47^{th} Scientific Meeting of the Italian Statistical Society

P R O C E E D I N G S

Editors: S. Cabras, T. Di Battista and W. Racugno

CUEC EDITRICE

ISBN: 978-88-8467-874-4
47^{th} SIS Scientific Meeting of the Italian Statistica Society
Cagliari, June 10-14, 2014.

CUEC Cooperativa Universitaria Editrice Cagliaritana
via Is Mirrionis 1, 09123 Cagliari
Tel. e Fax 070271573
www.cuec.eu
info@cuec.eu

ISBN 978-88-84-67-874-4

Program Committee

The Program Committee of the 47^{th} SIS Scientific Meeting is composed by:

Tonio Di Battista (*Chair*)	Università di Pescara
Marco Alfò	Sapienza Università di Roma
Giulio Barcaroli	ISTAT
Gianni Betti	Università di Siena
Daniela Cocchi	Università di Bologna
Stefano Gattone	Università di Roma "Tor Vergata"
Andrea Giommi	Università di Firenze
Salvatore Ingrassia	Università di Catania
Michele La Rocca	Università di Salerno
Letizia Mencarini	Università di Torino
Sonia Petrone	Università Commerciale Bocconi
Alessandra Petrucci	Università di Firenze
Donato Posa	Università di Lecce
Walter Racugno	Università di Cagliari
Giovanna Ranalli	Università di Perugia
Marco Riani	Università di Parma
Roberto Rocci	Università di Roma "Tor Vergata"
Arjuna Tuzzi	Università di Padova

Preface

The edition of this volume gave us the opportunity to perceive that, together with many well-known Italian statisticians belonging to the national and international community, many young researchers are emerging. They presented, at the *47th Scientific Meeting of the Italian Statistical Society*, their remarkable contributions, both from the methodological and the applicative point of views.

Although some papers may not look, in their actual form, fully mature from the scientific and communicative point of views, we decided - in agreement with the referees - to publish them since promising and full of ideas. In this respect, the contributions published in this volume provide a comprehensive overview of the current Italian scientific researches in theoretical and applied statistics.

This volume also contains several contributions presented by foreign researchers, highlighting the fact that the Italian Statistical Society has an attractive role in the international scientific community.

Finally, we would like to emphasize that, even from the abstracts of the contributions, the wideness of the collaborations between the statisticians and the experts from other fields emerges. This denotes that, also in Italy, statistical methods are spreading in the different fields of the scientific researches.

Stefano Cabras
Tonio Di Battista
Walter Racugno

Cagliari, June 11, 2014

Table of Contents

List of papers

List of abstracts

You can access to the full paper by clicking on their link.

PLENARY SESSIONS

President's Invited Speaker

- Antonio Golini. **Demographic crisis and economic crisis for Italian Mezzogiorno: an iceberg detached from the continent ?**

Plenary Session A

- Adrian Smith FRS. **The Bayesian 21st century. An appreciation of the contributions of Dennis Lindley: 1923-2013**

Plenary Session B

- Chiara Sabatti. **In the mist of the data deluge, how to let the interesting findings surface? Tales from genetics.**

Plenary Session C

- Andrea Cerioli. **How to marry robustness and applied statistics** (Full paper: 3104.pdf)

SPECIALIZED SESSIONS

SP1 - Recent Advances In Biostatistics

- Anne Presanis, Paul Birrell, Daniela De Angelis. **Statistical challenges in epidemic modelling** (Full paper: 3056.pdf)
- Chiara Romualdi. **Recent advances in genomic studies** (Full paper: 3055.pdf)
- Federico Rotolo, Stefan Michelis. **Global interaction tests for predictive gene signatures in randomized clinical trials.** (Full paper: 3054.pdf)

SP2 - Clustering Real Time Data Streams

- Antonio Balzanella, Rosanna Verde, Antonio Irpino. **Monitoring spatially dependent data streams** (Full paper: 3085.pdf)
- Angela De Sanctis, Francesca Fortuna. **Functional classification on convex function spaces** (Full paper: 3141.pdf)
- Donato Malerba, Annalisa Appice. **Discovering trend clusters in sensor data streams** (Full paper: 3140.pdf)

SP3 - Bayesian Nonparametrics: Methods And Applications

- Raffaele Argiento, Ilaria Bianchini, Alessandra Guglielmi. **A Bayesian nonparametric model for density and cluster estimation: the ε-NGG process mixture** (Full paper: 3032.pdf)
- Antonio Lijoi, Bernardo Nipoti, Igor Prunster. **A Bayesian nonparametric model for combining data from different experiments** (Full paper: 3033.pdf)
- Judith Rousseau. **On consistency issues in Bayesian nonparametric testing - a review** (Full paper: 3034.pdf)

SP4 - Recent Advances In Time Series Analysis

- Lara Fontanella, Luigi Ippoliti, Pasquale Valentini. **Regressions in Spatially Dynamic Factor Models** (Full paper: 3082.pdf)
- Liang Peng, Yadong Li, Jinguo Gong, Qiwei Yao. **Estimation of Extreme Quantiles for Functions of Dependent Random Variables** (Full paper: 3080.doc)
- Tommaso Proietti, Alessandra Luati. **Generalised Linear Cepstral Models for the Spectrum of a Time Series** (Full paper: 3081.pdf)

SP5 - New Challenges In Survey Sampling

- Fulvia Mecatti. **Multiple Frame Surveys: a simplified and unified review** (Full paper: 2990.pdf)
- Jean Opsomer, Wade Herndon, F. Jay Breidt. **Testing for Informativeness in Analytic Inference from Complex Surveys** (Full paper: 2998.pdf)
- Vijay Verma. **On sampling elusive populations** (Full paper: 2989.pdf)

SP6 - Directional Data

- Alan Gelfand, Fangpo Wang. **Analyzing spatial and spatio-temporal angular and linear data using Gaussian processes** (Full paper: 2991.pdf)
- Giovanna Jona Lasinio, Gianluca Mastrantonio, Alan Gelfand. **Models for space-time directional data using Wrapped Gaussian processes** (Full paper: 2995.pdf)
- Francesco Lagona. **Regression analysis of correlated circular data** (Full paper: 2870.pdf)

SP7 - Scoring Rules And Pseudo

- A.p. Dawid. **Proper Scoring Rules in Statistical Inference** (Full paper: 3068.pdf)
- Monica Musio, Valentina Mameli, Alexander Dawid. **Comparison of approaches to inference in stationary AR(1) models** (Full paper: 3035.pdf)
- Erlis Ruli, Nicola Sartori, Laura Ventura. **Approximate Bayesian Computation with proper scoring rules** (Full paper: 2973.pdf)

