SIS2016

Università degli Studi di Salerno

June 8th – June 10th, 2016

PROCEEDINGS

of the 48th scientific meeting of the
Italian Statistical Society

Editors: Monica Pratesi and Cira Pena

ISBN: 9788861970618

PLENARY SESSIONS

(A) E. Baldacci — Financial Crises and their Impacts: Data Gaps and Innovation in Statistical Production.

(B) D. Dunson — Probabilistic inference from big and complex data.

(C) S. Strozza — Foreign immigration in Italy: a forty-year-old history.

SPECIALIZED SESSION (SPE)

(SPE-01) Inference, sampling and survey design

P. Conti — Resampling from finite populations under complex designs: the pseudo-population approach. (Co-author(s): F. Andreis, D. Marella, F. Mecatti)

P. Righi — A joint use of model based and design based frameworks for defining optimal sampling designs. (Co-author(s): P. D. Falorsi)

A. Ruiz-Gazen — A unified approach for robustness in survey sampling. (Co-author(s): J. Beaumont, D. Haziza)

(SPE-02) Multivariate models for risk assessment

M. Billio — A Bayesian nonparametric approach to macroeconomic risk. (Co-author(s): R. Casarin, M. Costola, M Guindani)

P. Cerchiello — Bank risk contagion:an analysis through big data. (Co-author(s): P. Giudici, G. Nicola)

L. De Angelis — A Markov-switching regression model with non-Gaussian innovations for systemic risk measurement. (Co-author(s): C. Viroli)

(SPE-03) Bayesian nonparametrics

D. Durante — Bayesian Nonparametric Modeling of Dynamic International Relations. (Co-author(s): D. Dunson)

A. Guglielmi — Bayesian autoregressive semiparametric models for gap times of recurrent events. (Co-author(s): G. Paulon, M. De Iorio)

A. Rodriguez — Restricted Nonparametric Mixtures models for Disease Clustering. (Co-author(s): T. Xifara)

(SPE-04) Statistical methods for the analysis of gene-environment interaction in the study of complex pathologies

C. Angelini — An introduction to next generation sequencing for studying omic-environment interactions.

L. Calciano — Statistical approaches for the evaluation of genetic associations in complex diseases: the heterogeneity of asthma phenotypes. (Co-author(s): L. Portas, S. Accordini)

Y. Pankaj — Improved case-only approach to study genome-wide gene-environment interaction. (Co-author(s): S. Freitag-Wolf, A. Dempfle, W. Lieb, M. Krawczak)

(SPE-05) Nonlinear time series

M. Niglio — Probabilistic properties of Self Exciting Threshold Autoregressive processes. (Co-author(s): F. Giordano, C. D. Vitale)

T. Proietti — Optimal prediction of stochastic trends. (Co-author(s): A. Giovannelli)

H. Tong — On model selection from a finite family of possibly misspecified models. (Co-author(s): H. Hsu, C. Ing)

(SPE-06) Spatial analyses in demography

F. Heins — Measuring residential segregation with spatial indices: an appraisal and applications for the metropolitan area of Rome. (Co-author(s): F. Benassi, F. Lipizzi, E. Paluzzi)

A. Mazza — Immigrants' settlement patterns in the city of Naples. (Co-author(s): G. Gabrielli, S. Strozza)

L. Natale — Native Immigration and Pull Factor Evolution in Italy: a Spatial Approach. (Co-author(s): A. Santacroce, F. G. Truglia)

(SPE-07) Recent developments in Volatility modeling

R. Casarin — Dynamic Model Averaging for Quantile Regression. (Co-author(s): M. Bernardi, B. Maillet, L. Petrella)

A. Rahbek — Testing volatility: consistency of bootstrap testing for a parameter on the boundary of the parameter space.

E. Ruiz — Asymmetric Stochastic Volatility Models: Properties and Estimation. (Co-author(s): V. Czellar, X. Mao, H. Veiga)

(SPE-08) Advances in ordinal contingency table analysis

L. D'Ambra — Dimensionality reduction methods for contingency tables with ordinal variables. (Co-author(s): P. Amenta, A. D'Ambra)

R. Lombardo — Modelling Trends in Ordered Three-Way Non-Symmetrical Correspondence Analysis. (Co-author(s): P. Kroonenberg, E. Beh)

M. Riani — Using Collapsing and Multiple Comparisons to Detect Association in Two Way Contingency Tables. (Co-author(s): S. Arsenis)

(SPE-09) Statistical models for directional and circular data

C. Ley — The WeiSSVM: a tractable, parsimonious and flexible model for cylindrical data.

G. Mastrantonio — The multivariate projected-skew normal distribution: Bayesian estimation and a hidden Markov model application.

A. Panzera — Circular density estimation via matching local trigonometric moments. (Co-author(s): M. Di Marzio, S. Fensore, C. C. Taylor)

(SPE-10) The interplay between frequentist and bayesian inference

C. Grazian — Classical inference for intractable likelihoods.

J. Hannig — Fusion learning for Interlaboratory Comparison. (Co-author(s): Q. Feng, H. Iyer, C. Wang, X. Liu)

F. Pauli — p-value in science: a review of issues and proposed solutions.

(SPE-11) Société Française de Statistique

B.H. Avner — Stochastic Block Model for Multiplex network: an application to a multilevel network of researchers..

Y. Bennani — Nonnegative Matrix Factorization for Transfer Learning. (Co-author(s): I. Redko)

T. Laloe — Detection of dependence patterns with delay.

J. Poggi — Disaggregated Electricity Forecasting using Wavelet-Based Clustering of Individual Consumers. (Co-author(s): J. Cugliari, Y. Goude)

(SPE-12) National accounts

A. Coli — The European Welfare State in times of crisis according to macroeconomic official statistics. (Co-author(s): E. Micheletti, B. Pacini)

C. Martelli — National Account and Open Data: a new semantic approach.

G. Oneto — New information contents of the National Accounts for the monitoring of the economic situation.

(SPE-13) Statistical tools for monitoring the educational system and assessing students' performances

L. Grilli — Evaluation of university students' performance through a multidimensional finite mixture IRT model. (Co-author(s): S. Bacci, F. Bartolucci, C. Rampichini)

G. Leckie — Monitoring school performance using value-added and value-table models: Lessons from the UK.

P. Sarnacchiaro — A statistical model to assess teacher performance. (Co-author(s): I. Camminatiello, R. Palma)

(SPE-14) Robust inference by bounded estimating functions

A.C. Monti — M Estimation based Inference for Ordinal Response Model.

