Università di Pisa

Sant'Anna
Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Book of Short Papers
SIS 2020

SIS 2020
Pisa

SIS
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
Tiziana Tuoto and Marco Fortini

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
Ilaria Benedetti and Federico Crescenzi

Communication of Uncertainty of Official Statistics. 16
Edwin de Jonge and Gian Luigi Mazzi

Measuring uncertainty for infra-annual macroeconomic statistics. 22
George Kapetanios, Massimiliano Marcellino and Gian Luigi Mazzi

Bayesian methods in biostatistics 27

Network Estimation of Compositional Data. 28
Nathan Osborne, Christine B. Peterson and Marina Vannucci

Using co-data to empower genomics-based prediction and variable selection. 34
Magnus M. Münch, Mirrelijn M. van Nee and Mark A. van de Wiel

Data integration versus privacy protection: a methodological challenge? 40

Statistical Disclosure Control for Integrated Data. 41
Natalie Shlomo

The Integrated System of Statistic Registers: first steps towards facing privacy issues. 47
Mauro Bruno and Roberta Radini

Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics. 53
Fabio Ricciato, Kostas Giannakouris, Albrecht Wirthmann and Martina Hahn

Designing adaptive clinical trials 59

Optimal designs for multi-arm exponential trials. 60
Rosamarie Frieri and Marco Novelli

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
Isabella Sulis, Silvia Columbu and Mariano Porcu

Pull factors for university students' mobility: a gravity model approach. 73
Giovanni Boscaino and Vincenzo Giuseppe Genova

Spatial autoregressive gravity models to explain the university student mobility in Italy. 79
Silvia Bacci, Bruno Bertaccini and Chiara Bocci

Environmental Statistics (GRASPA-SIS) 85

A Time Clustering Model for Spatio-Temporal Data. 86
Clara Grazian, Gianluca Mastrantonio and Enrico Bibbona

Reconstruction of sparsely sampled functional time series using frequency domain functional principal components. 93
Amira Elayouty, Marian Scott and Claire Miller

Methods for High Dimensional Compositional Data Analysis 98

Algorithms for compositional tensors of third-order. 99
Violetta Simonacci

High-dimensional regression with compositional covariates: a robust perspective. 105
Gianna Serafina Monti and Peter Filzmoser

Three-way compositional analysis of energy intensity in manufacturing. 111
Valentin Todorov and Violetta Simonacci

Modern Statistics for Physics Discoveries 117

Identification of high-energy λ-ray sources via nonparametric clustering. 118
Giovanna Menardi, Denise Costantin, and Federico Ferraccioli

Statistical Analysis of Macroseismic Data for a better Evaluation of Earthquakes Attenuation Laws. 124
Marcello Chiodi, Antonino D'Alessandro, Giada Adelfio and Nicoletta D'Angelo

Network Modelling in Biostatistics...130

Natural direct and indirect relative risk for mediation analysis. 131
Monia Lupparelli and Alessandra Mattei

New issues on multivariate and univariate quantile regression.................137

Mixtures of quantile regressions for longitudinal data: an R package. 138
Maria Francesca Marino, Maria Giovanna Ranalli and Marco Alfò

Multivariate Mixed Hidden Markov Model for joint estimation of multiple quantiles. 144
Luca Merlo, Lea Petrella and Nikos Tzavidis

Recent methodological advances in finite mixture modeling
with applications (CLADAG-SIS) 150

Aggregating Gaussian mixture components. 151
Roberto Rocci

Local and overall coefficients of determination for mixtures of generalized linear models. 157
Roberto Di Mari, Salvatore Ingrassia and Antonio Punzo

Statistical Analysis of Satellite Data (SDS-SIS) 163

Functional Data Analysis for Interferometric Syntethic Aperture Radar Data Post-Processing:
The case of Santa Barbara mud volcano. 164
Matteo Fontana, Alessandra Menafoglio, Francesca Cigna and Deodato Tapete

Recent Contributions to the Understanding of the Uncertainty in Upper-Air Reference Measurements. 170
Alessandro Fassò

Statistical models and methods for Business and Industry 176

Modelling and monitoring of complex 3D shapes: a novel approach for lattice structures. 177
Bianca Maria Colosimo, Marco Grasso and Federica Garghetti

Open data powered territorial planning - Case study: The Turin historical center. 183
Silvia Casagrande, Gianmaria Origgi, Alberto Pasanisi, Martina Tamburini, Pascal Terrien, Tania Cerquitelli and Alfonso Capozzoli

Process optimization in Industry 4.0: Are all data analytics models useful? 189
Alberto Ferrer

Technology and demographic behaviours (AISP-SIS) 195

Internet and the Timing of Births. 196
Maria Sironi, Osea Giuntella and Francesco C. Billari

The Internetization of Marriage: Effects of the Diffusion of High-Speed Internet on Marriage, Divorce, and Assortative Mating. 202
Francesco C. Billari, Osea Giuntella and Luca Stella

Solicited Sessions

Advanced Statistical Methods in Health Analytics 209

Assessing the impact of the intermediate event in a non-markovian illness-death model. 210
Davide Paolo Bernasconi, Elena Tassistro, Maria Grazia Valsecchi and Laura Antolini

Big data and AI: challenges and opportunities in healthcare. 216
Vieri Emiliani, Gian Luca Cattani and Fabrizio Selmi

Statistical methodology for volume-outcome studies. 222
Marta Fiocco and Floor van Oudenhoven

Advances in textual data mining 228

Distance measures for exploring pairs of novels in a large corpus of Italian literature. 229
Matilde Trevisani and Arjuna Tuzzi

Supervised vs Unsupervised Latent Dirichlet Allocation: topic detection in lyrics. 235
Mariangela Sciandra, Alessandro Albano and Irene Carola Spera

Advances in the interaction between artificial intelligence and official statistics 241

Automated Land Cover Maps from Satellite Imagery by Deep Learning. 242
Fabrizio De Fausti, Francesco Pugliese and Diego Zardetto

CROWD4SDG: Crowdsourcing for sustainable developments goals. 248
Barbara Pernici

Permanent Population Census: evaluation of the effects of regional strategies on the process efficiency.
The direct experience of Tuscany. 253
Linda Porciani, Luisa Francovich, Luca Faustini and Alessandro Valentini

Capture-recapture methods 259

Bayesian Model Averaging for Latent Class Models in Capture-Recapture. 260
Davide Di Cecco

Combining “signs of life” and survey data through latent class models to consider over-coverage in Capture-Recapture estimates of population counts. 266
Marco Fortini, Antonella Bernardini, Marco Caputi and Nicoletta Cibella

Population size estimation with interval censored counts and external information. 272
Alessio Farcomeni

Changes in environment extremes and their impacts 278

FPCA Clustering of rainfall events. 279
Gianluca Sottile, Antonio Francipane, Leonardo Noto and Giada Adelfio

Trends in rainfall extremes in the Venice lagoon catchment. 285
Ilaria Prosdocimi and Carlo Gaetan

Copulas: models and inference 291

Analysis of district heating demand through different copula-based approaches. 292
F. Marta L. Di Lascio and Andrea Menapace

CoVaR and backtesting: a comparison between a copula approach and parametric models. 298
Michele Leonardo Bianchi, Giovanni De Luca and Giorgia Rivieccio