SP8 - Quantile And M

- Giulia Cavrini, Sara Piombo, Alessandra Samoggia. **A Multilevel Quantile Regression Analysis of Chronic Diseases Affects on Physical and Mental Well-being** (Full paper: 3067.docx)
- Marco Geraci, Mc Jones. **Prediction of conditional quantiles on the half line and the unit interval using transformation models** (Full paper: 2825.pdf)
- Maria Francesca Marino, Nikos Tzavidis. **Mixed hidden Markov models for quantiles** (Full paper: 3070.pdf)

SP9 - Methodological Issues For Constructing Composite Indicators

- Filomena Maggino, Marco Fattore. **Evaluating subjective suffering in Italy with partially ordered sets** (Full paper: 3016.pdf)
- Matteo Mazziotta, Adriano Pareto. **A Non-compensatory Composite Index for Comparisons over Time** (Full paper: 3011.pdf)
- Francesco Vidoli, Elisa Fusco, Claudio Mazziotta. **Non-compensability in composite indicators: a robust directional frontier method** (Full paper: 3010.pdf)

SP10 - Parametric And Nonparametric Mixed Effect Models

- Silvia Cagnone, Francesco Bartolucci. **Continuous versus discrete latent structures in dynamic latent variable models** (Full paper: 3065.pdf)
- Paul H.c. Eilers. **Generalized linear mixed models for over-dispersed counts** (Full paper: 3064.pdf)
- Sara Viviani. **Smooth random effect distribution in joint models with an application to cardiomiopathy data.** (Full paper: 3053.pdf)

SP11 - Demography And Environmental Emergency

- Elena Ambrosetti, Enza Petrillo. **Environmental Change, Migration and Displacement. Insights and developments from L'Aquila** (Full paper: 3039.docx)
- Wolfgang Lutz. **Population – Environment Interactions** (Full paper: 3041.pdf)
- Valerio Manno, Giada Minelli, Susanna Conti. **The use of current data to evaluate the health impact of environmental pollution: the "SENTIERI approach" and the case study of Taranto** (Full paper: 3040.docx)

SP12 - Statistical Analysis And Big Data

- Barteld Braaksma. **Big Data and official statistics: local experiences and international initiatives** (Full paper: 3126.pdf)
- Salvatore Rinzivillo, Fosca Giannotti. **Understanding Human Mobility with Big Data** (Full paper: 3027.pdf)
- Simone Vantini, Piercesare Secchi, Paolo Zanini. **The Virtuous Cycle of Big Data and Big Cities: a Case Study from Milan** (Full paper: 3000.pdf)

SOLICITED SESSIONS

SL1 - Bayesian Models For Complex Problems

- Raffaele Argiento, Alessandra Guglielmi. **Bayesian principal curve clustering by species-sampling mixture models** (Full paper: 2877.pdf)
- Laura Azzimonti, Laura Sangalli, Piercesare Secchi. **Modeling prior knowledge on complex phenomena behaviors via partial differential equations** (Full paper: 2923.pdf)

- Michele Guindani, Linlin Zhang, Francesco Versace, Marina Vannucci. **A Bayesian Variable Selection Model for the Clustering of Time Courses in FMRI data** (Full paper: 2879.pdf)
- Andrea Tancredi, Brunero Liseo. **Bayesian Inference with Linked Data** (Full paper: 2878.pdf)

SL2 - Geostatistics And Environmental Applications

- Claudia Cappello, Sandra De Iaco, Donato Posa. **Computing non-separability for space-time covariance functions: a case study on PM10 data** (Full paper: 3138.pdf)
- Sandra De Iaco, Sabrina Maggio, Monica Palma, Donato Posa. **Modeling environmental quality: a case study** (Full paper: 3137.pdf)
- J. Jaime Gòmez-Hernàndez, Teng Teng Xu. **When it is not normal to be Gaussian** (Full paper: 3135.pdf)
- Ana Russo, Amílcar O. Soares. **Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach** (Full paper: 3139.pdf)

SL3 - Robust Methods For The Analysis Of Complex Data

- Anthony Atkinson, Aldo Corbellini. **Introducing Prior Information into the Forward Search for Regression** (Full paper: 2829.pdf)
- Andrea Cerasa, Francesca Torti, Domenico Perrotta. **Analysis of complex data in official statistics** (Full paper: 3047.pdf)
- Pietro Coretto, Christian Hennig. **Robust covariance matrix estimation with regularization** (Full paper: 3046.pdf)
- Luis Garcıa-Escudero, Alfonso Gordaliza, Carlos Matran, Agustın Mayo-Iscar. **Robust model-based clustering and mixture** (Full paper: 3058.pdf)

SL4 - Statistics For Environmental Phenomena And Their Interactions

- Paola Berchialla, Marta Blangiardo, Michela Cameletti, Francesco Finazzi, Maria Villoria, Rosaria Ignaccolo. **Spatial modeling for air pollution epidemiology: hospital admission risk for cardio-respiratory diseases in Torino province** (Full paper: 3051.pdf)
- Rosaria Ignaccolo, Maria Franco Villoria. **Kriging for functional data: uncertainty assessment** (Full paper: 2953.pdf)
- Luigi Ippoliti, Lara Fontanella, Pasquale Valentini, Annalina Sarra, Sergio Palermi. **Bayesian structural equation modeling for factors influencing residential radon levels** (Full paper: 3023.pdf)
- Elena Scardovi, Francesca Bruno, Fedele Greco. **Assessment of bayesian models for rainfall field reconstruction** (Full paper: 3013.pdf)

SL5 - Mixture And Latent Variable Models For Causal Inference And Analysis Of Socio

- Silvia Bacci, Francesco Bartolucci. **A multidimensional finite mixture SEM for non-ignorable missing responses to test items** (Full paper: 2831.pdf)
- Paolo Giordani. **Finite mixtures for multivariate mixed data: a Parafac-based approach** (Full paper: 3072.pdf)

- Lucia Modugno, Silvia Cagnone, Simone Giannerini. **A multilevel model for repeated cross-sectional data with stochastic volatility** (Full paper: 3125.pdf)
- Leonardo Grilli, Fulvia Pennoni, Carla Rampichini, Isabella Romeo. **Multivariate multilevel modelling of student achievement data** (Full paper: 2959.pdf)

SL6 - Equity And Sustainability: Theory And Relationships

- Simone Bastianoni, Luca Coscieme, Federico Pulselli. **Categorization of National Economies through Environmental, Social and Economic Indicators.** (Full paper: 2981.pdf)
- Alessandra Ferrara, Angela Ferruzza, Nicoletta Pannuzi. **Environmental resources, landscape and cultural heritage, economic conditions: fundamental components to measure well-being** (Full paper: 3006.doc)
- Tommaso Luzzati, Bruno Cheli. **Measuring the sustainability performances of the Italian regions** (Full paper: 2987.pdf)
- Andrea Regoli, Antonella D'Agostino, Francesca Gagliardi, Laura Neri. **A framework for monitoring country sustainability** (Full paper: 2980.pdf)