E. Ruli — Approximate Robust Bayesian Inference with an Application to Linear Mixed Models. (Co-author(s): N. Sartori, L. Ventura)

J. Valeinis — Some robust methods using empirical likelihood for two samples. (Co-author(s): M. Velina, E. Cers, G. Luta)

SOLICITED SESSION (SOL)

(SOL-01) Subjective wellbeing and demographic events over the life course

G. Fuochi — Cultural and institutional drivers of basic psychological needs satisfaction. (Co-author(s): P. Conzo, A. Aassve, L. Mencarini)

L. Mencarini — Five reasons to be happy about childbearing. (Co-author(s): A. Aassve, F. Luppi)

B. Nowok — Migration motivations and migrants' satisfaction in the life course: A sequence analysis of geographical mobility trajectories in the United Kingdom.

A. Pirralha — Does becoming a parent change the meaning of happiness and life satisfaction? Evidence from the European Social Survey. (Co-author(s): H. Dobewall)

(SOL-02) Statistics for equitable and sustainable development

E. di Bella — Wellbeing and sustainable development: a multi-indicator approach to evaluate urban waste management systems. (Co-author(s): B. Cavalletti, M. Corsi)

C. Giusti — Small Area Estimation for Local Welfare Indicators in Italy. (Co-author(s): S. Marchetti, L. Faustini, L. Porciani)

T. Laureti — Does socio-economic variables influence the Italians' adherence towards a sustainable diet?. (Co-author(s): L. Secondi)

F. Riccardini — Sustainability of wellbeing: an analysis of resilience and vulnerability through subjective indicators. (Co-author(s): M. Bachelet, F. Maggino)

(SOL-03) New approaches to treat undercoverage and nonresponse

F. Andreis — Methodological perspectives for surveying rare and clustered population: towards a sequentially adaptive approach.

E. Furfaro — Dealing with under-coverage bias via Dual/Multiple Frame designs: a simulation study for telephone surveys.

D. Haziza — Weight adjustment procedures for the treatment of unit nonresponse in surveys. (Co-author(s): É. Lesage)

E. Kabzinska — Empirical likelihood multiplicity adjusted estimator for multiple frame surveys. (Co-author(s): Y. G. Berger)

(SOL-04) Statistical models and methods for network data

M. Cugmas — Measuring stability of co-authorship structures in time. (Co-author(s): A. Ferligoj)

J. Koskinen — A dynamic discrete-choice model for movement flows. (Co-author(s): T. Mueller, T. Grund)

G. Ragozini — Prototyping and Comparing Networks through Archetypal Analysis. (Co-author(s): D. De Stefano, M.R. D'Esposito)

S. Zaccarin — Modeling network dynamics: evidence from policy-driven innovation networks. (Co-author(s): A. Caloffi, D. De Stefano, F. Rossi, M. Russo)

(SOL-05) Recent developments in computational statistics

R. Argiento — A conditional algorithm for Bayesian finite mixture models via normalized point process.

S. Favaro — Thompson sampling for species discovery. (Co-author(s): M. Battiston, Y. Teh)

A. Mira — An application of Reinforced Urn Process to advice network data. (Co-author(s): S. Peluso, P. Muliere, F. Pallotti, A. Loni)

N. Sartori — Bootstrap prepivoting in the presence of many nuisance parameters. (Co-author(s): R. Bellio, I. Kosmidis, A. Salvan)

(SOL-06) Statisticians meet naturalists: issues on ecological and environmental statistics

F. Ferretti — Estimating the abundance of wildlife ungulate populations in Mediterranean areas: methods, problems and findings. (Co-author(s): A. Sforzi)

M. Ferretti — The monitoring of forests in Europe: methods, problems and proposals.

D. Rocchini — The power of generalized entropy for biodiversity assessment by remote sensing: an open source approach. (Co-author(s): L. Delucchi, G. Bacaro)

(SOL-07) From survey data to new data sources and big data in official statistics

G. Barcaroli — Machine learning and statistical inference: the case of Istat survey on ICT. (Co-author(s): G. Bianchi, R. Bruni, A. Nurra, S. Salamone, M. Scarnò)

S. Falorsi — Forecasting Italian Youth Unemployment Rate Using Online Search Data. (Co-author(s): S. Loriga, A. Naccarato, A. Pierini)

B. Liseo — Bayesian nonparametric methods for record linkage. (Co-author(s): A. Tancredi)

T. Tuoto Exploring solutions for linking Big Data in Official Statistics. (Co-author(s): L. Di Consiglio, D. Fusco)

(SOL-08) Symbolic data analysis methods and applications

E. Diday Explanatory and discriminatory power of variables in Symbolic Data Analysis.

M.B. Ferraro Fuzzy and possibilistic approach to clustering of imprecise data. (Co-author(s): P. Giordani)

L. Grassini Symbolic data analysis approach for monitoring the stability of monuments.. (Co-author(s): B. Bertaccini, G. Biagi, A. Giusti)

M. Ichino Similarity and Dissimilarity Measures for Mixed Feature-type Symbolic Data. (Co-author(s): K. Umbleja)

(SOL-09) Compositional analysis

L. Crosato Forecasting CPI weights through compositional VARIMA: an application to Italian data.. (Co-author(s): F. Lovisolo, B. Zavanella)

J. A. Martín-Fernández Understanding association rules from a compositional data approach. (Co-author(s): M. Vives-Mestres, R. Kenett)

A. Menafoglio Object Oriented Geostatistical Simulation of Functional Compositions via Dimensionality Reduction in Bayes spaces. (Co-author(s): A. Guadagnini, P. Secchi)

V. Simonacci Fitting CANDECOMP-PARAFAC model for compositional data: a combined SWATLD-ALS algorithm. (Co-author(s): M. Di Palma, V. Todorov)

(SOL-10) Sustainable development: theory, measures and applications

F. Riccardini Measuring sustainable development goals from now to 2030.

F. Riccardini How the nexus of food/water/energy can be seen with the perspective on well-being of people and the Italian BES framework. (Co-author(s): D. De Rosa)

T. Rondinella An innovative methodology for the analysis of sustainability, inclusion and smartness of growth through Europe2020 indicators.. (Co-author(s): E. Grimaccia)

P. Ungaro The Italian population behaviours toward environmental sustainability: a study from Istat surveys. (Co-author(s): I. Mingo, V. Talucci)