Estimating Asymmetric Dependence via Empirical Checkerboard Copulas. 304
Wolfgang Trutschnig and Florian Griessenberger

Strong Convergence of Multivariate Maxima. 310
Michael Falk, Simone A. Padoan and Stefano Rizzelli

Data Science: when different expertise meet 316

Bayesian stochastic modelling of the temporal evolution of seismicity. 317
Elisa Varini and Renata Rotondi

Cluster Analysis for the Characterization of Residential Personal Exposure to ELF Magnetic Field. 323
Gabriella Tognola, Silvia Gallucci, Marta Bonato, Emma Chiaramello, Isabelle Magne, Martine Souques, Serena Fiocchi, Marta Parazzini and Paolo Ravazzani

Statistical Assessment and Validation of Ship Response in High Sea State by Computational Fluid Dynamics. 328
Andrea Serani, Matteo Diez and Frederick Stern

Uncertainty Quantification for PDEs with random data using the Multi-Index Stochastic Collocation method. 334
Lorenzo Tamellini and Joakim Beck

Emerging challenges in official statistics:
new data sources and methods 340

Small area poverty indicators adjusted using local spatial price indices. 341
Stefano Marchetti, Luigi Biggeri, Caterina Giusti and Monica Pratesi

Smart solutions for trusted smart statistics: the European big data hackathon experience. 347
Francesco Amato, Mauro Bruno, Tania Cappadozzi, Fabrizio De Fausti and Manuela Michelini

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. 360
Maurizio Carpita, Enrico Ciavolino and Paola Pasca

A framework of cumulated chi-squared type statistics for ordered correspondence analysis. New tools and properties. 366
Antonello D'Ambra, Pietro Amenta and Luigi D'Ambra

Exploring drug consumption via an ultrametric correlation matrix. 372
Giorgia Zaccaria and Maurizio Vichi

Ranking extraction in ordinal multi-indicator systems. 378
Marco Fattore and Alberto Arcagni

Gender statistics 384

Gender differences in Italian STEM degree courses: a discrete-time competing-risks model. 385
Marco Enea and Massimo Attanasio

Some Challenges and Results in Measuring Gender Inequality. 391
Fabio Crescenzi and Francesco Di Pede

How Deep is Your Plot? Young SIS and deep statistical learning (ySIS)..397

A modal approach for clustering matrices. 398
Federico Ferraccioli and Giovanna Menardi

A Note on Detection of Perturbations in Biological Networks. 404
Vera Djordjilović

Bayesian inference for DAG-probit models. 410
Federico Castelletti

Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process. 416
Sirio Legramanti

Measuring poverty and vulnerability 421

Choosing the vulnerability threshold using the ROC curve. 422
Chiara Gigliarano and Conchita D'Ambrosio

New advances in applications, a Bayesian nonparametric perspective 428

Bayesian Mixture Models for Latent Class Analysis. 429
Raffaele Argiento, Bruno Bodin and Maria De Iorio

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
Fabio Centofanti and Matteo Fontana

Statistical process monitoring of multivariate profiles from ship operating conditions. 440
Christian Capezza

Prior choice in Bayesian Modelling (SISbayes)..446

Bayesian Learning of Multiple Essential Graphs. 447
Luca La Rocca, Federico Castelletti, Stefano Peluso, Francesco Claudio Stingo and Guido Consonni

Bayesian post-processing of Gibbs sampling output for variable selection. 453
Stefano Cabras

Priors on precision parameters of IGRMF models. 459
Aldo Gardini, Fedele Greco and Carlo Trivisano

Sequence Analysis: methods and applications 465

Internal migration, family formation and social stratification in Europe. A life course approach. 466
Roberto Impicciatore, Gabriele Ballarino and Nazareno Panichella

Socio economic integration of migrants 472

A study on the characteristics of spouses who intermarry in Italy. 473
Agnese Vitali and Romina Fraboni

Statistical Analysis for mobility and transportation 479

A multilevel Analysis of University attractiveness in the network flows from Bachelor to Master's degree. 480
Silvia Columbu and Ilaria Primerano

Analysis of mobility data through a novel Cheng and Church algorithm for functional data. 486
Marta Galvani, Agostino Torti and Alessandra Menafoglio

Bridge closures in a transportation network: analysis of the impacts in the region of Lombardy. 491
Agostino Torti, Marika Arena, Giovanni Azzone, and Piercesare Secchi

Statistical Methods and Applications in Social Network Analysis 496

A clustering procedure for ego-networks data: an application to Italian elders living in couple. 497
Elvira Pelle and Roberta Pappadà

Analysing the mediating role of a network: a Bayesian latent space approach. 503
Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Network-time autoregressive models for valued network panel. 509
Viviana Amati

University student mobility flows and network data structures. 515
Maria Prosperina Vitale, Giuseppe Giordano and Giancarlo Ragozini

Statistical Methods in Psychometrics 521

A simple probabilistic model to evaluate questionable interim analysis strategies. 522
Francesca Freuli and Luigi Lombardi

Incorporating Expert Knowledge in Structural Equation Models: Applications in Psychological Research. 528
Gianmarco Altoè, Claudio Zandonella Callegher, Enrico Toffalini and Massimiliano Pastore

Predicting social media addiction from Instagram profiles: A data mining approach. 534
Antonio Calcagnì, Veronica Cortellazzo, Francesca Guizzo, Paolo Girardi, Natale Canale

Structural entropy based modeling for psychological measurement. 540
Enrico Ciavolino, Mario Angelelli, Paola Pasca and Omar Carlo Gioacchino Gelo

Statistical modelling in environmental epidemiology..................................546

A Time Varying Coefficient Model to Estimate the Short-Term Effects of Air Pollution on Human Health. 547
Pasquale Valentini, Luigi Ippoliti and Clara Grazian

Joint Analysis of Short and Long-Term Effects of Air Pollution. 551
Annibale Biggeri, Dolores Catelan, Giorgia Stoppa and Corrado Lagazio

Statistical Modelling of Scientific Evidence for Forensic Investigation and Interpretation..557

DNA mixtures with related contributors. 558
Peter J. Green and Julia Mortera

Forensic Statistics: How to estimate life expectancy after injury. 564
Jane L Hutton

The additional contribution of combining genetic evidence from multiple samples in a complex case. 570
Giampietro Lago

The history of forensic inference and statistics: a thematic perspective. 576
Franco Taroni and Colin Aitken

Topological learning: interpretable representations of complex data......581

Comparing Neural Networks via Generalized Persistence. 582
Mattia G. Bergomi and Pietro Vertechi

On the topological complexity of decision boundaries. 588
António Leitão and Giovanni Petri

Persistence-based Kernels for Data Classification. 594
Ulderico Fugacci

Topological and Mixed-type learning of Brain Activity. 600
Tullia Padellini, Pierpaolo Brutti, Riccardo Giubilei

Contributed papers and Posters

Bayesian Statistics 607

A Bayesian approach for modelling dependence among mixture densities. 608
Mario Beraha, Matteo Pegoraro, Riccardo Peli and Alessandra Guglielmi

A change of glasses strategy to solve the rare type match problem. 614
Giulia Cereda and Fabio Corradi