SL7 - Advances In Bayesian Statistics

- Stefano Cabras, María Eugenia Castellanos Nueda, Erlis Ruli, Mario Pirastu, Maria Pina Concas. **An ABC/quasi-likelihood approach for linkage/GWAS study of a Sardinian genetic isolate** (Full paper: 3030.pdf)
- Pierpaolo De Blasi, Stephen Walker. **Bayesian nonparametric estimation and asymptotics with misspecified density models** (Full paper: 3087.pdf)
- Edward George, Veronika Ročkóva. **Determinantal Priors for Variable Selection** (Full paper: 3086.pdf)
- Brunero Liseo. **Approximate Bayesian Inference for Copula Models** (Full paper: 3084.pdf)

SL8 - Statistical Models For The Analysis Of Energy Markets

- Claudia Furlan, Mariangela Guidolin, Renato Guseo. **Are there any effects of Fukushima accident on the diffusion of nuclear energy?** (Full paper: 2826.pdf)
- Luigi Grossi, Fany Nan. **Robust forecasting of electricity prices with nonlinear models and exogenous regressors** (Full paper: 2920.pdf)
- Matteo Pelagatti, Lucia Parisio. **First results on the Italian-Slovenian electricity market coupling** (Full paper: 2958.pdf)
- Alessandro Sapio. **Renewable flows and congested lines in the Italian power grid: Binary time series and vector autoregressions** (Full paper: 2970.pdf)

SL9 - Recent Developments In Sampling Theory

- Pier Luigi Conti. **Weak convergence and empirical processes in survey sampling** (Full paper: 3066.pdf)
- Pier Francesco Perri, Lucio Barabesi, Giancarlo Diana. **On the linearization of inequality indexes in the design-based framework** (Full paper: 2992.pdf)
- Caterina Pisani, Lorenzo Fattorini. **Sampling strategies for diversity indexes estimation in presence of rare species** (Full paper: 3020.pdf)

- Emilia Rocco. **Spatially balanced adaptive web sampling** (Full paper: 3022.pdf)

SL10 - Functional Data Analysis

- Aldo Goia. **Some advances on semi-parametric functional data modelling** (Full paper: 3007.pdf)
- Elvira Romano, Jorge Mateu, Iulian Teodor Vlad. **A Functional Model for Detecting Changes in Evolving Shapes Brain Tumors** (Full paper: 3045.pdf)
- Laura Sangalli. **Functional data analysis in spaces of surfaces** (Full paper: 2986.pdf)
- Simone Vantini, Alessia Pini. **Hypothesis Testing in Functional Data Analysis: a Non-parametric Approach** (Full paper: 3001.pdf)

SL11 - Extremes And Dependent Sequences

- Paola Bortot, Carlo Gaetan. **Latent process models for temporal extremes with an application to rainfall data** (Full paper: 3142.pdf)
- Fabrizio Laurini. **The clustering of extreme values for some asymmetric GARCH-type models** (Full paper: 3143.pdf)
- Giulia Marcon, Simone A. Padoan, Philippe Naveau, Pietro Muliere. **Inference of multivariate dependence structures in extreme value theory** (Full paper: 2996.pdf)

SL12 - Issues In Ecological Statistics

- Linda Altieri, Marian Scott, Janine Illian. **A changepoint analysis on spatio-temporal point processes** (Full paper: 3009.pdf)
- Antonella Bodini, Gianni Gilioli. **Linking metapopulation modelling and Information Theory for area-wide pest management** (Full paper: 2983.pdf)
- Ilaria Rosati, Alessio Pollice, Serena Arima, Giovanna Jona Lasinio, Alberto Basset. **Uncertainty of methodologies assessing ecological status in transitional water systems** (Full paper: 3074.pdf)
- Luca Tardella, Danilo Alunni Fegatelli. **Modelling and inferential issues for behavioral patterns in capture-recapture data** (Full paper: 3059.pdf)

SL13 - Computations With Intractable Likelihood

- Isadora Antoniano-Villalobos, Stephen Walker. **Exact Bayesian inference for discretely observed diffusions** (Full paper: 3005.pdf)
- Luigi Augugliaro, Angelo Mineo. **An efficient algorithm to estimate the sparse group structure of an high-dimensional generalized linear model** (Full paper: 3002.pdf)
- Francesca Greselin, Luis Garcìa-Escudero, Alfonso Gordaliza, Salvatore Ingrassia, Agustìn Mayo-Iscar. **An adaptive method to robustify ML estimation in Cluster Weighted Modeling** (Full paper: 3004.pdf)
- Antonietta Mira, Alberto Caimo. **Delayed rejection algorithm to estimate Bayesian social networks** (Full paper: 3003.pdf)

SL14 - Geographical Information In Sampling And Estimation

- Roberto Benedetti, Federica Piersimoni, Paolo Postiglione. **Spatially balanced samples for land use/land cover surveys** (Full paper: 2997.pdf)
- Chiara Bocci, Emilia Rocco, Patrizia Lattarulo. **Modeling the location decisions of firms** (Full paper: 3136.pdf)
- Elisabetta Carfagna, Simone Maffei, Andrea Carfagna. **Geo-referenced information for agricultural statistics** (Full paper: 3069.pdf)
- Stefano Marchetti, Caterina Giusti, Nicola Salvati. **The use of geographic information under the area-level approach to small area estimation** (Full paper: 3024.pdf)

SL15 - Clinical Designs

- Andrea Ghiglietti, Anna Maria Paganoni. **Statistical properties of urn designs in clinical trials** (Full paper: 2852.pdf)
- Stefania Gubbiotti, Pierpaolo Brutti, Fulvio De Santis. **A predictive look at Bayesian Bandits** (Full paper: 2926.pdf)
- Paola Rebora, Maria Grazia Valsecchi. **Optimal two-phase design and incidence estimation in cohort studies** (Full paper: 2848.pdf)

SL16 - Bayesian Inference For High

- Veronika Ročkóva, Edward I. George. **Fast EM Inference for Bayesian Factor Analysis with Indian Buffet Process Prior** (Full paper: 3048.pdf)
- Mahlet Tadesse, Alberto Cassese, Michele Guindani, Francesco Falciani, Marina Vannucci. **A Unified Method for CNV Detection and Association with Gene Expression** (Full paper: 3049.pdf)
- Lorenzo Trippa. **Bayesian Adaptive Trials** (Full paper: 3057.pdf)

SL17 - Use Of Big Data For The Production Of Statistical Information

- Stefano Falorsi, Fabio Bacchini, Michele D'Alò, Andrea Fasulo, Carmine Pappalardo. **Does Google index improve the forecast of Italian labour market?** (Full paper: 3019.pdf)
- Dino Pedreschi, Roberta Vivio, Fosca Giannotti, Mirco Nanni, Barbara Furletti, Giuseppe Garofalo, Lorenzo Gabrielli, Letizia Milli,. **Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach** (Full paper: 3026.pdf)
- Dino Pedreschi. **Big Data, Social Mining, Diversity, and Wellbeing** (Full paper: 3107.pdf)
- Marco Scarnò, Sergio Salamone, Monica Scannapieco. **Web scraping and web mining: new tools for Official Statistics** (Full paper: 3008.pdf)

