and Alessandra Petrucci</i>	1296
On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. <i>Caterina May and Chiara Tommasi</i>	1302
Pseudo-population based resamplings for two-stage design. <i>Pier Luigi Conti, Daniela Marella and Vincenzina Vitale</i>	1308
Models and methods - Theoretical Issues in Statistical Inference	1314
A new mixture model for three-way data. <i>Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato</i>	1315
A Sequential Test for the Cpmk Index. <i>Michele Scagliarini</i>	1320
Probability Interpretations and the Selection of the Most Effective Statistics Method. <i>Paolo Rocchi</i>	1326
Robust Composite Inference. <i>Valentina Mamei, Monica Musio, Erlis Ruli and Laura Ventura</i>	1332
Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques. <i>Massimiliano Giacalone and Demetrio Panarello</i>	1338
Stochastic dependence with discrete copulas. <i>Fabrizio Durante and Elisa Perrone</i>	1344
Models and methods - Time Series and Longitudinal Data.....	1350
Bootstrap test in Poisson-INAR models. <i>Lucio Palazzo and Riccardo Ievoli</i>	1351
Continuous Time-Interaction Processes for Population Size Estimation. <i>Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini</i>	1357
Longitudinal data analysis using PLS-PM approach. <i>Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia and Marina Marino</i>	1363
Long-memory models for count time series. <i>Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto</i>	1369

Combining multiple frequencies in Realized GARCH models. <i>Antonio Naimoli and Giuseppe Storti</i>	1375
Models with Time-Varying Parameters for Realized Covariance. <i>Luc Bauwens and Edoardo Otranto</i>	1381
Pitman-Yor mixture models for survival data stratification. <i>Riccardo Corradin, Luis Enrique Nieto Barajas and Bernardo Nipoti</i>	1387
Prediction is not everything, but everything is prediction. <i>Leonardo Egidi</i>	1393
The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data. <i>Francesca Martella, Antonello Maruotti and Francesco Tursini</i>	1399
Trends and long-run relations in cointegrated time series observed with noise. <i>Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti</i>	1405
Population and society	1411
A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach. <i>Simone Del Sarto, Michela Gnalzi, Yara Maasri and Edouard Legoupil</i>	1412
Accounting for Interdependent Risks in Vulnerability Assessment of Refugees. <i>Daria Mendola, Anna Maria Parroco and Paolo Li Donni</i>	1418
Active ageing in China: What are the domains that most affect life satisfaction in the elderly? <i>Ilaria Rocco</i>	1424
Analyzing the waiting time of academic publications: a survival model. <i>Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi and Francesco Porro</i>	1430
Clustering of food choices in a large sample of students using university canteen. <i>Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli and Roberto Barale, Giuseppe Turchetti</i>	1436
Cruise passengers' expenditure at destinations: Review of survey techniques and data collection. <i>Caterina Sciortino, Stefano De Cantis, Mauro Ferrante and Szilvia Gyimóthy</i>	1442
Educational integration of foreign citizen children in Italy: a synthetic indicator. <i>Alessio Buonomo, Stefania Capecchi and Rosaria Simone</i>	1448
Estimating the Change in Housework Time of the Italian Woman after the Retirement of the Male Partner: An Approach Based on a Two-Regime Model Estimated by ML. <i>Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini</i>	1454
First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. <i>Antonella D'Agostino, Giulio Ghellini and Gabriele Lombardi</i>	1460
Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study. <i>Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari</i>	1466
Gender and Monetary Policy Preferences: a Diff-in-Diff Approach. <i>Donata Favaro, Anna Giraldo and Ina Gollikja</i>	1472
Headcount based indicators and functions to evaluate the effectiveness of Italian university education. <i>Silvia Terzi and Francesca Petrarca</i>	1478
Identify the speech code through statistics: a data-driven approach. <i>Andrea Briglia, Massimo Mucciardi and Jérémi Sauvage</i>	1484
Inspecting cause-specific mortality curves by simplicial functional data analysis. <i>Marco Stefanucci and Stefano Mazzucco</i>	1490
Intertemporal decision making and childless couples. <i>Daniela Bellani, Bruno Arpino and Daniele Vignoli</i>	1495
Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables. <i>Laura Bocci and Isabella Mingo</i>	1501