A new prior distribution on the simplex: the extended flexible Dirichlet. 620
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

ABC model choice via mixture weight estimation. 626
Gianmarco Caruso, Luca Tardella and Christian P. Robert

An ABC algorithm for random partitions arising from the Dirichlet process. 632
Mario Beraha and Riccardo Corradin

Bayesian Inference of Undirected Graphical Models from Count Data. 638
Pier Giovanni Bissiri, Monica Chiogna and Nguyen Thi Kim Hue

Bayesian IRT models in NIMBLE. 644
Sally Paganin, Chris Paciorek and Perry de Valpine

Bayesian modelling of Facebook communities via latent factor models. 650
Emanuele Aliverti

Bayesian nonparametric adaptive classification with robust prior information. 655
Francesco Denti, Andrea Cappozzo and Francesca Greselin

Choosing the right tool for the job: a systematic analysis of general purpose MCMC software. 661
Mario Beraha, Giulia Gualtieri, Eugenia Villa, Riccardo Vitali and Alessandra Guglielmi

Empirical Bayes estimation for mixture models. 667
Catia Scricciolo

Improving ABC via Large Deviations Theory. 673
Cecilia Viscardi, Michele Boreale and Fabio Corradi

Learning Bayesian Networks for Nonparanormal Data. 679
Flaminia Musella and Vincenzina Vitale

Measuring well-being combining different data sources: a Bayesian networks approach. 685
Federica Cugnata, Silvia Salini and Elena Siletti

Penalising the complexity of extensions of the Gaussian distribution. 691
Diego Battagliese and Brunero Liseo

Predictive discrepancy of credible intervals for the parameter of the Rayleigh distribution. 697
Fulvio De Santis and Stefania Gubbiotti

Small-area statistical estimation of claim risk. 702
Francesca Fortunato, Fedele Greco and Pierpaolo Cristaudo

Subject-specific Bayesian Hierarchical model for compositional data analysis. 708
Matteo Pedone and Francesco C. Stingo

Wasserstein consensus for Bayesian sample size determination. 714
Michele Cianfriglia, Tullia Padellini and Pierpaolo Brutti

Biostatistics 720

A comparison of the CAR and DAGAR spatial random effects models with an application to diabetics rate estimation in Belgium. 721
Vittoria La Serra, Christel Faes, Niel Hens and Pierpaolo Brutti

A functional approach to study the relationship between dynamic covariates and survival outcomes: an application to a randomized clinical trial on osteosarcoma. 727
Marta Spreafico, Francesca Ieva and Marta Fiocco

A Statistical Approach to the Alignment of fMRI Data. 733
Angela Andreella, Ma Feilong, Yaroslav Halchenko, James Haxby and Livio Finos

Adaptive clinical trials: Bayesian decision-theoretic and frequentist approaches for cost-effectiveness analysis. 739
Martin Forster and Marco Novelli

Bootstrap corrected Propensity Score: Application for Anticoagulant Therapy in Haemodialysis Patients. 745
Maeregu W. Arisido, Fulvia Mecatti and Paola Rebora

Combining multiple sources to overcome misclassification bias in epidemiological database studies. 751
Francesca Beraldi, Rosa Gini, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini

Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics. 756
Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Nicola Rares Franco, Tiziana Rancati and Catharine West

Graphical models for count data: an application to single-cell RNA sequencing. 762
Nguyen Thi Kim Hue, Monica Chiogna and Davide Risso

Interregional mobility, socio-economic inequality and mortality among cancer patients. 768
Claudio Rubino, Mauro Ferrante, Antonino Abbruzzo, Giovanna Fantaci and Salvatore Scondotto

PET radiomics-based lesions representation in Hodgkin lymphoma patients. 774
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

Prediction of late radiotherapy toxicity in prostate cancer patients via joint analysis of SNPs sequences. 780
Nicola Rares Franco, Michela Carlotta Massi, Francesca Ieva, Anna Maria Paganoni, Andrea Manzoni, Paolo Zunino, Tiziana Rancati and Catharine West

Predictive versus posterior probabilities for phase II trial monitoring. 785
Valeria Sambucini

Profile networks for precision medicine. 791
Andrea Lazzerini, Monia Lupparelli and Francesco C. Stingo

Proton-Pump Inhibitor Provider Profiling via Funnel Plots and Poisson Regression. 797
Dario Delle Vedove, Francesca Ieva and Anna Maria Paganoni

Selecting optimal thresholds in ROC analysis with clustered data. 803
Duc Khanh To, Gianfranco Adimari and Monica Chiogna

Environment, Physics and Engineering 809

A hidden semi-Markov model for segmenting environmental toroidal data. 810
Francesco Lagona and Antonello Maruotti

An experimental analysis on quality and security about green communication. 816
Vito Santarcangelo, Emilio Massa, Davide Scintu, Michele Di Lecce and Massimiliano Giacalone

An improved sensitivity-data based method for probabilistic ecological risk assessment. 822
Sonia Migliorati and Gianna Serafina Monti

Comparing predictive distributions in EMOS. 828
Giummolè Federica and Mameli Valentina

Compositional analysis of fish communities in a fast changing marine ecosystem. 834
Pierfrancesco Alaimo Di Loro, Marco Mingione, Giovanna Jona Lasinio, Sara Martino and Francesco Colloca

FDA dimension reduction techniques and components separation in Fourier-transform infrared spectroscopy. 840
Francesca Di Salvo, Elena Piacenza and Delia Francesca Chillura Martino

Functional Data Analysis for Spectroscopy Data. 846
Mara S. Bernardi, Matteo Fontana, Alessandra Menafoglio, Diego Perugini, Alessandro Pisello, Marco Ferrari, Simone De Angelis, Maria Cristina De Sanctis and Simone Vantini

Functional graphical model for spectrometric data analysis. 852
Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini

Local LGCP estimation for spatial seismic processes. 857
Nicoletta D'Angelo, Marianna Siino, Antonino D'Alessandro and Giada Adelfio

Observation-driven models for storm counts. 863
Mirko Armillotta, Alessandra Luati and Monia Lupparelli

Statistical control of complex geometries, with application to Additive Manufacturing. 869
Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio, Piercesare Secchi

Tree attributes map by 3P sampling in a design-based framework. 875
Lorenzo Fattorini and Sara Franceschi

Unsupervised classification of texture images by gray-level spatial dependence matrices and genetic algorithms. 880
Roberto Baragona and Laura Bocci

Finance, business and official statistics 886

A discrete choice approach to analyze contractual attributes in the durum wheat sector in Italy. 887
Stefano Ciliberti, Simone Del Sarto, Giulia Pastorelli, Angelo Frascarelli and Gaetano Martino

A fuzzy approach to the measurement of the employment rate. 893
Bruno Cheli, Alessandra Coli and Andrea Regoli

A proposal to model credit risk contagion using network count-based models. 898
Arianna Agosto and Daniel Felix Ahelegbey

A similarity matrix approach to empower ESCO interfaces for testing, debugging and in support of users' experience. 904
Adham Kahlawi, Cristina Martelli, Lucia Buzzigoli, Laura Grassini

Adding MIDAS terms to Linear ARCH models in a Quantile Regression framework. 910
Vincenzo Candila and Lea Petrella