SL18 - Measuring The Smart City

- Roberta De Santis, Alessandra Fasano, Nadia Mignolli, Anna Villa. **Smart City: measuring a multidimensional topic** (Full paper: 3029.doc)
- Carlo Giovannella. **Smart Territory Analytics: toward a shared vision** (Full paper: 3031.pdf)
- Mariagrazia Dotoli, Vincenzo Patruno, Luigi Ranieri, Aldo Scarnera. **Staying On The Smart side. Measuring The Smart Communities** (Full paper: 3021.doc)
- Paolo Testa, Nicolò Marchesini. **Measuring the territory to build up the Smart City** (Full paper: 3037.doc)

SL19 - Forecasting Economic And Financial Time Series

- Roberto Baragona, Domenico Cucina. **Outliers in Time Series: an Empirical Likelihood Approach** (Full paper: 3093.pdf)
- Luc Bauwens, Manuela Braione, Giuseppe Storti. **Long term component dynamic models for realized covariance matrices** (Full paper: 3092.pdf)
- Roberto Casarin, Tilmann Gneiting, Francesco Ravazzolo. **Probabilistic Calibration of Predictive Distributions** (Full paper: 3134.pdf)
- Simone Giannerini, Esfandiar Maasoumi, Estela Bee Dagum. **Testing for nonlinear serial dependence in time series with surrogate data and entropy measures** (Full paper: 3123.pdf)

CONTRIBUTED PAPER SESSIONS

CP1 - Demography

- Gustavo De Santis, Mauro Maltagliati, Silvana Salvini. **A measure of the distance between countries based on individual data** (Full paper: 2880.doc)
- Antonino Di Pino, Maria Gabriella Campolo, Marcantonio Caltabiano. **Retirement and Intra-Household Labour Division of Italian Couples: A Simultaneous Equation Approach** (Full paper: 2840.pdf)
- Lucia Leporatti, Paolo Cremonesi, Enrico Di Bella, Luca Persico. **Demographic change and future sustainability of emergency departments: a pilot study for Liguria** (Full paper: 2932.pdf)
- Luigi Marcone, Francesco Borrelli, Stefania Di Domenico. **Error models for weighting estimators in the 15th Italian Population and Household Census** (Full paper: 2917.docx)
- Silvia Meggiolaro, Fausta Ongaro. **Father-child contact after marital dissolution. Evidence from Italy** (Full paper: 2847.pdf)

CP2 - Statistics In Finance

- Alessandra Amendola, Vincenzo Candila. **The use of loss functions in assessing the VaR measures** (Full paper: 2909.pdf)
- Paola Cerchiello, Paolo Giudici. **Systemic risk models** (Full paper: 2977.pdf)
- Giancarlo Ferrara, Francesco Vidoli, Arianna Campagna, Jacopo Canello. **Labour-use efficiency in the Italian machinery industry: a non-parametric stochastic frontier perspective** (Full paper: 2898.doc)

- Andrea Pierini. **Chain Graph for VAR and MARCH parameters reduction. EU index returns case.** (Full paper: 2882.pdf)

- Marianna Succurro, G. Damiana Costanzo. **Predicting financial bankruptcy by a (Robust) Principal Component Analysis based model: an empirical investigation.** (Full paper: 2935.docx)

CP3 - Statistics In Medicine

- Antonio Canale, Dipankar Bandyopadhyay. **Bayesian nonparametric spatial modelling of ordinal periodontal data** (Full paper: 2845.pdf)

- Silvia Liverani, Chris Cantwell. **Elicitation and visualisation of uncertainty in electrograms for activation time maps** (Full paper: 2875.pdf)

- Monia Lupparelli, Alberto Roverato. **Log-mean linear regression models for assessing the effect of HIV-infection on multimorbidity in a case-control study** (Full paper: 2951.pdf)

- Lorenzo Maragoni, Monica Chiogna. **A Quantile-based Test for Detecting Differential Expression in Microarray Data** (Full paper: 2964.pdf)

- Paolo Zanini, Piercesare Secchi, Simone Vantini. **EEG signals decomposition: a multi-resolution analysis** (Full paper: 2843.pdf)

CP4 - Clustering Methods: Theory And Applications

- Marzia Cremona, Pier Giuseppe Pelicci, Laura Riva, Laura Sangalli, Piercesare Secchi, Simone Vantini. **Cluster analysis on shape indices for ChIP-Seq data** (Full paper: 2904.pdf)

- Paolo Giordani, Maria Brigida Ferraro. **fclust: an R package for fuzzy clustering** (Full paper: 2832.pdf)

- Giovanna Menardi, Domenico De Stefano. **Modal Clustering of Social Networks** (Full paper: 2940.pdf)

- Andrea Pastore, Stefano Tonellato. **Identification of multiple clusterings using Gaussian mixtures** (Full paper: 2914.pdf)

CP5 - Functional Data Analysis

- Enea Bongiorno, Aldo Goia. **A clustering method for functional data** (Full paper: 2963.pdf)

- Tommaso Lando, Lucio Bertoli-Barsotti. **A class of bibliometric indices based on a sum of increasing and concave functions** (Full paper: 2902.docx)

- Alessia Pini, Konrad Abramowicz, Charlotte Häger, Kim Hébert-Losier, Lina Schelin, Johan Strandberg, Simone Vantini. **Anterior Cruciate Ligament Rupture: Functional Data Analysis of Knee Motion** (Full paper: 2842.pdf)

- Matteo Ruggiero, Omiros Papaspiliopoulos, Dario Spanò. **Bayesian inference for dynamically evolving distributions** (Full paper: 2820.pdf)

- Nicholas Tarabelloni, Rachele Biasi, Francesca Ieva, Anna Paganoni. **Depth measures for multivariate functional data with data-driven weights** (Full paper: 2834.pdf)

CP6 - Forensic Statistics

- Carlo Cusatelli, Massimiliano Giacalone. **Statistical comparison of European judicial systems according to ICT** (Full paper: 2968.doc)
- Antonia Rosa Gurrieri, Marilene Lorizio. **Efficiency of justice and justice's demand** (Full paper: 2954.doc)
- Laura Antonucci, Francesco D'Ovidio, Domenico Viola. **Measures of Courts Performances and Stochastic Frontier Models** (Full paper: 2974.doc)
- Annamaria Stramaglia, Marilene Lorizio. **The efficiency of justice in an economic perspective: the role of the supply side** (Full paper: 2899.doc)