LI-CoD Model. From Lifespan Inequality to Causes of Death. <i>Andrea Nigri and Susanna Levantesi</i>	1507
Modeling Well-Being through PLS-SEM and K-M. <i>Venera Tomaselli, Mario Fordellone and Maurizio Vichi</i>	1513
News life-cycle: a multiblock approach to the study of information. <i>Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza</i>	1519
Short-term rentals in a tourist town. <i>Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti and Veronica Sula</i>	1525
Sportstat: a playful activity to developing statistical literacy. <i>Alessandro Valentini and Francesca Paradisi</i>	1531
Statistical modeling for some features of Airbnb activity. <i>Giulia Contu and Luca Frigau</i>	1537
Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University. <i>Cristina Giudici, Donatella Vicar and Eleonora Trappolini</i>	1543
The Causal Effect of Immigration Policies on Income Inequality. <i>Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù</i>	1549
The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea and Mandatory Notices of the Ministry of Labour. <i>Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci</i>	1557
The joint effect of childcare services and flexible female employment on fertility rate in Europe. <i>Viviana Cocuccio and Massimo Mucciardi</i>	1565
The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs? <i>Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano and Antonella Rocca</i>	1571
The probability to be employed of young adults of foreign origin. <i>Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza</i>	1577
The risk of inappropriateness in geriatric wards: a comparison among the Italian regions. <i>Paolo Mariani, Andrea Marietta, Marcella Mazzoleni and Mariangela Zenga</i>	1583
The role of the accumulation of poverty and unemployment for health disadvantages. <i>Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle</i>	1589
Unemployment and fertility in Italy. A regional level data panel analysis. <i>Gabriele Ruiu and Marco Breschi</i>	1595
University drop out and mobility in Italy. First evidences on first level degrees. <i>Nicola Tedesco and Luisa Salaris</i>	1601
Worthiness-based Scale Quantifying. <i>Giulio D'Epifanio</i>	1607
Young people in Southern Italy and the phenomenon of immigration: what is their perception? <i>Nunziata Ribecco, Angela Maria D'Ugento and Angela Labarile</i>	1613

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution

Nuovi algoritmi e diagnostiche di bontà di adattamento per la modellizzazione dei dati di ranking con la distribuzione di Plackett-Luce estesa

Cristina Mollica and Luca Tardella

Abstract The *forward order* assumption postulates that the ranking process of the items is carried out by assigning the positions from the top (most-liked) to the bottom (least-liked) alternative. This assumption has been recently relaxed in the *Extended Plackett-Luce model* (EPL) through the introduction of the discrete *reference order* parameter, describing the rank attribution path. By starting from two formal properties of the EPL, we derive novel diagnostic tools for testing appropriateness of the EPL assumption. We also show how one of the two statistics can be exploited to construct a heuristic method, that surrogates the maximum likelihood approach for inferring the underlying reference order. The performance of the proposals was compared with more conventional approaches through an extensive simulation study.

Abstract *L'ipotesi di ordinamento in avanti postula che il processo di ranking venga eseguito assegnando le posizioni dalla prima all'ultima. Questa ipotesi è stata recentemente rilassata nel modello di Plackett-Luce esteso (EPL) attraverso l'introduzione del parametro discreto detto ordine di riferimento, che descrive il percorso di attribuzione del rango. Partendo da due proprietà formali dell'EPL, deriviamo nuovi strumenti diagnostici per testare l'adeguatezza del modello EPL. Mostriamo anche come una delle due statistiche possa essere impiegata per costruire un metodo euristico, che surroga l'approccio di massima verosimiglianza per l'inferenza sull'ordine di riferimento. La performance delle proposte è stata confrontata con approcci più convenzionali attraverso un ampio studio di simulazione.*

Key words: ranking data, Plackett-Luce model, model specification, goodness-of-fit assessment, bootstrap, heuristic methods