Company requirements in Italian tourism sector: an analysis for profiles. 916
Paolo Mariani, Andrea Marletta, Lucio Masserini and Mariangela Zenga

Determinants of Firms' Default Risk after the 2008 and 2011 Economic Crises: a Latent Growth Models Approach. 921
Lucio Masserini, Matilde Bini and Alessandro Zeli

Double Asymmetric GARCH-MIDAS model - new insights and results. 927
Alessandra Amendola, Vincenzo Candila and Giampiero M. Gallo

European SMEs and Circular Economy Activities: Evaluating the Advantage on Firm Performance through the Estimation of Average Treatment Effects. 933
Luca Secondi

Financial Spillover Measures to Assess the Stability of Basket-based Stablecoins. 939
Paolo Pagnottoni

Forecasting Banknote Flows in BdI Branches: Speed-up with Machine Learning. 945
Marco Brandi, Monica Fusaro, Tiziana Laureti and Giorgia Rocco

Fully reconciled GDP forecasts from Income and Expenditure sides. 951
Luisa Bisaglia, Tommaso Di Fonzo and Daniele Girolimetto

GLASSO Estimation of Commodity Risks. 957
Beatrice Foroni, Saverio Mazza, Giacomo Morelli and Lea Petrella

Measuring the Effect of Unconventional Policies on Stock Market Volatility. 963
Giampiero M. Gallo, Demetrio Lacava and Edoardo Otranto

Multidimensional versus unidimensional poverty measurement. 969
Michele Costa

Multiple outcome analysis of European Agriculture in 2000-2016: a latent class multivariate trajectory approach. 975
Alessandro Magrini

Nowcasting GDP using mixed-frequency based composite confidence indicators. 981
Maria Carannante, Raffaele Mattera, Michelangelo Misuraca, Germana Scepi and Maria Spano

On the tangible and intangible assets of Initial Coin Offerings. 987
Paola Cerchiello and Anca Mirela Toma

Seasonality variation of electricity demand: decompositions and tests. 993
Luigi Grossi and Mauro Mussini

SMEs circular economy practices in the European Union: Implications for sustainability. 999
Nunzio Tritto, Josè G. Dias and Francesca Bassi

Tax Incentives' Effect on the Provision of Occupational Welfare in Italian Enterprises. 1005
Alessandra Righi

The determinants of eco-innovation: a country comparison using the community innovation survey. 1011
Ida D'Attoma and Silvia Pacei

World ranking of urban sustainability through composite indicators. 1017
Elena Grimaccia, Alessia Naccarato and Silvia Terzi

Machine Learning and Data Science..1023

A novel approach for Artificial Intelligence through Lorenz zonoids and Shapley Values. 1024
Paolo Giudici and Emanuela Raffinetti

A warning signal for variable importance interpretation in tree-based algorithms. 1030
Anna Gottard and Giulia Vannucci

Assessment of the effectiveness of digital flyers: analysis of viewing behavior using eye tracking. 1036
Gianpaolo Zammarchi, Claudio Conversano and Francesco Mola

At risk mental status analysis: a comparison of model selection methods for ordinal target variable. 1042
Elena Ballante, Silvia Molteni, Martina Mensi and Silvia Figini

Categorical Encoding for Machine Learning. 1048
Agostino Di Ciaccio

Dynamic Quantile Regression Forest. 1054
Mila Andreani and Lea Petrella

Estimating the UK Sentiment Using Twitter. 1059
Stephan Schlosser, Daniele Toninelli and Michela Cameletti

Forecasting local rice prices from crowdsourced data in Nigeria. 1065
Ilaria Lucrezia Amerise and Gloria Solano Hermosilla

Generalized Mixed Effects Random Forest: does Machine Learning help in predicting university student dropout? 1071
Massimo Pellagatti, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

HateViz: a textual dashboard Twitter data-driven. 1077
Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Nicola Canestrari

How to perform cyber risk assessment via cumulative logit models. 1083
Silvia Facchinetti, Silvia Angela Osmetti and Claudia Tarantola

Machine learning prediction for accounting system. 1087
Chiara Bardelli and Silvia Figini

Teaching statistics: an assessment framework based on Multidimensional IRT and Knowledge Space Theory. 1093
Cristina Davino, Rosa Fabbricatore, Carla Galluccio, Daniela Pacella, Domenico Vistocco, Francesco Palumbo

The weight of words: textual data versus sentiment analysis in stock returns prediction. 1099
Riccardo Ferretti and Andrea Sciandra

Unsupervised Energy Trees: clustering with complex and mixed-type variables. 1105
Riccardo Giubilei, Tullia Padellini and Pierpaolo Brutti

Using anchoring vignettes to adjust self-reported life satisfaction: a nonparametric approach leading to a Semantic Differential scale. 1111
Sara Garbin, Serena Berretta, Maria Iannario and Omar Paccagnella

Variable selection for robust model-based learning from contaminated data. 1117
Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy

Variable Selection in Text Regressions: Back to Lasso? 1123
Marzia Freo and Alessandra Luati

Web Usage Mining and Website Effectiveness. 1129
Maria Francesca Cracolici and Furio Urso

Models and methods - Categorical, Ordinal, Rank Data 1135

Aberration for the analysis of two-way contingency tables. 1136
Roberto Fontana and Fabio Rapallo

An investigation of the paradoxical behaviour of κ-type inter-rater agreement coefficients for nominal data. 1142
Amalia Vanacore and Maria Sole Pellegrino

Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach. 1148
Luigi Lombardi and Antonio Calcagnì

BOD – min range: A Robustness Analysis Method for Composite Indicators. 1154
Emiliano Seri, Leonardo Salvatore Alaimo and Vittoria Carolina Malpassuti

Comparing classifiers for ordinal variables. 1160
Silvia Golia and Maurizio Carpita

Discovering Interaction Effects Between Subject-Specific Covariates: A New Probabilistic Approach For Preference Data. 1166
Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio, Mark De Rooij and Elise Dusseldorp

Hybrid random forests for ordinal data. 1171
Rosaria Simone and Gerhard Tutz

Model-based approach to biclustering ordinal data. 1177
Monia Ranalli and Francesca Martella

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution. 1183
Cristina Mollica and Luca Tardella

Non-metric unfolding on augmented data matrix: a copula-based approach. 1189
Marta Nai Ruscone and Antonio D'Ambrosio

Ordinal probability effect measures for dyadic analysis in cumulative models. 1194
Maria Iannario and Domenico Vistocco

Simulated annealing for maximum rater agreement. 1200
Fabio Rapallo and Maria Piera Rogantin

Models and methods – Regression..1206

A Clusterwise regression method for Distributional-valued Data. 1207
Rosanna Verde, Francisco de A. T. de Carvalho and Antonio Balzanella

A nonparametric approach for nonlinear variable screening in high-dimensions. 1213
Francesco Giordano, Sara Milito and Lucia Maria Parrella

Adjusted scores for inference in negative binomial regression. 1219
Euloge C. Kenne Pagui, Alessandra Salvan and Nicola Sartori

Estimation of the treatment effect variance in a difference-in-differences framework. 1224
Marco Doretti and Giorgio E. Montanari