(SOL-11) Detecting heterogeneity in ordinal data surveys

E. Di Nardo CUB models: a preliminary Fuzzy approach to heterogeneity. (Co-author(s): R. Simone)

S. Giordano Modelling uncertainty in bivariate models for ordinal responses. (Co-author(s): R. Colombi, A. Gottard, M. Iannario)

M. Manisera — Treatment of "don't know" responses in rating data: effects on the heterogeneity of the CUB distribution. (Co-author(s): P. Zuccolotto)

F. Pennoni — Modelling a multivariate hidden Markov process on survey data.

(SOL-12) Active ageing: age management and lifelong learning strategies

P. E. Cardone — Age management in Italian companies. Findings of two Isfol surveys. (Co-author(s): M. Aversa, L. D'Agostino)

A. Lorenti — Working after Retirement in Europe.

C. Polli — Older low-skilled workers and economic crisis in Italy. (Co-author(s): R. Angotti)

G. Rivellini — Population ageing and human resources management. A chance for Applied Demography. (Co-author(s): F. Marcaletti, F. Racioppi)

(SOL-13) Statistical models for evaluating policy impact

M. Bia — Evaluation of Training Programs by exploiting secondary outcomes in Principal Stratification frameworks: the case of Luxembourg. (Co-author(s): F. Li, A. Mercatanti)

G. Cerulli — Testing Stability of Regression Discontinuity Models. (Co-author(s): Y. Dongz, A. Lewbel, A. Poulsen)

R. P. Mamede — Counterfactual Impact Evaluation of Vocational Education in Portugal. (Co-author(s): D. Cruz, T. Fernandes)

G. Pellegrini — Italian public guarantees to SME: the impact on regional growth. (Co-author(s): M. De Castris)

(SOL-14) Usage of geocoded micro data in the economic analysis

M. Dickson — Spatial sampling methods with locational errors. (Co-author(s): D. Filipponi)

D. Giuliani — Spatial Micro-Econometrics Models with Locational Errors. (Co-author(s): S. Cozzi, G. Espa)

F. Santi — Three-Year Survival Probability of Italian Start-up Businesses in Healthcare Industry: an Empirical Investigation through Logistic Multilevel Modelling. (Co-author(s): M. M. Dickson, D. Giuliani, D. Piacentino)

(SOL-15) Statistical models in functional data analysis

G. Adelfio — Space-time FPCA Algorithm for clustering ofmultidimensional curves. (Co-author(s): F. Di Salvo, M. Chiodi)

C. Miller — Functional data analysis approaches for satellite remote sensing applications. (Co-author(s): R. O'Donnell, M. Gong, M. Scott)

E. Romano — Order statistics for spatially dependent functional data. (Co-author(s): A. Balzanella, R. Verde)

L. M. Sangalli — A penalized regression model for functional data with spatial dependence. (Co-author(s): M. S. Bernardi, G. Mazza, J. O. Ramsay)

(SOL-16) Forecasting economic and financial time series

G. Goracci — Asymptotics and power of entropy based tests of dependence for categorical data. (Co-author(s): S. Giannerini)

M. M. Pelagatti — Forecasting electricity load and price: a comparison of different approaches. (Co-author(s): F. Lisi)

G. Storti — Flexible Realized GARCH Models. (Co-author(s): R. Gerlach)

(SOL-17) Immigrations and integration in Italy

O. Casacchia — Minorities internal migration in Italy: an analysis based on gravity models. (Co-author(s): C. Reynaud, S. Strozza, E. Tucci)

C. Conti — Growing generations and new models of integration.

N. Tedesco — Measurement of segregation in the labour market. An alternative approach. (Co-author(s): L. Salaris)

L. Terzera — Family behaviours among first generation migrants. (Co-author(s): E. Barbiano di Belgiojoso)

(SOL-18) Open data, linked data and big data in public administration and official statistics

G. Di Bella — Linked Administrative Data in Official Statistics: a Positive Feedback for the Quality?. (Co-author(s): G. Garofalo)

C. Martelli — Generating high quality administrative data: new technologies in a national statistical reuse perspective. (Co-author(s): M. Calzaroni, A. Samaritani)

V. Santarcangelo — An innovative approach about the analysis of quality and efficiency in Italian law. (Co-author(s): A. Buondonno, A. Romano, M. Giacalone, C. Cusatelli)

B. Squittieri — Prato municipality experience towards a high integration between administrative and statistical data.

(SOL-19) Evaluation of prognostic biomarkers

F. Ambrogi — Combining Clinical and Omics data: hope or illusion?. (Co-author(s): P. Boracchi)

L. Antolini — Graphical representations and summary indicators to assess the performance of risk predictors. (Co-author(s): D. Bernasconi)

P. Chiodini — Multivariable prognostic model: external validation and model recalibration with application to non-metastatic renal cell carcinoma. (Co-author(s): L. Cindolo)

(SOL-20) Models for studying the mobility of students

S. Balia — Modelling inter-regional patient mobility: evidence from the Italian NHS. (Co-author(s): R. Brau, E. Marrocu)

A. D'Agostino — University mobility at enrollment: geographical disparities in Italy. (Co-author(s): G. Ghellini, S. Longobardi)

M. Enea — From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree.

F. Giambona — Measuring territory student-attractiveness in Italy. Longitudinal evidence.

CONTRIBUTED SESSION (CON)

(CON-01) Bayesian statistics (1)

F. Giummolè — Reference priors based on composite likelihoods. (Co-author(s): V. Mameli, L. Ventura)

B. Nipoti — On Bayesian nonparametric inference for discovery probabilites. (Co-author(s): J. Arbel, S. Favaro, Y. W. Teh)

R. Pappadà — Relabelling in Bayesian mixture models by pivotal units. (Co-author(s): L. Egidi, F. Pauli, N. Torelli)

C. Scricciolo — On Deconvolution of Dirichlet-Laplace Mixtures.

(CON-02) Statistical modeling

P. Faroughi — A New Bivariate Regression Model for Count Data with Excess Zeros. (Co-author(s): N. Ismail)

B. Francis — Dynamic latent class profiles in cross-sectional surveys: some preliminary results. (Co-author(s): V. Hoti)

P. M. Kroonenberg — The use of deviance plots for non-nested model selection in loglinear models, structural equations, three-mode analysis.