CP7 - Economic Phenomena

- Alberto Arcagni, Michele Zenga. **Decomposition by sources of the ξ inequality index** (Full paper: 2853.pdf)
- Stefano Falorsi, Michele D'Alò, Silvia Loriga. **LFS quarterly small area estimation of youth unemployment at provincial level** (Full paper: 2945.pdf)
- Francesco Porro, Michele Mario Zenga. **The decomposition by subgroups of the inequality curve Z(p) and the inequality index ξ** (Full paper: 2972.pdf)
- Emanuela Raffinetti, Elena Siletti, Achille Vernizzi. **Inequality measures and the issue of negative incomes** (Full paper: 2894.pdf)
- Daniele Toninelli, Martin Beaulieu, Catalin Dochitoiu. **The Effects of a new Aggregation Structure on Consumer Price Index Estimates** (Full paper: 2892.pdf)

CP8 - Advances In Statistical Modelling

- Alessandra Brazzale, Valentina Mameli. **Small-sample likelihood asymptotics for the equi-correlated bivariate normal model** (Full paper: 2867.pdf)
- Carlo Orsi, Andrea Ongaro. **On Non-central Beta distributions** (Full paper: 2933.pdf)
- Marco Perone Pacifico. **Nonparametric Mode Hunting** (Full paper: 2871.pdf)
- Mario Romanazzi. **Discriminant analysis of von Mises - Fisher distributions** (Full paper: 2895.pdf)

CP9 - Developments In Bayesian Inference

- Pierpaolo Brutti. **Bayesian Inference for the Intrinsic Dimension** (Full paper: 2922.pdf)
- Maria Eugenia Castellanos, Stefano Cabras. **Zellner-Siow Priors for variable selection with censored data** (Full paper: 2886.pdf)
- Francesca Condino, Filippo Domma, Giovanni Latorre. **Likelihood and Bayesian estimation of P(Y < X) using lower record values from a general class of distribution** (Full paper: 2887.pdf)
- Elias Moreno, Francisco Vázquez-Polo, Miguel Negrín. **An objective Bayesian procedure for meta–analysis of binomial data** (Full paper: 2862.pdf)
- Bernardo Nipoti, Stefano Favaro. **Uncertainty quantification for Bayesian nonparametric estimators of rare species variety** (Full paper: 2957.pdf)

CP10 - Educational Statistics

- Matilde Bini, Lucio Masserini. **A finite mixture model approach on the first year university drop-out probability** (Full paper: 2900.doc)
- Stefano Cabras, Juan De Dios Tena Horrillo. **A Bayesian nonparametric modelling to estimate students' response to ICT investment** (Full paper: 2839.pdf)
- Dalit Contini, Elisa Grand. **On the development of school achievement inequalities with cross-sectional data** (Full paper: 2967.docx)
- Franca Crippa, Marcella Mazzoleni, Mariangela Zenga. **The role of the membership function to model university students' flow** (Full paper: 2955.pdf)
- Francesca Ieva, Tommaso Agasisti, Anna Paganoni. **Multilevel modeling of heterogeneity in math achievements: different class- and school-effects across Italian regions** (Full paper: 2833.pdf)

CP11 - Sanitary Statistics And Epidemiology

- Domenico De Stefano, Stefano Campostrini. **The 2008 Great Recession and Health in Italy. A Study on the Surveillance Data System PASSI** (Full paper: 2962.pdf)
- Marco Geraci, Alessio Farcomeni. **A probabilistic approach to the estimation of principal components with nonignorable missing data: Applications in accelerometer-based physical activity studies** (Full paper: 2881.pdf)
- Stefano Mazzuco, Bruno Scarpa, Lucia Zanotto. **A mortality model based on a mixture distribution function** (Full paper: 2897.pdf)
- Inad Nawajah, Raffaele Argiento, Alessandra Guglielmi, Ettore Lanzarone. **Joint Prediction of Demand and Care Duration in Home Care Patients: a Bayesian Approach** (Full paper: 2893.pdf)
- Emiliano Sironi, Massimo Cannas. **Hospital Differences in Caesarean Deliveries in Sardinia: A Multilevel Analysis** (Full paper: 2948.doc)

CP12 - Survey Methodology

- Alessio Farcomeni. **Heterogeneity for a general class of recapture models based on equality constraints on the conditional capture probabilities** (Full paper: 2822.pdf)
- Pier Luigi Conti, Daniela Marella, Mauro Scanu. **Uncertainty in statistical matching for complex sample surveys** (Full paper: 2859.pdf)
- Flaminia Musella, Daniela Marella, Paola Vicard. **Learning Bayesian networks in complex survey sampling** (Full paper: 2865.pdf)
- Leo Pasquazzi, Lucio De Capitani. **Quantile estimation with auxiliary information** (Full paper: 2857.pdf)
- Silvia Polettini, Serena Arima. **Small Area Estimation with Covariates Perturbed for Disclosure Limitation** (Full paper: 2915.pdf)

CP13 - Statistical Methods For The Analysis Of Fertility And Health

- Bianca Destavola, Lorenzo Richiardi, Daniela Zugna, Rhian Daniel, Rossella Murtas. **Birth Order, Birth Weight and Asthma: how to assess mediation and the presence of Unmeasured Confounding** (Full paper: 2944.pdf)
- Alessio Fornasin, Marco Breschi, Matteo Manfredini, Massimo Esposito. **Reproductive Change in Transitional Italy: Insights from the Italian Fertility Survey of 1961** (Full paper: 2835.pdf)
- Haftu Gebremeskel, Stefano Mazzuco. **Implementing Hierarchical Bayesian Model to Fertility Data: the case of Ethiopia** (Full paper: 2960.pdf)
- Stanislao Mazzoni, Lucia Pozzi, Marco Breschi. **Fertility and Child Mortality in the Sardinian Demographic Transition. Alghero (1866-1935)** (Full paper: 2901.doc)
- Daria Mendola, Annalisa Busetta, Daniele Vignoli. **Persistent Employment Instability and Fertility Intentions** (Full paper: 2907.docx)

CP14 - Advanced In Compositional Data Analysis

- Domenico De Stefano, Maria Rosaria D'Esposito, Giancarlo Ragozini. **Multiple Factor Analysis to Visually Explore Collaboration Structures: the Case of Technological Districts** (Full paper: 2971.pdf)
- Josep Martin-Fernandez, Josep Daunis-i-Estadella, Santiago Thió-Henestrosa. **Information Provided by Absolute, Essential and Structural Zeros in Compositional Data Sets** (Full paper: 2906.docx)
- Alessandra Menafoglio, Alberto Guadagnini, Piercesare Secchi. **Kriging prediction for functional compositional data and application to particle-size curves** (Full paper: 2836.pdf)
- Gianna Monti, Gloria Mateu-Figueras, Vera Pawlowsky-Glahn, Juan José Egozcue. **Scaled-Dirichlet regression for compositional data** (Full paper: 2928.pdf)
- Nickolay Trendafilov, Michele Gallo. **Sparse PCA for compositional data** (Full paper: 2863.pdf)