Exploring multicollinearity in quantile regression. 1230
Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco

Generalized M-quantile random effects model. 1236
Francesco Schirripa Spagnolo and Vincenzo Mauro

Goodness-of-fit assessment in linear quantile regression. 1242
Ilaria Lucrezia Amerise and Agostino Tarsitano

Joint Redundancy Analysis by a multivariate linear predictor. 1248
Laura Marcis and Renato Salvatore

M-quantile regression shrinkage and selection via the lasso. 1254
M. Giovanna Ranalli, Lea Petrella and Francesco Pantalone

New insights into the Conditioning and Gain Score approaches in multilevel analysis. 1260
Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto and Carla Rampichini

Simultaneous confidence regions and curvature measures in nonlinear models. 1265
Claudia Furlan and Cinzia Mortarino

Models and methods – Sampling 1271

Design-based consistency of the Horvitz-Thompson estimator for spatial populations. 1272
Lorenzo Fattorini, Marzia Marcheselli, Caterina Pisani and Luca Pratelli

Empirical likelihood in the statistical matching for informative samples. 1278
Daniela Marella and Danny Pfeffermann

Evaluating a Hybrid One-Staged Snowball Sampling through Bootstrap Method on a Simulated Population. 1284
Venera Tomaselli and Giulio Giacomo Cantone

How optimal subsampling depends on guessed parameter values. 1290
Laura Deldossi and Chiara Tommasi

Indicators for risk of selection bias in non-probability samples. 1296
Emilia Rocco and Alessandra Petrucci

On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. 1302
Caterina May and Chiara Tommasi

Pseudo-population based resamplings for two-stage design. 1308
Pier Luigi Conti, Daniela Marella and Vincenzina Vitale

Models and methods - Theoretical Issues in Statistical Inference 1314

A new mixture model for three-way data. 1315
Salvatore D. Tomarchio, Antonio Punzo and Luca Bagnato

A Sequential Test for the Cpmk Index. 1320
Michele Scagliarini

Probability Interpretations and the Selection of the Most Effective Statistics Method. 1326
Paolo Rocchi

Robust Composite Inference. 1332
Valentina Mameli, Monica Musio, Erlis Ruli and Laura Ventura

Statistical hypothesis testing within the Generalized Error Distribution: Comparing the behavior of some nonparametric techniques. 1338
Massimiliano Giacalone and Demetrio Panarello

Stochastic dependence with discrete copulas. 1344
Fabrizio Durante and Elisa Perrone

Models and methods - Time Series and Longitudinal Data.....................1350

Bootstrap test in Poisson–INAR models. 1351
Lucio Palazzo and Riccardo Ievoli

Continuous Time-Interaction Processes for Population Size Estimation. 1357
Linda Altieri, Alessio Farcomeni, Danilo Alunni Fegatelli and Francesco Palini

Longitudinal data analysis using PLS-PM approach. 1363
Rosanna Cataldo, Corrado Crocetta, Maria Gabriella Grassia and Marina Marino

Long-memory models for count time series. 1369
Luisa Bisaglia, Massimiliano Caporin and Matteo Grigoletto

Combining multiple frequencies in Realized GARCH models. 1375
Antonio Naimoli and Giuseppe Storti

Models with Time-Varying Parameters for Realized Covariance. 1381
Luc Bauwens and Edoardo Otranto

Pitman-Yor mixture models for survival data stratification. 1387
Riccardo Corradin, Luis Enrique Nieto Barajas and Bernardo Nipoti

Prediction is not everything, but everything is prediction. 1393
Leonardo Egidi

The Generalized Dynamic Mixtures of Factor Analyzers for clustering multivariate longitudinal data. 1399
Francesca Martella, Antonello Maruotti and Francesco Tursini

Trends and long-run relations in cointegrated time series observed with noise. 1405
Angelica Gianfreda, Paolo Maranzano, Lucia Parisio and Matteo Pelagatti

Population and society 1411

A dimensionality assessment of refugees' vulnerability through an Item Response Theory approach. 1412
Simone Del Sarto, Michela Gnaldi, Yara Maasri and Edouard Legoupil

Accounting for Interdependent Risks in Vulnerability Assessment of Refugees. 1418
Daria Mendola, Anna Maria Parroco and Paolo Li Donni

Active ageing in China: What are the domains that most affect life satisfaction in the elderly? 1424
Ilaria Rocco

Analyzing the waiting time of academic publications: a survival model. 1430
Francesca De Battisti, Giuseppe Gerardi, Giancarlo Manzi and Francesco Porro

Clustering of food choices in a large sample of students using university canteen. 1436
Valentina Lorenzoni, Isotta Triulzi, Irene Martinucci, Letizia Toncelli, Michela Natilli and Roberto Barale, Giuseppe Turchetti

Cruise passengers' expenditure at destinations: Review of survey techniques and data collection. 1442
Caterina Sciortino, Stefano De Cantis, Mauro Ferrante and Szilvia Gyimóthy

Educational integration of foreign citizen children in Italy: a synthetic indicator. 1448
Alessio Buonomo, Stefania Capecchi and Rosaria Simone

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. 1454
Giorgio Calzolari, Maria Gabriella Campolo, Antonino Di Pino and Laura Magazzini

First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. 1460
Antonella D'Agostino, Giulio Ghellini and Gabriele Lombardi

Future Scenarios and Support Interventions for the Family: Involving Experts' Participation through a Mixed-Method Research Study. 1466
Mario Bolzan, Simone Di Zio, Manuela Scioni and Morena Tartari

Gender and Monetary Policy Preferences: a Diff-in-Diff Approach. 1472
Donata Favaro, Anna Giraldo and Ina Golikja

Headcount based indicators and functions to evaluate the effectiveness of Italian university education. 1478
Silvia Terzi and Francesca Petrarca

Identify the speech code through statistics: a data-driven approach. 1484
Andrea Briglia, Massimo Mucciardi and Jérémi Sauvage

Inspecting cause-specific mortality curves by simplicial functional data analysis. 1490
Marco Stefanucci and Stefano Mazzuco

Intertemporal decision making and childless couples. 1495
Daniela Bellani, Bruno Arpino and Daniele Vignoli

Italian Households' Material Deprivation: Multi-Objective Genetic Algorithm approach for categorical variables. 1501
Laura Bocci and Isabella Mingo

LI-CoD Model. From Lifespan Inequality to Causes of Death. 1507
Andrea Nigri and Susanna Levantesi

Modeling Well-Being through PLS-SEM and K-M. 1513
Venera Tomaselli, Mario Fordellone and Maurizio Vichi

News life-cycle: a multiblock approach to the study of information. 1519
Rosanna Cataldo, Marco Del Mastro, Maria Gabriella Grassia, Marina Marino and Rocco Mazza

Short-term rentals in a tourist town. 1525
Silvia Bacci, Bruno Bertaccini, Gianni Dugheri, Paolo Galli, Antonio Giusti and Veronica Sula

SportIstat: a playful activity to developing statistical literacy. 1531
Alessandro Valentini and Francesca Paradisi

Statistical modeling for some features of Airbnb activity. 1537
Giulia Contu and Luca Frigau

Tertiary students with migrant background: evidence from a cohort enrolled at Sapienza University. 1543
Cristina Giudici, Donatella Vicar and Eleonora Trappolini