A. Lucadamo — Variable selection through Multinomial LASSO for PCMR. (Co-author(s): L. Greco)

O. Paccagnella — Integrating CUB Models and Vignette Approaches. (Co-author(s): S. Pavan, M. Iannario)

(CON-03) Demographics and social statistics (1)

D. Bellani — Gender egalitarianism, education and life-long singlehood: A multilevel analysis. (Co-author(s): G. Esping-Andersen, L. Nedoluzhko)

L. Colangelo — Fear of Crime and Victimization among Sexual Harassed Women: Evidence from Italy. (Co-author(s): P. Mancini)

S. De Cantis — A survival approach for the analysis of cruise passengers' behavior at the destination. (Co-author(s): M. Ferrante, A. Parroco, N. Shoval)

A. Di Pino — Retirement of the Male Partner and the Housework Division in the Italian Couples: Estimation of the Causal Effects. (Co-author(s): M. Campolo)

F. Lariccia — Many women start, but few continue: determinants of breastfeeding in Italy. (Co-author(s): A. Pinnelli)

(CON-04) Environmental statistics

F. Bono — Measuring sustainable economic development through a multidimensional Gini index. (Co-author(s): M. Giacomarra, R. Giaimo)

C. Calculli — Modeling multi-site individual corals growth. (Co-author(s): B. Cafarelli, D. Cocchi, E. Pignotti)

F. Di Salvo — GAMs and functional kriging for air quality data. (Co-author(s): A. Plaia, M. Ruggieri)

F. Durante — The Kendall distribution and multivariate risks.

(CON-05) Health statistics

E. di Bella — Dental care systems across Europe: the case of Switzerland. (Co-author(s): L. Leporatti, I. Krejci, S. Ardu)

F. Gasperoni — Multi-state models for hospitalizations of heart failure patients in Trieste. (Co-author(s): F. Ieva, G. Barbati)

F. Grossetti — Multi-state Approach to Administrative Data on Patients affected by Chronic Heart Failure. (Co-author(s): F. Ieva, S. Scalvini, A. M. Paganoni)

G. Montanari — Evaluation of health care services through a latent Markov model with covariates. (Co-author(s): S. Pandolfi)

(CON-06) Labor market statistics

A. Bianchi — Multifactor Partitioning: an analysis of employment and firm size. (Co-author(s): S. Biffignandi)

G. Busetta — Ugly Betty looks for a job. Will she ever find it in Italy?. (Co-author(s): F. Fiorillo)

G. Busetta — No country for foreigners: an analysis of hiring process in Italian labor market. (Co-author(s): M. Campolo, D. Panarello)

F. Crippa — Know your audience. Towards a partnership between employers and university. (Co-author(s): M. Zenga)

I. Vannini — Online Job Vacancies: a big data analysis. (Co-author(s): D . Rotolone, C. Di Stefano, A. P. Paliotta, D. F. Iezzi)

(CON-07) Robust statistics

F. Greselin — Robust estimation of mixtures of skew-normal distributions. (Co-author(s): L. García-Escudero, A. Mayo-Iscar, G. McLachlan)

M. Musio — Renyi's Scoring Rules. (Co-author(s): A. F. Dawid)

A. Paganoni — Robust classification of multivariate functional data. (Co-author(s): F. Ieva)

G. C. Porzio — A robust estimator for the mean direction of the von Mises-Fisher distribution. (Co-author(s): T. Kirschstein, S. Liebscher, G. Pandolfo, G. Ragozini)

F. Palumbo — Robust Partial Possibilistic Regression Path Modeling. (Co-author(s): R. Romano)

(CON-08) Sampling methods

A. Ghiglietti — Adaptive Randomly Reinforced Urn design and its asymptotic properties.

D. Marella — PC algorithm from complex sample data. (Co-author(s): P. Vicard)

S. Missiroli — Optimal Adaptive Group Sequential Procedure for Finite Populations in the Presence of a Cost Function. (Co-author(s): E. Carfagna)

E. Pelle — The Rao regression-type estimator in ranked set sampling. (Co-author(s): P. Perri)

M. Ruggiero — Modelling stationary varying-size populations via Polya sampling. (Co-author(s): P. De Blasi, S. Walker)

(CON-09) Economic data analysis

M. Brunetti — Getting older and riskier: the effect of Medicare on household portfolio choices. (Co-author(s): M. Angrisani, V. Atella)

E. Ciavolino — Modelling the Public Opinion on the European Economy with the HO-MIMIC Model. (Co-author(s): M. Carpita)

G. D'Epifanio — Indexing the Worthiness of Social Agents. To norm index on conventional specifications.

G. Guagnano — An econometric model for undeclared work. (Co-author(s): M. Arezzo)

M. Mussini — A spatial shift-share decomposition of energy consumption variation. (Co-author(s): L. Grossi)

(CON-10) Quantile methods

M. Bernardi — Bayesian inference for L_p–quantile regression models. (Co-author(s): V. Bignozzi, L. Petrella)

V. Bignozzi — On the L_p-quantiles and the Student t distribution. (Co-author(s): M. Bernardi, L. Petrella)

M. Marino — M-quantile regression for multivariate longitudinal data. (Co-author(s): M. Alfò, M. Ranalli, N. Salvati)

D. Vistocco Comparing Prediction Intervals in Quantile and OLS Regression. (Co-author(s): C. Davino)

(CON-11) Statistical algorithms

N. Loperfido An Algorithm for Finding Projections with Extreme Kurtosis. (Co-author(s): C. Franceschini)

L. Scrucca Poisson change-point models estimated by Genetic Algorithms.

A. Stamm Maximum Likelihood Estimators of Brain White Matter Microstructure. (Co-author(s): O. Commowick, S. Vantini, S. K. Warfield)

(CON-12) Statistics for medicine

G. Barbati Competing risks between mortality and heart failure hospital re-admissions: a community-based investigation from the Trieste area. (Co-author(s): F. Ieva, A. Scagnetto, G. Sinagra, A. Di Lenarda)

C. Brombin Evaluating association between emotion recognition and Heart Rate Variability indices. (Co-author(s): F. Cugnata, R. M. Martoni, M. Ferrario, C. Di Serio)

M. Ferrante Socio-economic deprivation, territorial inequalities and mortality for cardiovascular diseases in Sicily. (Co-author(s): A. Milito, A. Parroco)

M. Giacalone The use of Permutation Tests on Large-Sized Datasets. (Co-author(s): A. Alibrandi, A. Zirilli)

(CON-13) Statistics for the education system

G. Boscaino Further considerations on a new indicator for higher education student performance. (Co-author(s): G. Adelfio, V. Capursi)