Cristina Mollica
Dipartimento MEMOTEF, Sapienza Università di Roma
e-mail: cristina.mollica@uniroma1.it

Luca Tardella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: luca.tardella@uniroma1.it

1 Introduction

Let us consider an experiment in which a sample of N judges is asked to rank a set $I = \{1, \dots, K\}$ of K labeled alternatives, namely *items*, according to a certain criterion. A *ranking* is a vector $\pi = (\pi(1), \dots, \pi(K))$ collecting the ranks assigned to each item, specifically the entry $\pi(i)$ indicates the position attributed to the i -th alternative. Equivalently, data can be recorded in the *ordering* format $\pi^{-1} = (\pi^{-1}(1), \dots, \pi^{-1}(K))$, where the generic component $\pi^{-1}(j)$ indicates the item ranked in the j -th position. Therefore, ranking data are multivariate ordinal data taking values in the set of permutations \mathcal{S}_K of the first K integers.

This work concentrates on the parametric family of stagewise ranking models, in particular on the *Extended Plackett-Luce model* (EPL) introduced by [1]. The EPL generalizes the *Plackett-Luce model* (PL) [2, 3] by relaxing the implicit forward order assumption with the introduction of the *reference order parameter* $\rho = (\rho(1), \dots, \rho(K))$. It indicates the rank assignment order, i.e., $\rho(t)$ is the position attributed at the stage t . The probability of an ordering π^{-1} under the EPL is

$$\mathbf{P}_{\text{EPL}}(\pi^{-1} | \rho, \underline{p}) = \mathbf{P}_{\text{PL}}(\pi^{-1} \rho | \underline{p}) = \prod_{t=1}^K \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^K p_{\pi^{-1}(\rho(v))}} \quad \pi^{-1} \in \mathcal{S}_K, \quad (1)$$

where ρ is the discrete parameter, specifically a permutation of the first K integers, and the positive quantities p_i 's are referred to as *support parameters*, proportional to the probabilities for each item to be ranked in the position indicated by the first entry of ρ . Hereinafter, we will shortly refer to (1) as $\text{EPL}(\rho, \underline{p})$. Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the Maximum Likelihood Estimation (MLE) perspective in [1], via the hybrid Expectation-Maximization-Minimization (EMM) algorithm. Recently, [4] introduced the Bayesian version of the EPL, where a tuned joint Metropolis-within-Gibbs sampling was developed to conduct approximate posterior inference. The MCMC algorithm was also adapted to infer on the EPL with order constraints on ρ [6] and on the Bayesian EPL mixture [7].

2 Novel EPL diagnostics and comparative evaluation

Specific diagnostic tools to evaluate model adequacy of multistage ranking models are very limited in the ranking literature and their effectiveness has not been deeply explored. The present work aims at providing some contributions in this direction.

Let us suppose that we have some data simulated from an $\text{EPL}(\rho, \underline{p})$. We expect the marginal frequencies of the items at the first stage to be ranked according to the order of the corresponding support parameter component. On the other hand, we expect the marginal frequencies of the items at the last stage to be ranked according to the reverse order of the corresponding support parameter component. One can then derive that the ranking of the marginal frequencies of the items corresponding

to the first and last stage should sum up to $(K + 1)$, no matter what their support is. Of course, this is less likely to happen when the sample size is small or when the support parameters are not so different of each other. In any case, one can define a test statistic by considering, for each couple of integers (j, j') candidate to represent the first and the last stage ranks, namely $\rho(1)$ and $\rho(K)$, a discrepancy measure $T_{jj'}(\pi)$ between $K + 1$ and the sum of the observed ranks of the frequencies corresponding to the same item extracted in the first and in the last stage. Formally, let $\underline{r}_j^{[1]} = (r_{j1}^{[1]}, \dots, r_{jK}^{[1]})$ and $\underline{r}_{j'}^{[K]} = (r_{j'1}^{[K]}, \dots, r_{j'K}^{[K]})$ be the marginal item frequency distributions for the j -th and j' -th positions, to be assigned respectively at the first [1] and last [K] stage. In other words, the generic entry $r_{ji}^{[s]}$ is the number of times that item i is ranked j -th at the s -th stage. The proposed EPL diagnostic relies on the following discrepancy