CP15 - Spatial And Spatio

- Manuela Cattelan, Cristiano Varin. **Composite likelihood estimation in spatial logistic regression** (Full paper: 2924.pdf)
- Enrico Foscolo, Marta Disegna, Fabrizio Durante. **A copula model for tourists' spending behavior** (Full paper: 2949.doc)
- Margherita Gerolimetto, Luisa Bisaglia, Paolo Gorgi. **Estimation and forecasting for binomial and negative binomial INAR(1) time series** (Full paper: 2913.pdf)
- Luca Romagnoli, Luigi Ippoliti, Richard Martin. **Kalman Filter for Estimating Bivariate GMRFs on Regular Lattice** (Full paper: 2919.pdf)
- Grazia Vicario, Giovanni Pistone. **A note on semivariogram** (Full paper: 2925.pdf)

CP16 - Environmental And Poverty Data Analysis

- Lucia Paci, Daniela Cocchi, Alan Gelfand. **Quantifying uncertainty associated with a numerical model output** (Full paper: 2824.pdf)
- Giuliana Passamani, Paola Masotti. **Smoothed Common Trend in Multivariate Time Series Air Pollution Data** (Full paper: 2952.doc)
- Silvia Perra, Stefano Cabras, Alberto Serci, Alessandra Mura, Antonella De Arca, Stefano Renoldi, Antonello Podda. **IDMS: The Sardinian Index of Multiple Deprivation** (Full paper: 2860.doc)
- Alessio Pollice, Vito Muggeo, Federico Torretta, Rocco Bochicchio, Mariana Amato. **Growth curves of sorghum roots via quantile regression with P-splines** (Full paper: 2956.pdf)
- Abel Rodriguez, Fernado Quintana. **On species sampling sequences induced by residual allocation models** (Full paper: 2885.pdf)

CP17 - Topics In Regression Models

- Matilde Bini, Vieri Del Panta, Margherita Velucchi. **Mixtures of Logit Regressions Detection with Forward Search** (Full paper: 2982.pdf)
- Silvia Columbu, Matteo Bottai. **Conditional concordance of the signs of the residuals of quantiles regressions of multivariate outcomes** (Full paper: 2965.pdf)
- Roberto Fontana, Fabio Rapallo, Maria Piera Rogantin. **Indicator functions and saturated fractions for factorial designs: a case study** (Full paper: 2888.pdf)
- Pia Clara Pafundi, Gianmarco Vacca. **Complex Redundancy Analysis models with covariate effect: a simulation study** (Full paper: 2869.doc)
- Mariangela Sciandra. **Variable selection in mixed models: a graphical approach** (Full paper: 2931.pdf)

CP18 - Bayesian Methods And Models

- Julyan Arbel, Kerrie Mengersen, Judith Rousseau. **On diversity under a Bayesian nonparametric dependent model** (Full paper: 2936.pdf)
- Antonio Canale, Bruno Scarpa. **Skew-normal nonparametric mixture models** (Full paper: 2844.pdf)
- Giulia Roli, Meri Raggi. **Bayesian hierarchical models for misaligned data: a simulation study** (Full paper: 2927.pdf)
- Catia Scricciolo, Sophie Donnet, Vincent Rivoirard, Judith Rousseau. **Posterior contraction rates for empirical Bayes procedures with applications** (Full paper: 2918.pdf)

POSTER SESSION

- Antonio Arcos, María Del Mar Rueda, David Molina, Maria Giovanna Ranalli. **Frames2: an R package for estimation in dual frames** (Full paper: 2993.pdf)
- Simona Arcuti, Alessio Pollice, Crescenza Calculli, Nunziata Ribecco, Angelo Tursi. **Spatial smoothing on complex regions: a case study on the median length of deep water rose shrimps in the North-Western Ionian Sea** (Full paper: 2921.pdf)

- Filippa Bono, Marcella Giacomarra. **The effect of support schemes on Photovoltaic installed capacity in Europe: a WDEA-STATIS analysis** (Full paper: 2988.docx)
- Maria Caterina Bramati, Flaminia Musella. **Bayesian Network structural learning in multivariate time series** (Full paper: 2856.pdf)
- Maurizio Brizzi, Alessia Orrù. **Gender Differentiation of Human Longevity in Sardinian Provinces** (Full paper: 2943.doc)
- Simone Del Sarto, M. Giovanna Ranalli, Elena Stanghellini, Davide Cappelletti, Beatrice Moroni, Stefano Crocchianti, Silvia Castellini. **Modelling the effect of vehicular–traffic and meteorology on fine particle concentration using Additive Mixed Models: the case of the town of Perugia** (Full paper: 3148.pdf)
- Rosa Falotico, Paolo Mariani. **Outsourcing in the Italian NHS: a statistical measure of mismatch between private supply and public demand** (Full paper: 2929.doc)
- Francesca Fortuna, Fabrizio Maturo. **Functional analysis of variance for parametric functional data** (Full paper: 3150.pdf)
- Massimiliano Giacalone, Paolo Carmelo Cozzucoli. **A performance comparison of the Lp-norm methods in multicollinearity situations, supposing a generalized normal distribution errors** (Full paper: 2969.doc)
- Clara Grazian, Christian Robert. **Jeffreys Priors for Mixture Models** (Full paper: 2891.pdf)
- Fabrizio Maturo, Francesca Fortuna. **Bell shaped fuzzy numbers associated with the normal curve** (Full paper: 3149.pdf)
- Daria Mendola, Annalisa Busetta, Philippe Van Kerm, Anna Maria Milito. **Material deprivation among foreigners in Italy** (Full paper: 2890.doc)
- Eugenia Nissi, Annalina Sarra. **Local Spatial Analysis of Cardiovascular Diseases in Canadian Health Regions** (Full paper: 3151.pdf)
- Zoe Nivolianitou, Myrto Konstandinidou, Chrys Caroni, Irini Kefalogianni. **Analysis of the causal factors in incidents of the Greek petrochemical industry** (Full paper: 2866.doc)
- Anna Pinto, Giulia Mascarello, Nicoletta Parise, Silvia Bonaldo, Stefania Crovato, Licia Ravarotto. **A classification of Italian consumers based on a proposed measure of their attitudes towards food risks** (Full paper: 2896.pdf)
- Giovanni Pistone. **A version of the geometry of the multivariate Gaussian model, with applications** (Full paper: 2855.pdf)
- Pasquale Recchia, Ernesto Toma. **Family size and educational outcomes: empirical evidence through multilevel approach** (Full paper: 2884.docx)
- Maria Piera Rogantin, Giovanni Pistone. **Fractionalization and Polarization** (Full paper: 2910.pdf)
- Elvira Romano, Maria Dolores Ruiz-Medina, Rosa M. M. Espejo. **A spatial functional approach for curve classification** (Full paper: 3110.pdf)
- Giorgio Russolillo, Laura Trinchera. **An extension of Non-Metric approach to inwards directed PLS Path Models** (Full paper: 3036.pdf)
- Aldo Solari, Jelle Goeman. **Adapting Benjamini-Hochberg by Simes Inequality** (Full paper: 3083.pdf)
- Susanna Zaccarin, Domenico De Stefano, Vittorio Fuccella, Maria Vitale. **Co-authorship Patterns of Italian Statisticians by Combining Different Data Sources** (Full paper: 2941.pdf)