C. Masci Analysis of pupils' INVALSI achievements by means of bivariate multilevel models. (Co-author(s): A. Paganoni, F. Ieva, T. Agasisti)

A. Valentini Promoting statistical literacy to university students: a new approach adopted by Istat. (Co-author(s): G. De Candia, M . Carbonara)

(CON-14) Testing procedures

E. Cascini A Reliability Problem: Censored Tests.

G. De Santis Testing the Gamma-Gompertz-Makeham model. (Co-author(s): G. Salinari)

M. M. Pelagatti A nonparametric test of independence.

A. Pini Functional Data Analysis of Tongue Profiles. (Co-author(s): L. Spreafico, S. Vantini, A. Vietti)

A. Vagheggini On the asymptotic power of the statistical test under Response-Adaptive randomization. (Co-author(s): A. Baldi Antognini, M. Zagoraiou)

(CON-15) Time series analysis

C. Cappelli — Robust Atheoretical Regression Tree to detect structural breaks in financial time series. (Co-author(s): P. D'Urso, F. Di Iorio)

P. Chirico — Prediction intervals for heteroscedastic series by Holt-Winters methods.

M. Costa — Inequality decomposition for financial variables evaluation.

G. De Luca — Three-stage estimation for a copula-based VAR model. (Co-author(s): G. Rivieccio)

(CON-16) Forecasting methods

M. Andreano — Forecasting with Mixed Data Sampling Models (MIDAS) and Google trends data: the case of car sales in Italy. (Co-author(s): R. Benedetti, P. Postiglione)

V. Candila — Probability forecasts in the market of tennis betting: the CaSco normalization. (Co-author(s): A. Scognamillo)

S. Vantini — Daily Prediction of Demand and Supply Curves. (Co-author(s): A. Canale)

(CON-17) Bayesian statistics (2)

G. Marchese — Bayesian hierarchical models for analyzing and forecasting football results. (Co-author(s): P. Brutti, S. Gubbiotti)

L. Paci — Bayesian modeling of spatio-temporal point patterns in residential property sales. (Co-author(s): A. E. Gelfand, M. Beamonte, P. Gargallo, M. Salvador)

V. Vitale — Non-parametric Bayesian Networks for Managing an Energy Market. (Co-author(s): V. Guizzi, F. Musella, P. Vicard)

(CON-18) Business statistics

E. Bartoloni — How do firms perceive their competitiveness? Measurement and determinants.

C. Bocci — An evaluation of export promotion programmes with repeated multiple treatments. (Co-author(s): M. Mariani)

A. Righi — The inter-enterprise relations in Italy. (Co-author(s): A. Nuccitelli, G. Barbieri)

(CON-19) Clustering and classification

C. Drago — Dendrograms Stability Analysis of Sub-periods Time Series Clustering. (Co-author(s): R. Ricciuti)

G. Menardi — Stability-based model selection in nonparametric clustering.

T. Padellini — Topological signatures for classification. (Co-author(s): P. Brutti)

(CON-20) Demographics and social statistics (2)

M. Antonicelli — Ecolabels: informin or confusing customers? Evidences form the agrifood sector. (Co-author(s): D. Calace, D. Morrone, A. Russo, V. Vastola)

B. Arpino — What makes you feeling old? An analysis of the factors influencing perceptions of ageing. (Co-author(s): V. Bordone, A. Rosina)

G. De Santis — A (partial) solution to the intractability of APC models. (Co-author(s): M. Mucciardi)

G. Gabrielli — Partner reunification of first generation immigrants in Lombardy. (Co-author(s): A. Paterno, L. Terzera)

(CON-21) Statistical inference

E. Kenne Pagui — Median bias reduction of maximum likelihood estimates in binary regression models. (Co-author(s): A. Salvan, N. Sartori)

N. Lunardon — On penalized likelihood and bias reduction. (Co-author(s): G. Adimari)

A. Maruotti — Population size estimation and heterogeneity in capture-recapture count data. (Co-author(s): O. Anan, D. Böhning)

(CON-22) Survey methods

A. Pinto — Italian consumers' food risks perception: an approach based on the correspondence analysis. (Co-author(s): E. Ruli, S. Crovato, L. Ventura, L. Ravarotto)

R. Salvatore — Spatial-temporal multivariate small area estimation. (Co-author(s): F. Cappuccio)

D. Toninelli — Is the Smartphone Participation Affecting the Web Survey Experience?. (Co-author(s): M. Revilla)

POSTER SESSION (POS)

M. Bernardi — Non-conjugate Variational Bayes Approximation. (Co-author(s): E. Ruli)

M. Bernardi — The Multivariate Fuzzy Skew Student–t distribution.

M. Bini — Quality of Educational Services, Institutional Image, Students' Satisfaction and Loyalty in Higher Education. (Co-author(s): L. Masserini, M. Pratesi)

L. Bisaglia — Estimation of INAR(p) models using bootstrap. (Co-author(s): M. Gerolimetto)

D. Bossoli — Effect of internet-based cognitive therapy on children anxiety disorders: results from a marginal logistic quantile regression.

C. Calì — Some mathematical properties of the ROC curve. (Co-author(s): M. Longobardi)

M. Cannas — Machine learning for the estimation of the propensity score: a simulation study. (Co-author(s): B. Arpino)

M. Cannas — An R package for propensity score matching with clustered data. (Co-author(s): B. Arpino, C. Conversano)

A. Coli — Mapping local social protection data in Italy. (Co-author(s): B. Pacini, A. Valentini, S. Venturi)

A. Cosma — Indirect inference for nonlinear panel data.

I. L. Danesi — Cluster Analysis of Transactional Data in the Frequency Domain. (Co-author(s): F. M. Pons, C. Rea)

L. Gabrielli — Using purchase market behavior to estimate collective well-being. (Co-author(s): G. Riccardi, L. Pappalardo)

F. Giambona — The Bifactor Item Response Theory Model for the analysis of repeated measurements. An application to the measurement of Italian households' well-being. (Co-author(s): M. Porcu, I. Sulis)

A. Lepore — A Bayesian short-term strategy for site specific wind potential assessment. (Co-author(s): P. Erto, B. Palumbo, M. Lepore)

A. Magrini — Distributed-Lag Structural Equation Modelling: An Application to Impact Assessment of Research Activity on European Agriculture. (Co-author(s): F. Bartolini, A. Coli, B. Pacini)