The Causal Effect of Immigraton Policies on Income Inequality. 1549
Irene Crimaldi, Laura Forastiere, Fabrizia Mealli and Costanza Tortù

The job condition of academic graduates: a joint longitudinal analysis of AlmaLaurea
and Mandatory Notices of the Ministry of Labour. 1557
Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci

The joint effect of childcare services and flexible female employment on fertility rate in Europe. 1565
Viviana Cocuccio and Massimo Mucciardi

The Left Behind Generation: How the current Early School Leavers affect tomorrow's NEETs? 1571
Giovanni De Luca, Paolo Mazzocchi, Claudio Quintano and Antonella Rocca

The probability to be employed of young adults of foreign origin. 1577
Alessio Buonomo, Francesca Di Iorio and Salvatore Strozza

The risk of inappropriateness in geriatric wards: a comparison among the Italian regions. 1583
Paolo Mariani, Andrea Marletta, Marcella Mazzoleni and Mariangela Zenga

The role of the accumulation of poverty and unemployment for health disadvantages. 1589
Annalisa Busetta, Daria Mendola, Emanuela Struffolino and Zachary Van Winkle

Unemployment and fertility in Italy. A regional level data panel analysis. 1595
Gabriele Ruiu and Marco Breschi

University drop out and mobility in Italy. First evidences on first level degrees. 1601
Nicola Tedesco and Luisa Salaris

Worthiness-based Scale Quantifying. 1607
Giulio D'Epifanio

Young people in Southern Italy and the phenomenon of immigration: what is their perception? 1613
Nunziata Ribecco, Angela Maria D'Uggento and Angela Labarile

SIS2020

Pisa

Preface

The COVID-19 pandemic is putting our society under incredible health, emotional, and economic stress. Facing its harmful effects and their uncertainty, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee, to ensure the highest level of safety for members and delegates, deliberated to cancel the 50th Meeting of the Italian Statistical Society originally planned to be held in Pisa in June 2020 and to postpone the conference to June 2021. The Executive Board and the Local Organizing Committee continue to monitor closely the pandemic evolving situation, and keep the members of SIS and the researchers informed about the potential new dates for the next meeting. To give value to the work of those who prepared their presentation for the conference, the Program Committee decided to publish the volume *Book of short papers - SIS 2020* despite the conference cancellation.

The conference program included 4 plenary sessions, 16 specialized sessions, 24 solicited sessions, 32 contributed sessions and the poster exhibition. Plenary sessions concerned with robust statistics, human longevity, statistical models for climate changes and small area estimation for educational poverty. The meeting had to host also 2 round tables on data privacy and innovation in statistics. Activities focused on topics of interest for a wider audience included two round tables on Teaching Statistics and on the SIS journal Statistical Methods & Applications, and the Stats Under the Stars (SUS6) competition for young statisticians. The SUS6 event attracted many sponsors from statistical, financial and editorial firms as well as numerous students. The conference committee had registered 345 accepted submissions, including 143 to be presented in invited plenary, specialized and solicited sessions, and 202 spontaneously submitted for oral and poster sessions.

This book includes most of the scientific contributions that had to be presented at the 50th Meeting of the Italian Statistical Society. It is organized into 49 chapters corresponding to 15 specialized, 23 solicited sessions, and to 11 general topics for contributed papers and posters. All 268 contributions provide a wide overview of the state-of-the-art of the subjects, from methodological and theoretical contributions, to applied works and case studies. The result is a very lively picture of the Italian statistical community with its international connections.

We would like to thank all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Although the Conference did not take place, the organization went on until cancellation was decided for safety reasons. It would have been impossible without the joint effort of Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members these three institutions took part actively in the Local Organizing Committee. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

This book is our contribution to encourage the scientific community and the network of the Italian Statistical Society to go on and transform this difficult period into an opportunity of scientific debate for better statistics in a better world.

Alessio Pollice
Università degli Studi di Bari Aldo Moro
Chair of the Program Committee

Nicola Salvati
Università di Pisa
Chair of the Local Organizing Committee

Francesco Schirripa Spagnolo
Università di Pisa

Program Committee: Alessio Pollice (Chair), Serena Arima, Marilena Barbieri, Alessandra Brazzale, Eugenio Brentari, Alessia Caponera, Antonio Lepore, Antonella Plaia, Tommaso Proietti, MarcoRiani, Nicola Salvati, Pasquale Sarnacchiaro, Mauro Scanu, Manuela Stranges, Valentina Tocchioni, Simone Vantini, Massimo Ventrucci, Paola Vicard, Donatella Vicari.

Local Organizing Committee: Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Alessandra Coli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

Organizers of Specialized and Solicited Sessions: Giada Adelfio, Bruno Arpino, Emanuele Aliverti, Nicoletta Balbo, Mara Bernardi, Silvia Bozza, Pierpaolo Brutti. Annalisa Busetta, Michela Cameletti, Carlo Cavicchia, Fabrizio Durante, Leonardo Egidi, Pietro D. Falorsi, Francesco Finazzi, Livio Finos, Stefania Galimberti, Michele Gallo, Caterina Giusti, Francesca Greselin, Alessandra Guglielmi, Francesca Ieva, Tiziana Laureti, Achille Lemmi, Brunero Liseo, Fabio Massimo Lo Verde, Daria Mendola, Roberta Pappadà, Lea Petrella, Alessandra Petrucci, Alessia Pini, Sabrina Prati, Maria Giovanna Ranalli, Davide Risso, Fabrizio Ruggeri, Silvana Salvini, Monica Scannapieco, Francesco Stingo, Luca Tardella, Grazia Vicario, Susanna Zaccarin, Maroussa Zagoraiou.

SIS2020
Pisa

Specialized sessions

Aggregating Gaussian mixture components

Come aggregare le componenti gaussiane di un miscuglio

Roberto Rocci[1]

Abstract The finite mixture of Gaussians is a well-known model frequently used to classify a sample of observations. It considers the sample as drawn from a heterogeneous population where each subpopulation, cluster, is Gaussian and corresponds to one component of the mixture. Whenever such assumption is false, the model may use two or more Gaussians to describe a single cluster. In this case, the researcher has the problem of how to identify the clusters starting from the estimated components. This work proposes to solve this problem by aggregating the components in clusters by optimizing an appropriate criterion based on their posterior probabilities.

Abstract *I modelli miscuglio di gaussiane sono spesso utilizzati nell'analisi dei gruppi. L'idea è quella di considerare la popolazione che ha generato il campione come formata da sottopopolazioni, gruppi, ognuna ben descritta da una componente del miscuglio. Quando questa assunzione risulta falsa, il modello tende ad utilizzare due o più componenti per rappresentare un unico gruppo, creando così il problema di come identificare i gruppi a partire dalle componenti stimate. In questo lavoro proponiamo di risolvere il problema aggregando le componenti in modo da ottimizzare un criterio basato sulle loro probabilità a posteriori.*

Key words: Unsupervised Classification, Finite mixtures of Gaussians, Within and Between deviances.

1 Introduction

In cluster analysis, or unsupervised classification, quite frequently observations are classified by using a finite mixture of Gaussians (see for example Hennig et al., 2015). The idea is to consider the population as heterogeneous, i.e. formed by sub-populations, clusters, which are well represented by the mixture components.