$$T_{jj'}(\pi) = \sum_{i=1}^K |\text{rank}(\underline{r}_j^{[1]})_i + \text{rank}(\underline{r}_{j'}^{[K]})_i - (K + 1)|, \quad (2)$$

implying that the smaller the value $T_{jj'}(\pi)$, the larger the plausibility that the two integers (j, j') represent the first and the last components of the reference order. In this sense, $T_{jj'}(\pi)$ is a measure of the closeness of the positions j and j' in ρ . To globally assess the conformity of the sample with the EPL, we consider the statistic

$$T_m(\pi) = \min_{j < j'} T_{jj'}(\pi). \quad (3)$$

With the aim at further enlarging the collection of diagnostics for the EPL class, we focus our attention also on another specific property of the EPL, known as *Luce's choice axiom* or *independence of irrelevant alternatives* (IIA). The IIA hypothesis implies that the probability ratio of selecting item i over item i' is constant over the stages of the ranking process (constant ratio rule), as long as the two items are both still available. Formally, it implies that the expected paired comparison frequency at stage t of choosing item i over item i' is $\tau_{ii't}^* = T_{ii't} p_i / (p_i + p_{i'})$, given by the product between the total number $T_{ii't}$ of pairwise comparisons between i and i' at stage t and the theoretical pairwise comparison probability under the EPL. Hence, a chi-squared statistic for the IIA assumption can be defined as follows

$$X_{\text{IIA}}^2 = \sum_{t=1}^{K-1} \sum_{i < i'} \frac{(\tau_{ii't} - \tau_{ii't}^*)^2}{\tau_{ii't}^*},$$

where $\tau_{ii't}$ is the observed paired comparison frequency at stage t .

After introducing novel test statistics, one should enquire into their power, for instance through a bootstrap approach, and preferably compare it with that of some standard goodness-of-fit tools for ranking models. To this aim, we conducted a simulation study under alternative model specifications, involving in the comparison the chi-squared discrepancies based on the top frequencies, the pairwise comparisons and the first-order marginals, given respectively by

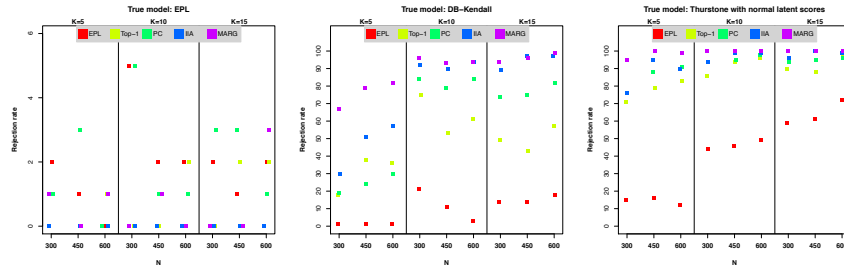


Fig. 1: Rejection rates of the EPL assumption for alternative goodness-of-fit diagnostics computed on simulated data from different model scenarios.

$$X_{\text{TOP}}^2 = \sum_{i=1}^K \frac{(m_{1i} - m_{1i}^*)^2}{m_{1i}^*} \quad X_{\text{PC}}^2 = \sum_{i < i'} \frac{(\tau_{ii'} - \tau_{ii'}^*)^2}{\tau_{ii'}^*} \quad X_{\text{M}}^2 = \sum_{j=1}^K \sum_{i=1}^K \frac{(m_{ji} - m_{ji}^*)^2}{m_{ji}^*}$$

where the expected frequencies are $m_{1i}^* = N p_i$ and $\tau_{ii'}^* = N \frac{p_i p_{i'}}{p_i + p_{i'}}$, whereas m_{ji}^* were estimated with a Monte Carlo simulation.