G. Mastrantonio — A multivariate circular-linear hidden Markov model for site-specific assessment of wind predictions by an atmospheric simulation system. (Co-author(s): A. Pollice, F. Fedele)

F. Musella — Bayesian networks for supporting the digitization process in Italian schools. (Co-author(s): S. Capogna, M.C. De Angelis)

B. Palumbo — Statistical approach in aerospace industry innovation. (Co-author(s): P. Erto, F. Tagliaferri, G. De Chiara, R. Marrone, C. Leone, S. Genna)

B. Palumbo — Ship fuel consumption control and engineering approach to fault-detection. (Co-author(s): P. Erto, A. Lepore, L. Vitiello, C. Capezza, D. Bocchetti, A. D'Ambra, B. Antonelli)

A. Petrucci — Small area model-based direct estimator for spatial data. (Co-author(s): C. Bocci, E. Rocco)

F. Poggioni — Dynamic Quantile Lasso Regression. (Co-author(s): L. Petrella, M. Bernardi)

A. Pramov — Estimating dependence within neuropsychological models for designing risk profiles of decision-makers..

A. Pramov — Confidence intervals for a partially identified parameter with bounds estimated by the minimum and the maximum of two correlated and normally distributed statistics.

I. Primerano — Semantic Knowledge Detection in Open-ended Questionnaire. (Co-author(s): G. Giordano)

G. Ragozini — A joint approach to the analysis of time-varying affiliation networks. (Co-author(s): D. D'Ambrosio, M. Serino)

M. Restaino — Non-parametric estimators for estimating bivariate survival function under randomly censored and truncated data. (Co-author(s): H. Dai, H. Wang)

G. Riccardi — Bayesian M-quantile regression in Small Area Estimation.

E. Ruli — Optimal B-robust posterior distributions for operational risk. (Co-author(s): I. Danesi, F. Piacenza, L. Ventura)

F. Schirripa Spagnolo — Estimating income of immigrant communities in Italy using small area estimation methods. (Co-author(s): N. Salvati, A. D'Agostino)

M. Soscia — The Switching Skew-GARCH Model. (Co-author(s): M. Bernardi, L. Petrella)

S. Spina — Inference on a non-homogeneous Gompertz process with jumps as model of tumor dynamics. (Co-author(s): V. Giorno, P. Román-Román, F. Torres-Ruiz)

G. Storti — Combining multiple frequencies in multivariate volatility forecasting. (Co-author(s): A. Amendola, V. Candila)

D. Toninelli — An Enhanced Measure of Well-being through Structural Equation Modeling: a Cross-Country Approach. (Co-author(s): M. Cameletti)

A. Vanacore — Statistics for knowledge improvement of an innovative manufacturing process and quality cost management. (Co-author(s): B. Palumbo, F. Del Re, P. Corrado, M. Lanza, G. La Sala, M. Mastrovita)

A. Vanacore — Statistics for Safety and Ergonomics in Design. (Co-author(s): A. Lanzotti, C. Percuoco, A. Capasso, F. Liccardo, B. Vitolo)

L. Zanin — Modelling transition probabilities in a flexible hierarchical logit framework: evidence from the Italian labour market. (Co-author(s): R. Calabrese)

Robust estimation of mixtures of skew-normal distributions

Stima robusta di misture di normali asimmetriche

L.A. García-Escudero, F. Greselin, A. Mayo-Iscar, and G. McLachlan

Abstract Recently, observed departures from the classical Gaussian mixture model in real datasets motivated the introduction of mixtures of skew t, and remarkably widened the application of model based clustering and classification to great many real datasets. Unfortunately, when data contamination occurs, classical inference for these models could be severely affected. In this paper we introduce robust estimation of mixtures of skew normal, to resist sparse outliers and even pointwise contamination that may arise in data collection. Hence, in each component, the skewed nature of the data is explicitly modeled, while any departure from it is dealt by the robust approach. Some applications on real data show the effectiveness of the proposal.

Sommario *Recentemente, a fronte di dataset reali multimodali con asimmetria e code pesanti, è stato introdotto il modello mistura di t asimmetriche, ampliando considerevolmente il campo di applicazione delle classiche misture di distribuzioni Gaussiane. La stima di questi modelli non è però robusta rispetto a contaminazioni e/o errori che possano accadere nella raccolta dei dati. In questo lavoro si introduce uno stimatore robusto per le misture di normali asimmetriche, in grado di resistere a valori anomali e a contaminazione puntuale. La natura asimmetrica dei dati è esplicitamente modellata in ciascuna componente, mentre la stima robusta consente*

L.A. García-Escudero
Department of Statistics and Operations Research and IMUVA, University of Valladolid, Valladolid, Spain, e-mail: lagarcia@eio.uva.es

F. Greselin
Department of Statistics and Quantitative Methods, Milano-Bicocca University, Milano, Italy e-mail: francesca.greselin@unimib.it

A. Mayo-Iscar
Department of Statistics and Operations Research and IMUVA, University of Valladolid, Valladolid, Spain e-mail: agustinm@eio.uva.es

G. McLachlan
Department of Mathematics, University of Queensland, Brisbane, Australia e-mail: g.mclachlan@uq.edu.au

di gestire ogni allontanamento dal modello. Applicazioni su dati reali documentano l'efficacia della proposta.

Key words: Clustering, Robustness, Trimming, Constrained estimation, Skew data, model-based classification, Finite mixture models.

1 Introduction

Finite mixtures of distributions have been widely used as a powerful tool to model heterogeneous data and to approximate complex probability densities, presenting multimodality, skewness and heavy tails. During the last decade, there has been an increasing interest in finding more flexible methods to accurately represent observed data and to reduce unrealistic assumptions. This very active and stimulating context has seen the appearance of many contributions. Among the available proposals in the literature, mixtures of skew normal can incorporate asymmetry in components (see f.i., [2]). On the other hand, mixtures of t distributions can model heavier tails by down-weighting the contribution of extremal observations, as shown in [4, 7]. Mixtures of skew t may accomodate for both asymmetry and leptokurtosis in the grouped data, and therefore remarkably widened the application of model based clustering and classification (see, for example, [6]).