[1] Department of Statistical Science, Sapienza University of Rome; email: roberto.rocci@uniroma1.it

Such approach works well in practice unless one or more subpopulations have a distribution different from, or not well approximated by, a single component of the mixture. In this case, the model may use more than one component, i.e. a sub-mixture of two or more Gaussians, to describe a single cluster destroying the one to one correspondence between clusters and components. This problem is well recognized in practice and several solutions have been proposed.

The first idea is to assume for each component a functional form that is more flexible than the Gaussian (see McNicholas, 2016, for an excellent review about this approach). This solves the problem in many cases in practice. However, it cannot be considered as the definitive solution because even in this case we cannot exclude that more than one component could be necessary to represent a cluster. This derives from the identifiability of the finite mixture model. For example, a cluster that is a finite mixture of two components cannot be represented by only one and vice versa.

A different idea comes up by observing that, if the mixture fits well the data, then the information about the clustering structure is contained in the estimated model and it can be recovered by aggregating, in an opportune way, the components. This allow us to represent a very wide variety of possible distributions for the clusters and to relax the, usually made and sometimes restrictive, assumption of same functional form for the distribution of each cluster. Technically, the model would become a finite mixture of finite mixtures of Gaussians, i.e. a finite mixture where each component is a finite mixture of Gaussians. Unfortunately, such a model is not identified because different aggregations of the components give the same population distribution. The estimation is then possible only by using some constraints on model parameters making the model identified (see for example Di Zio et al. 2007), or by introducing a criterion determining the aggregation. On the latter approach, there are several proposals in the literature where the Gaussian components are hierarchically aggregated into clusters on the basis of a measure of proximity (see Hennig 2010, Comas-Cufi et al. 2017 and references there in for some examples). However, such methods are optimal only locally. They establish what is the best way to merge two components not what is the best way to aggregate, say G, components into, say K, clusters. It is not specified how the “internal cohesion” and “external isolation” (Cormack, 1971) are measured, related and optimized (e.g. total deviance = within deviance + between deviance in K-means). To achieve this goal a partitioning method should be adopted but, as far as we know, only Li (2005) considered a partitioning approach based on the application of the K-means on the mean components. This proposal makes clear the aforementioned aspects but it is based on a measure of dissimilarity between components depending only on their locations. Our purpose is to go beyond the limits of this proposal.

In our paper we are going to propose a new method to aggregate the Gaussian components of a finite mixture model making clear how the identified partition optimize the “internal cohesion” and “external isolation” of the clustering. The plan of the paper is the following. Our proposal, based on the use of the Kullback Leibler divergence to measure the dissimilarity among components, will be presented in section 2. In section 3, some insights on how to extend the technique are presented with particular reference to other dissimilarity measures and its hierarchical version.

2 Partitioning Gaussian components by the Kullback-Liebler divergence: the KL-components method

The finite mixture of Gaussians (McLachlan & Peel, 2000) is based on the assumption that the probability density of a multivariate observation is of the form

$$f(\mathbf{x}_i;\Theta)=\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_i;\vartheta_g)\ . \tag{1}$$

where $\mathbf{x}_i = [x_{i1}, x_{i2},\ldots,x_{iJ}]'$ is a random vector of J variables sampled from a population parametrized by $\Theta = \{\vartheta_1,\vartheta_2,\ldots,\vartheta_G, p_1,p_2,\ldots,p_G\}$, which consists of G groups, or subpopulations, in proportions $\pi_1, \pi_2,\ldots, \pi_G$, where π_g is the prior probability to sample one observation from group g. The density $f_g(\mathbf{x}_i;\vartheta_g)$ of $\mathbf{x}_i$ in the g^{th} group is multivariate normal (Gaussian). Let $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ be a sample of n independent and identically distributed observations, we can use the above model to classify the observations into G classes. First, we compute the posterior probabilities

$$\mathrm{P}(g\,|\,\mathbf{x}_i)=\pi_{g|i}=\pi_g f_g(\mathbf{x}_i;\vartheta_g)\,/\sum_{h=1}^{G}\pi_h f_h(\mathbf{x}_i;\vartheta_h)\,, \tag{2}$$

then, we use the MAP rule (Maximum A Posterior probability) to assign the observations to the Gaussian components. Usually, the parameter Θ and the number of components G are unknown and estimated from the data.

Our method, named the KL-components technique, originates from the observation that two components, say g and h, are equal, with respect to the data, if and only if their posterior probabilities are proportional. In formulas

$$g=h\Leftrightarrow \boldsymbol{\pi}_g=\mathrm{P}(g\,|\,\mathbf{x}_i)/\pi_g=\mathrm{P}(h\,|\,\mathbf{x}_i)/\pi_h=\boldsymbol{\pi}_h\,,\ g,h = 1,\ldots,G \text{ and } i = 1,\ldots,n. \tag{3}$$

It seems quite natural to measure the dissimilarity between components as a function of the diversity between the normalized posteriors, say profiles. In particular, we investigate the use of the Kullback-Leibler (KL) divergence

$$\mathrm{KL}(\boldsymbol{\pi}_g,\boldsymbol{\pi}_h)=\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_{g|i}}{\pi_g}\log\left(\frac{\pi_{g|i}}{\pi_g}\frac{\pi_h}{\pi_{h|i}}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{f_g(\mathbf{x}_i)}{f(\mathbf{x}_i)}\log\left(\frac{f_g(\mathbf{x}_i)}{f_h(\mathbf{x}_i)}\right), \tag{4}$$

which can be considered an estimate of

$$\mathrm{E}\left(\frac{f_g(\mathbf{x})}{f(\mathbf{x})}\log\left(\frac{f_g(\mathbf{x})}{f_h(\mathbf{x})}\right)\right)=\int f_g(\mathbf{x})\log\left(\frac{f_g(\mathbf{x})}{f_h(\mathbf{x})}\right)d\mathbf{x}\,, \tag{5}$$

i.e. the KL divergence between the two components. It is well known, and evident from formula (5), that the KL divergence is not symmetric. However, this is not a problem for us because we use it to define a sort of deviance rather than to measure the dissimilarity between two components. In particular, we define the Total deviance as the weighted sum of the KL divergences among the profiles of the mixture components and their barycenter. By noting that the latter quantity is

$$\sum_{g=1}^{G} \pi_g \frac{\pi_{g|i}}{\pi_g} = 1, \quad i = 1,2,...,n, \tag{6}$$

the Total deviance results to be

$$\mathrm{D}_T = \sum_{\mathrm{g}=1}^{\mathrm{G}} \pi_g \mathrm{KL}(\boldsymbol{\pi}_g, \mathbf{1}) = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n} \pi_{g|i} \log\left(\frac{\pi_{g|i}}{\pi_g}\right). \tag{7}$$