An adaptive method to robustify ML estimation in Cluster Weighted Modeling

Stima robusta per Cluster Weighted Model

L.A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar

Abstract Cluster-Weighted Models are a wide family of mixture distributions for modeling the joint probability of data coming from a heterogeneous population, and includes mixtures of distributions and mixtures of regressions as special cases. Unfortunately, they suffer from non-regular maximum likelihood issues, due to possible spikes and unboundedness in the target function. We propose an improved version of the Gaussian Cluster-Weighted estimation methodology, by trimming a portion α of the data and imposing constraints to the estimated variances. Trimming provides robustness properties to the estimators and constraints move the maximization problem to a well-posed setting and allow to avoid spurious solutions, i.e. fitting a small localized random pattern in the data rather than a proper underlying cluster structure. Theoretical results are illustrated using a few empirical studies.

Abstract *I modelli Cluster Weighted sono una ampia famiglia di misture che comprendono, come casi particolari, le misture di distribuzioni e di regressione e consentono di modellizzare dati eterogenei. Per questi modelli si osservano problemi di non regolarità della funzione di verosimiglianza, che può essere illimitata e avere spikes. In questo lavoro si introduce un nuovo metodo di stima del Cluster Weighted Gaussiano, che consente di eliminare una porzione α di dati contaminati e di imporre vincoli alla stima delle matrici di covarianza. Le proprietà di robustezza degli stimatori sono assicurate dall'eliminazione degli outliers, e la stima vincolata fa sì che il problema di massimizzazione sia ben posto e riduce le soluzioni spurie, ovvero l'adattarsi del modello ad un piccolo raggruppamento casuale di dati invece che ad una vera e propria componente della mistura. I risultati teorici sono illustrati anche mediante alcune analisi empiriche.*

Luis Angel García-Escudero · Alfonso Gordaliza · Agustín Mayo-Iscar
Departamento de Estadística e Investigación Operativa, Universidad de Valladolid (Spain) e-mail: lagarcia@eio.uva.es, alfonsog@eio.uva.es, agustinm@eio.uva.es

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

Salvatore Ingrassia
Department of Economics and Business, University of Catania (Italy). e-mail: s.ingrassia@unict.it

Key words: Constrained estimation, Cluster Weighted Modeling, Mixture of regression, Model-Based Clustering.

1 Introduction and Motivation

The analysis of mixture models is a fertile source of non-regular maximum likelihood problems. For instance, a two-component normal mixture incurs the problem of unbounded likelihood if the mean parameter of the first component is set to be one of the data values and the standard deviation σ is allowed to tend to 0. However, the singularity does not impose itself until σ becomes extremely small. In many normal mixture problems susceptible to unbounded likelihood, there is also, an asymptotically consistent local maximum (Redner and Walker, 1984), but still spurious solutions could drive the maximization of the target function far away from the true value of the parameter. Moreover, it is well known that a small fraction of outlying observations (background noise, pointwise contamination, unexpected minority patterns, etc.) could severely affect ML parameter estimation. With these considerations in mind, we approach Cluster-Weighted Models (CWMs), introduced in Gershenfeld (1997). They are a flexible family of mixture models for fitting the joint density of a pair $(\mathbf{X};Y)$ composed by a response variable Y and by a vector of covariates $\mathbf{X}$, assuming that data are coming from a heterogenous population. Ingrassia *et al.* (2012) show that Gaussian CWM includes, as special cases, mixtures of distributions and finite mixture of regression models. Our purpose is to modify the classical ML method, by adding trimming and constraints in such a way to make robust and free from non-regularity conditions the model estimation. We have organized the rest of the paper as follows. In Section 2 we recall the main ideas about the CWM and we discuss issues in EM estimation. In Section 3 we present the trimmed CWRM and a feasible algorithm for its implementation. Concluding remarks will end the paper.

2 Cluster Weighted Modeling

Let $p(\mathbf{x},y)$ be the joint density of $(\mathbf{X},Y)$. Suppose that Ω can be partitioned into G groups, say $\Omega_1,\ldots,\Omega_G$. In this work, we will focus on CWM linear models with Gaussian components: assuming $\mathbf{X}|\Omega_g \sim N_d(\mu_g,\Sigma_g)$, a linear relationship between Y and $\mathbf{x}$ in the g-th group written as $Y=\mathbf{b}_g'\mathbf{x}+b_g^0+\varepsilon_g$ where $\varepsilon_g \sim N(0,\sigma_g^2)$, and $Y|\mathbf{x},\Omega_g \sim N(\mathbf{b}_g'\mathbf{x}+b_g^0,\sigma_g^2)$, the *linear Gaussian CWM* has the following density $p(\mathbf{x},y;\theta)=\sum_{g=1}^G \phi(y;\mathbf{b}_g'\mathbf{x}+b_g^0,\sigma_g)\phi_d(\mathbf{x};\mu_g,\Sigma_g)\pi_g$.As usual, $\phi_d(\cdot;\mu_g,\Sigma_g)$ is the density of the d-variate Gaussian distribution with mean vector μ_g and covariance matrix Σ_g, and π_g is the weight of Ω_g in the mixture.The ML estimation of the Gaussian CWM suffers from a serious lack of robustness, which should be taken into account due to the common presence of noise sources in data. To illustrate this problem, Figure 1(a) shows a simulated data set, *Simdata1*, generated from

a linear Gaussian CWM with $G = 2$, 90 observations from each component. 20 contaminating observations have been added as either background noise, or pointwise contamination around the point $(15, 20)$. The true underlying regression lines are represented with dotted lines in Figure 1, which shows that contaminating data points seriously affected the estimation.

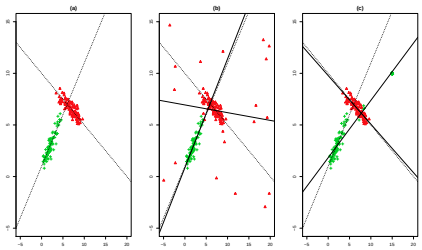

Fig. 1 *Simdata1*: (a) original data and Cluster Weighted Model fitting; (b) original data plus background noise and fitted model; (c) original data plus pointwise contamination and fitted model.