When dealing with model fitting, the elegant theory of likelihood inference provides estimators with desirable properties such as consistency and efficiency. However, these estimators are not robust and there is usually a trade-off between robustness and efficiency. Hence, due to the possible presence of contaminating data (background noise, pointwise contamination, unexpected minority patterns, etc.) a small fraction of outliers, (located far from the groups and, even, between them) could severely affect the model fitting, and a robust approach is needed. Surely, considering skew t distributions is an interesting proposal to achieve robustness with respect to uniform noise or a few sparse outliers. However, Hennig (2004) noted that they are not effective against gross outliers or pointwise contamination that may arise in data collection, their asymptotic breakdown point being zero.

In view of all these considerations, we introduce here a new proposal. To gain effective protection against all type of outliers we jointly use trimming and constrained estimation along the estimation of mixtures of skew Gaussian distributions. We apply our robust estimation to skew Gaussian components (instead of skew t) because they are more parsimonious in parameters and easier in estimation. Indeed the flexibility inherited by trimming does not require any assumption on the heaviness of the tails. The asymptotic breakdown point of the resulting method is strictly positive, an indication of robustness even against gross outliers. As final remark, due to its properties, our methods is offered as a very general tool for clustering heterogeneous skew populations.

2 Finite Mixtures of Canonical Fundamental Skew Normal

We consider here the location-scale variant of the Canonical Fundamental Skew Normal (CFUSN)[1], whose parameters allow to separately govern location, scale, correlation, and skewness. The model arises from a $p+q$ multivariate normal r.v. $(\mathbf{U},\mathbf{V})$, such that

$$\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} \sim \mathscr{N}_{q+p}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{I}_q & 0\\ 0 & \Sigma\end{bmatrix}\right)$$

where Σ is a positive definite scale matrix and 0 is a vector of zeros with appropriate dimension. Then, given a $p\times q$ matrix Δ and a p-dimensional vector μ, we arrive at a stochastic representation for $\mathbf{Y}$, obtained via a convolution, i.e.

$$\mathbf{Y} = \mu + \Delta|\mathbf{U}| + \mathbf{V},$$

which follows the CFUSN distribution, whose density is given by

$$f(\mathbf{y};\mu,\Sigma,\Delta) = 2^q \phi_p(\mathbf{y};\mu,\Omega)\Phi_q\big(\Delta^T\Omega^{-1}(\mathbf{y}-\mu);0,\Lambda\big), \tag{1}$$

where $\Omega = \Sigma + \Delta\Delta^T$ and $\Lambda = \mathbf{I}_q - \Delta^T\Omega^{-1}\Delta$. As usual, $\phi_p(\mathbf{y};\mu,\Sigma)$ denotes the p-dimensional density of the multivariate Gaussian with mean μ and scale Σ evaluated at $\mathbf{y}$, while $\Phi_q(\cdot)$ denotes the cumulative distribution function. The probability density function for a g-component mixture model of CFUSNs can be written as

$$\sum_{h=1}^{g}\pi_h f(\mathbf{y};\mu_h,\Sigma_h,\Delta_h), \qquad \pi_h \geq 0, \qquad \sum_{h=1}^{g}\pi_h = 1, \tag{2}$$

where $f(\mathbf{y};\mu_h,\Sigma_h,\Delta_h)$ denotes the h^{th} skew normal component with location parameter μ_h, scale matrix Σ_h and skew parameter Δ_h, given in (1). We denote the unknown parameter by $\theta = (\theta_1,\ldots,\theta_g)$, with $\theta_h = (\pi_h,\mu_h,\Sigma_h,\Delta_h)$ related to component h, π_h being the group weights, and adopt the acronym FM-CFUSN for (2).

3 Robust estimation for FM-CFUSN

Aiming at achieving robustness and obtaining good breakdown properties for the ML estimators, a constructive way to obtain a robust estimation is given by providing a feasible EM algorithm for model fitting, where we incorporate impartial trimming, just before the E-step, and constrained estimation along the M-step. The key idea in *trimming* is that a small portion of observations, which are highly unlikely to occur under the current fitted model, is discarded from contributing to the mixture estimates. In the maximization, therefore, we consider the following *trimmed* log-likelihood function [3, 8]

$$\ell_{trim} = \sum_{j=1}^{n} \zeta(\mathbf{y}_j) \log \left[\sum_{h=1}^{g} \phi_p(\mathbf{y}_j; \mu_h, \Omega_h) \Phi_q(\Delta_h^T \Omega_h^{-1}(\mathbf{y}_j - \mu_h); 0, \Lambda_h) \pi_h \right]. \quad (3)$$

By $\zeta(\cdot)$ we denote a 0-1 trimming indicator function that indicates whether observation $\mathbf{y}_j$ is trimmed off: $\zeta(\mathbf{y}_j)$=0, or not: $\zeta(\mathbf{y}_j)$=1. A fixed fraction α of observations, whose contributions to the likelihood are lower than their α-quantile, will be unassigned by setting $\sum_{j=1}^{n} \zeta(\mathbf{y}_j) = [n(1-\alpha)]$ just before each E-step, in such a way that they do not influence the parameter estimation (by $[\cdot]$ we denote the integer part of the argument). Hence α denotes the *trimming level*.

Furthermore - and this will be our second step - we implement a *constrained ML estimation* for the Σ_h matrices in the components of the mixture. The ML estimates $\hat{\theta}$ based on a set of i.i.d. observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is now rephrased into

$$\hat{\theta} = \arg\max_{\theta} \ell_{trim}(\theta|\mathbf{y}), \quad \text{for } \theta \in \Theta, \quad (4)$$

where Θ denotes the parameter space. Also in this case - the same happens for the non-robust case - the defining problem is ill-posed because the log-likelihood tends to ∞ when either $\mu_h = \mathbf{y}_j$ and $|\Sigma_h| \to 0$. As a trivial consequence, the EM algorithm can be trapped into non-interesting local maximizers, called "spurious" solutions.

For this reason, we set a constraint on the maximization of ℓ_{trim}, by imposing

$$\lambda_{l,h} \leq c\, \lambda_{m,k} \qquad \text{for} \quad 1 \leq l \neq m \leq p \quad \text{and} \quad 1 \leq h \neq k \leq g \quad (5)$$

where $\{\lambda_{l,h}\}_{l=1,\ldots,p}$ are the eigenvalues of Σ_h, for $h = 1, \ldots, g$ and $1 \leq c < +\infty$.

We will denote by Θ_c the constrained parameter space under requirement (5).