Given a partition of the G components into K clusters according to the binary row stochastic membership matrix $\mathbf{U} = [u_{gk}]$, where u_{gk} is equal to 1 if component g belong to cluster k and 0 otherwise, we note that the posteriors probabilities of cluster k are

$$p_{k|i} = \sum_{g=1}^{G} u_{gk} \pi_{g|i} \,, \tag{8}$$

the priors are

$$p_k = \sum_{g=1}^{G} u_{gk}\pi_g = \sum_{g=1}^{G} u_{gk} \frac{1}{n}\sum_{i=1}^{n} \pi_{g|i} = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G} u_{gk}\pi_{g|i} = \frac{1}{n}\sum_{i=1}^{n} p_{k|i} \,, \tag{9}$$

and the barycenter is

$$\frac{1}{\sum_{g=1}^{G} u_{gk}\pi_g} \sum_{g=1}^{G} u_{gk}\pi_g \frac{\pi_{g|i}}{\pi_g} = \frac{1}{p_k}\sum_{g=1}^{G} u_{gk}\pi_{g|i} = \frac{p_{k|i}}{p_k}, \quad i = 1,2,...,n. \tag{10}$$

According to (7), the Within deviance is defined as

$$\mathrm{D}_W(\mathbf{U}) = \sum_{k,g} u_{gk}\pi_g \mathrm{KL}(\boldsymbol{\pi}_g, \mathbf{p}_k) = \sum_{k,g} u_{gk}\pi_g \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_{g|i}}{\pi_g}\left[\log\left(\frac{\pi_{g|i}}{\pi_g}\right) - \log\left(\frac{p_{k|i}}{p_k}\right)\right]. \tag{11}$$

Formula (11) suggests in a very natural way a partitioning method based on its minimization with respect to $\mathbf{U}$. The minimization of (11) guarantees the maximum internal cohesion of the clusters. However, we should ask: what about the external isolation? To give an answer to this question, we should find a way to measure the external isolation, i.e. the cluster separation. In coherence with (7), we can define the Between deviance as the weighted sum of the KL divergences among the cluster profiles and their barycenter, that is the vector of ones even in this case. In formulas

$$\mathrm{D}_B(\mathbf{U}) = \sum_{k=1}^{K} p_k \mathrm{KL}(\mathbf{p}_k, \mathbf{1}) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} p_{k|i} \log\left(\frac{p_{k|i}}{p_k}\right). \tag{12}$$

Once again, we have a formula suggesting, in a very natural way, a partitioning method corresponding to the maximization of (12) with respect to **U**. However, it is possible to show that, as in the case of *K*-means, the minimization of D_W is equivalent to the maximization of D_B because the sum of the two is constant and equal to D_T. The Within deviance (11) can be minimized by using a coordinate descent algorithm that minimize (11) with respect to **U** and the centroids $\mathbf{p}_1$, $\mathbf{p}_2$, …, $\mathbf{p}_K$.

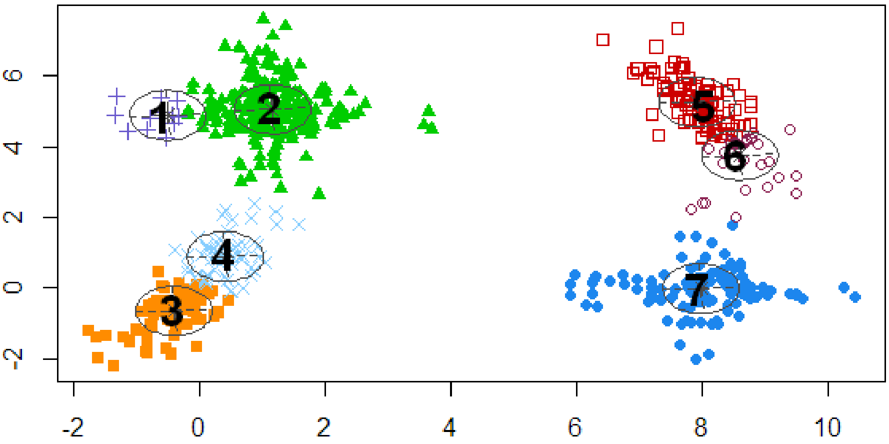

Figure 1: Simulated sample from a 6-component mixture along with the estimated classification in 7 components given by a homoscedastic mixture of Gaussians.

In order to check if the proposed method works properly, we considered the simulated dataset analysed in section 4.1 of Baudry et al. (2010). It is a sample from a finite mixture of 6 bivariate Gaussians. In Figure 1, the data is shown along with the classification in 7 components given by a homoscedastic mixture of Gaussians where the number of components have been selected by using BIC and the parameters estimated by maximum likelihood. Looking at the figure, it is not clear if the number of true clusters is 2 or 4 and then we considered both. The algorithm has been run for *K* = 2 and 4, from several different starting points. The technique aggregated the 7 components as {1,2,3,4} and {5,6,7} for *K* = 2 and as {1,2}, {3,4}, {5,6} and {7} for *K* = 4. From Figure 1, it is clear that in both cases, KL-components has been able to find the correct aggregation of the components.

3 Final comments and extensions

The method here presented has been extended to the use of divergences different from the KL. In particular, we have proven that all the properties of the KL-components technique shown in the previous section do hold if a Bregman divergence (Bregman, 1967) is used. The proofs are not reported here for the sake of space.

In practical applications, especially when the clusters of components are not well separated, the coordinate descent algorithm quite often remains trapped into a local optimum. Frequently, this problem can be simply solved by starting several times the algorithm from different random partitions. However, there is still the need to have a method able to produce good rational, non random, starting points. To this end, a hierarchical clustering procedure has been proposed, not shown here for the sake of space, where at each step two clusters are merged by minimizing the increment of Within deviance. The hierarchical solution is then used to start the partitioning algorithm.

The aim of our method is not to find the true number of clusters, even if it can help us in this task exploring the possible components aggregations. In this respect, further insights can be obtained by looking at the plot of D_W vs K, computing the Calinski-Harabasz index (1974) or any other index based on D_W and/or D_B.

We conclude by noting that the results presented here do not use the assumption that the components of the mixture are normally distributed. It follows that the proposed techniques can be used to cluster components that are not Gaussians.

References

1. Baudry, J.P., Raftery, A., Celeux, A., Lo, K., Gottardo, R.: Combining mixture components for clustering, *Journal of Computational and Graphical Statistics*, 19, 332-353 (2010).
2. Bregman, L. M.: The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 7: 200–217 (1967)
3. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis.Com-munications in Statistics, 3, no. 1:1–27 (1974).
4. Comas-Cufi M., Martin-Fernandez J.A., Mateu-Figueras G.: Merging the components of a finite mixture using posterior probabilities, *Statistical Modelling*, 19(2), 1–31 (2017)
5. Cormack, R. M.: A review of classification (with discussion). *Journal of the Royal Statistical Society*, A, 134, 321-67 (1971)
6. Di Zio M., Guarnera U., Rocci R.: A mixture of mixture models for a classification problem: the unity measure error. *Computational Statistics and Data Analysis*, 51, 5, 2573-2585 (2007)
7. Hennig, C.: Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* 4, 3–34 (2010)
8. Hennig, C., Meila, M., Murtagh, F., Rocci R.: *Handbook of Cluster Analysis*, Chapman and Hall/CRC (2015)
9. Li J.: Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 14:547–568 (2004)
10. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
11. McNicholas, P.D.: *Mixture Model-Based Classification*, Boca Raton FL: Chapman & Hall/CRC Press (2016)