A comparative evaluation of the competing model specification tools was carried out by means of an extensive simulation study. For each possible combination (K, N) , with values varying respectively in the grids $K \in \{5, 10, 15\}$ and $N \in \{300, 450, 600\}$, we drew 100 datasets with N orderings of K items from the following ranking distributions: i) EPL, ii) distance-based model (DB) with the Kendall metric (DB-Kend), iii) DB with the Hamming distance (DB-Ham) and iv) TH with normal latent scores (TH-norm). where the true parameter values were uniformly generated. To approximate the reference distribution of the test statistics under the EPL assumption, the bootstrap method was applied. Finally, for each model adequacy criterion, we estimated the mis- and correct rejection rates with the relative frequency of the times that the p -value was smaller than or equal to 0.05.

The simulation study revealed a satisfactory performance of all the considered diagnostics regarding the rates of mis-rejections, testified by estimated Type I error probabilities below 0.05 (Figure 1, left). On the other hand, noteworthy differences emerged in terms of the power. Firstly, measure (3) exhibited a consistent poor behavior of the estimated power under each considered model scenario, see for instance the plots in the middle and on the right of Figure 1. At least two motivations can be put forward to argue this evidence. The former is related to the formal definition of $T_m(\pi)$; in fact, (3) is a parameter-free measure based on the ranks of the expected marginal frequencies, rather than on the computation of the parameter-dependent first- and last-stage theoretical probabilities. This makes $T_m(\pi)$ by construction a rougher diagnostic in the comparison with the other statistics. Secondly, the remarkably low power of (3) under the DB with the Kendall metric (Figure 1, center) suggested that the monotonicity property of the first- and last-stage item probabilities is not specific of the EPL, but it is shared by other rankings models too. Another stable evidence highlighted by the comparative analysis concerns the

best-performing diagnostic, which turned to be the one relying on the marginal item distributions. However, it is no less apparent that, for higher values of K and N , the performance of the new IIA statistic is pretty much equivalent to that of the chi-squared based on the marginal distributions and, in general, always better than the remaining competing statistics typically used in the real-data applications.

3 Likelihood-free estimation of the reference order

Rather than from a model specification perspective, in this section we explored the utility of the statistic (3) from the inferential point of view.

Let $\mathbf{T}(\boldsymbol{\pi}) = (T_{jj'}(\boldsymbol{\pi}))$ be the $K \times K$ matrix with entries defined in (2). For each component $T_{jj'}(\boldsymbol{\pi})$, the inequality $T_{jj'}(\boldsymbol{\pi}) \leq u_K$ holds, where the upper bound u_K corresponds to the constant value in the main diagonal of the matrix, i.e., $u_K = T_{jj}(\boldsymbol{\pi}) = \sum_{l=1}^K |2l - (K+1)|$ depending on data only through K . Our heuristic method to estimate the unknown parameter ρ is composed of the following steps:

1. compute $\mathbf{D}(\boldsymbol{\pi}) = |\mathbf{T}(\boldsymbol{\pi}) - u_K \mathbf{J}_K|$, where \mathbf{J}_K is $K \times K$ all-ones matrix, so that each component $D_{jj'}(\boldsymbol{\pi})$ can be interpreted as a measure of the distance between positions j and j' in the sequential rank assignment process;
2. use the matrix $\mathbf{D}(\boldsymbol{\pi})$ as the input of a Principal Component Analysis (PCA);
3. estimate ρ by taking the non-decreasing ordering of the scores (s_1, \dots, s_K) of the K positions on the first PC.

The inferential effectiveness of the proposal to recover the true discrete parameter was explored by means of a simulation study. For each possible combination (K, N) , where $K \in \{5, 10, 15\}$ and $N \in \{50, 200, 1000, 10000\}$, we drew 100 datasets $\boldsymbol{\pi}_{(R)}^{-1}$ with $R = 1, \dots, 100$ from the EPL with uniformly generated parameters. For comparison purposes, we inferred the reference order of each simulated sample $\boldsymbol{\pi}_{(R)}^{-1}$ with the heuristic strategy described above, with the one replacing the PCA with the Multidimensional Scaling (MDS) and with the MLE approach [1], which is considered as the reference method for the present estimation task. Finally, the estimation performance of the competing strategies were compared in terms of: i) percentage of matching between $\hat{\rho}^{(R)}$ and $\hat{\rho}^{(R)}$ (% recoveries) and ii) average cograduation $\bar{r}_{\text{Spear}}(\hat{\rho}, \hat{\rho})$ between the estimated and the actual reference order computed with Spearman's rank correlation.