4 Applications to real data

We consider here the Australian Institute of Sports (AIS) dataset, consisting of $p = 11$ physical and hematological measurements on 202 athletes (100 females and 102 males) in different sports, and available within the R package *sn*. Our purpose is to provide a model for the entire dataset, and since the group labels (athletes gender) are provided in advance, the aim is to classify athletes by this feature. By applying the robust FM-CFUSN, with $p = 11$, $q = 1$, 50 starting values and stopping the EM after a maximum of 100 iterations, we got the results shown in Figure 1 (left panel). After the robust estimation, also the 20 trimmed observations can be classified, by using the Bayes' rule and assigning each unit to the component with maximum a posteriori probability, yielding finally to 4 misclassified units (Figure 1, right panel). Notice that this is a very encouraging result when compared to similar approaches available in the literature (see also [5], where a detailed analysis has been done), as the use of an ordinary normal mixture model yields 8 misclassifications, and the t-version of the FM-CFUSN model yields 4 misclassifications.

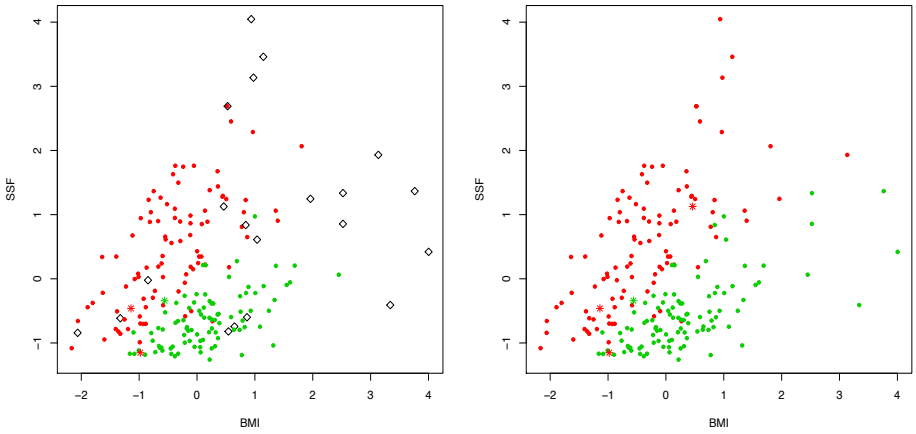

Fig. 1 Classification of the AIS dataset (left panel) by using robust FM-CFUSN (c=16, α=0.1, female data in red, male in green, represented as filled circles when right classification takes place, otherwise as stars; trimmed units are denoted by diamonds). Bivariate plots refer to variables weight/height2 (BMI) and sum of skin folds (SSF)

A second application has been developed on annual financial data of 66 American firms, considering the Ratio of Retained Earnings (RE) to total assets, and ratio of earnings before interest an taxes (EBiT) to total assets. The purpose is to classify firms who filed for bankruptcy. The bivariate sample is plotted in Figure 2, where bimodality and skewness are apparent, thus we fit a two component mixture to the data. We set $\alpha = 0.10$ and c=16. After estimating the model without the contribution of the 7 trimmed units (which are apparently located far from the cores of the components), we classified them as well, arriving at only 4 misclassified firms. This compares to 2 misclassifications with using the t-version of the FM-CFUSN model.

A third application has been done on real world natural images from the Berkeley's image segmentation dataset, where the aim is to segment pixels into background and foreground. We also applied our method to perform automated high-dimensional flow cytometric data analysis on real data. All results show that our method provides an effective approach for asymmetric, heavy tailed data in the mixture components. Simulated results show that the estimation is able to resist to noise as well as to the more dangerous pointwise contamination.

In conclusion, even if further study should be devoted along the lines of the present proposal, we introduced a very general robust tool for clustering heterogeneous skew populations, by using the parsimonious and well-known skew Gaussian model and by flexibly dealing with any departure from the skewed components' cores via the trimming approach.

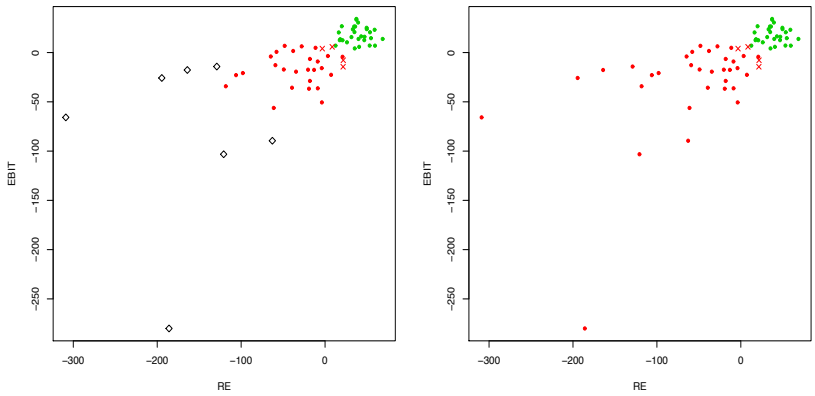

Fig. 2 Classification of the Bankruptcy dataset by using robust FM-CFUSN (bankrupted firms in red, solvent firms in green, represented as filled circles when right classification takes place, otherwise as crosses; In the left panel trimmed units are denoted by diamonds, while in the right panel also trimmed observations have been classified)

References

[1] Arellano-Valle, R. B., Genton, M. G.: On fundamental skew distributions. J. Multiv. Anal., **96** (1), 93–116 (2005).

[2] Branco, M. D., Dey, D. K.: A general class of multivariate skew-elliptical distributions. J. Multiv. Anal., **79** (1), 99-113 (2001).

[3] García-Escudero, L., Gordaliza, A., Mayo-Iscar, A.: A constrained robust proposal for mixture modeling avoiding spurious solutions. Adv. Data Anal. Classif., **8** (1), 27–43 (2014).

[4] Greselin, F., Ingrassia, S.: Constrained monotone EM algorithms for mixtures of multivariate t distributions. Stat. Comp., **20** (1), 9–22 (2010).

[5] Lee, S. X., McLachlan, G. J.: Model-based clustering and classification with non-normal mixture distributions. Statistical Methods & Appl., 22 (4), 427–454 (2013).

[6] Lin, T. I.: Robust mixture modeling using multivariate skew t distribution. Stat. Comp., **20**, 343–356 (2010).

[7] McLachlan, G. J., Peel, D.: Robust cluster analysis via mixtures of multivariate t-distributions. In: Advances in pattern recognition. Springer Berlin Heidelberg (1998).

[8] Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using the trimmed likelihood estimator. Comp. Stat. & Data Anal., **52** (1), 299–308 (2007).