As apparent in Table 1, PCA and MDS exhibited essentially the same ability. Compared with the MLE, one can appreciate very good results for the heuristic methods. The percentage of matching consistently grows with N and, by checking also the cases where there is not an exact correspondence, on average an analogous trend is highlighted for the relative Spearman correlation. Additionally, if we look at a fixed N , the percentage of recoveries shows a worse tendency for larger values of K . In this regard, the cases $K \in \{10, 15\}$ combined with a relatively very low ($N = 50$) and very high ($N = 10000$) sample size deserve some considerations to stress typical issues which can be encountered in a ranking data analysis. First, in a

Table 1: Inferential performance of the heuristic methods via PCA and MDS to estimate the reference order on simulated data compared to the MLE.

(K, N)	% recoveries			$\bar{r}_{\text{Spear}}(\hat{\rho}, \rho)$		
	PCA	MDS	MLE	PCA	MDS	MLE
(5, 50)	58	60	45	0.86	0.86	0.69
(5, 200)	79	80	77	0.89	0.89	0.95
(5, 1000)	91	90	100	0.92	0.93	1.00
(5, 10000)	97	97		0.94	0.94	
(10, 50)	3	5	3	0.90	0.89	0.87
(10, 200)	14	16	23	0.93	0.93	0.98
(10, 1000)	55	54	68	0.96	0.95	0.99
(10, 10000)	79	78		0.96	0.96	
(15, 50)	0	0	0	0.92	0.92	0.91
(15, 200)	0	1	3	0.94	0.94	0.97
(15, 1000)	13	15	26	0.97	0.97	0.99
(15, 10000)	48	51		0.97	0.97	

sparse data situation, all of the estimation techniques exhibit a great uncertainty in recovering the actual ρ , testified by the negligible values of the recovery percentage. On the other hand, although a better behavior of the MLE is expected for $N = 10000$, this has not been implemented since, without a specialized program, fitting the EPL to a large sample can be deeply computational demanding if not actually unfeasible. Moreover, the computational burden is further aggravated by the multiple initialization needed to address the issue of local maxima. In the light of these remarks, the likelihood-free approach can be motivated as a straightforward method that can be combined with the MLE or with an MCMC method in the Bayesian framework. In fact, without computational costs, it can be implemented as a preliminary step to obtain a promising initialization, that can guide the parameter space exploration towards the global optimum and substantially reduce the elaboration time.

References

1. Mollica, C., Tardella, L.: Epitope profiling via mixture modeling of ranked data. *Stat Med.* **33**(21), 3738–3758 (2014)
2. Luce, R.D.: Individual choice behavior: A theoretical analysis. John Wiley & Sons Inc (1959)
3. Plackett, R.L.: The analysis of permutations. *J Royal Stat Soc C.* **24**(2), 193–202 (1975)
4. Mollica, C., Tardella L.: Bayesian analysis of ranking data with the Extended Plackett-Luce model. *Statistical Methods and Applications*. doi: 10.1007/s10260-020-00519-5 (2020)
5. Mollica, C., Tardella L.: Bayesian mixture of Plackett-Luce models for partially ranked data. *Psychometrika.* **82**(2), 442–458 (2017)
6. Mollica, C., Tardella, L.: Constrained Extended Plackett-Luce model for the analysis of preference rankings. In: *Book of Short Papers SIS 2018*, (eds.) Springer Italia, pp 480-486, ISBN: 9788891910233 (2018)
7. Mollica, C., Tardella, L. Modelling unobserved heterogeneity of ranking data with the Bayesian mixture of Extended Plackett-Luce models. In: *Book of Short Papers CLADAG 2019*, pp. 346-349, ISBN: 978-88-8317-108-6 (2019)