Another important issue concerns the unboundedness of the target function $\sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \phi(y_i; \mathbf{b}_g' \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \pi_g \right]$, when no constraints are imposed on the scatter parameters. In this case, the defining problem is ill-posed because the target function tends to ∞ when either $\mu_g = \mathbf{x}_i$ and $|\Sigma_g| \to 0$ or $y_i = \mathbf{b}_g' \mathbf{x}_i + b_g^0$ and $\sigma_g^2 \to 0$. Moreover, as a trivial consequence of the unboundedness, the EM algorithms often applied to fit a CWM can be trapped into non-interesting local maximizers and the result of the EM algorithm strongly depends on the initialization of the algorithm. Spurious solutions may be due to very localized patterns in the explanatory variables, as it will be shown by a second simulated data set, *Simdata2*. In Figure 2, two sets of 90 observations for $\mathbf{X}$ were drawn from two bivariate spherical normal distributions centered, respectively, at $(2,2)$ and $(4,4)$. Furher, 20 almost collinear observations, were added close to the second group, centered at $(4,4)$. The same linear functions with equally distributed error terms have been considered, to generate the response variable Y. We can see in Figure 2 that the standard fit of the CWM yields to the determination of a "spurious" component with the 20 almost collinear observations and a second component joining together the two groups, with 90% of the observations. To overcome the issues we illustrated in these examples, in the next section we will propose an alternative methodology which incorporates trimming and constraints to the CWM.

3 Trimmed Cluster Weighted Restricted Modeling

For a given sample of n observations, the trimmed CWRM methodology is based on the maximization of the following log-likelihood function

$$\sum_{i=1}^{n} z(\mathbf{x}_i, y_i) \log \left[\sum_{g=1}^{G} \phi(y_i; \mathbf{b}_g' \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \pi_g \right],$$

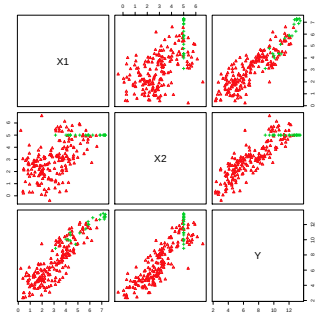

Fig. 2 *Simdata2*: Scatter plot matrix. Dots in green concern collinear observations.

where $z(\cdot,\cdot)$ is a 0-1 trimming indicator function that tell us whether observation $(y_i,\mathbf{x}_i)$ is trimmed off ($z(\mathbf{x}_i,y_i)$=0), or not ($z(\mathbf{x}_i,y_i)$=1). A fixed fraction α(trimming level) of observations can be unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i,y_i) = [n(1-\alpha)]$. Analogous approaches based on trimmed mixture likelihoods can be found in Neykov *et al.* (2007), Gallegos and Ritter (2009) and García-Escudero *et al.* (2013). Moreover, we introduced two further constraints, on the set of eigenvalues $\{\lambda_l(\Sigma_g)\}_{l=1,\ldots,d}$ of the scatter matrices Σ_g

$$\lambda_{l_1}(\Sigma_{g_1}) \leq c_X \lambda_{l_2}(\Sigma_{g_2}) \text{ for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G$$

and on the variances σ_g^2 of the regression error terms,

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \text{ for every } 1 \leq g_1 \neq g_2 \leq G \quad \text{ with } 0 < c_X, c_\varepsilon < +\infty.$$

These constraints can be seen as an extension to CWMs of those introduced in Ingrassia and Rocci (2007), García-Escudero *et al.* (2008) and Greselin and Ingrassia (2010) and go back to Hathaway (1985). Here, they allow for a specific treatment when modeling the marginal distribution of $\mathbf{X}$ and the regression error term, giving a high flexibility to the model.

Let us consider now the effects of trimming in the two data sets derived from *Simdata1*. In Figure 3 we can see that setting $\alpha = 0.1$ allows to restore the true structure of the data, by discarding the outlying observations, both in the case of background noise and huge pointwise contamination. Hence, trimming modifies the ML estimation in such a way that it is no more influenced by potential outliers and drives it far from the previous bad results shown in Figure 1. Commenting the use of constraints for *Simdata2*, we can see in Figure 3 how a moderate choice for c_X and c_ε allows to correctly detect the $G = 2$ main groups and to avoid the disturbing effect of the "spurious" pattern in the data. More information about the role played by the parameters α, c_X and c_ε could be given, omitted here for the sake of space.

3.1 Algorithm

The maximization of the target function on its parameters under the bounds given by c_X and c_ε is not an easy task, obviously. We will give a feasible algorithm ob-

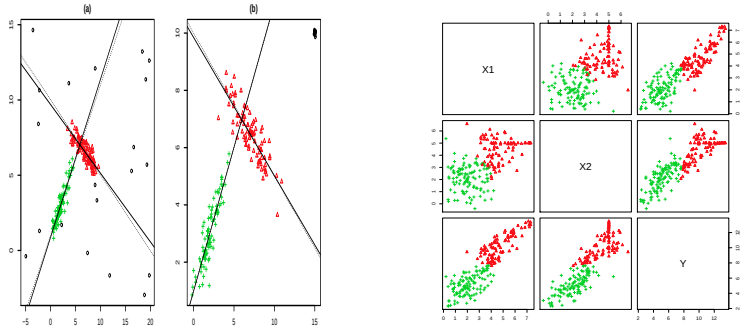

Fig. 3 Left panels:*Simdata1*: Results of fitting the trimmed CWRM with $\alpha = 0.1$, $c_X = c_\varepsilon = 20$ for the two data sets in Figure 1, panels (b) and (c). Trimmed points are denoted by black circles. Right panel: *Simdata2*: Results of fitting the trimmed CWRM with $\alpha = 0$, $c_X = c_\varepsilon = 20$ for data in Figure 2.

tained by combining the EM algorithm for CWM with that (with trimming and constraints) introduced in García-Escudero *et al.* (2013). The algorithm is initialized *nstart* times, by selecting different values of the initial parameter vector θ_0. Constraints on scatter matrices and variances of the error terms should be enforced (as described below, in the M-step). We will have adapted EM steps, alternatively executed until convergence. During the E-step:

- the current mixture density has to be evaluated at each observation in the sample, following the CWM methodology
- the proportion $1-\alpha$ of observations with highest values of the density is retained, so giving the subset I_α of untrimmed observations
- the posterior probabilities of the observations in I_α are computed, for trimmed observations they are set to 0.

During the M-step the parameters are updated, taking into account only the observations in I_α, i.e. tentatively discarding the observations suspicious to be outliers. Along the iterations, due to the updates, it may happen that the estimated scatter matrices T_g and the estimated variances of the error terms s_g^2 do not satisfy the constraints. To enforce them, the singular-value decomposition of $T_g = U_g' E_g U_g$ is considered, with U_g being an orthogonal matrix and $E_g = \text{diag}(e_{g1}, e_{g2}, ..., e_{gd})$ a diagonal matrix. The truncated eigenvalues are defined as $[e_{gl}]_m = \min(c_X \cdot m, (\max(e_{gl}, m))$, with m being some threshold value. The scatter matrices are finally updated as $\Sigma_g^{(l+1)} = U_g' E_g^* U_g$, with $E_g^* = \text{diag}\left([e_{g1}]_{m_{\text{opt}}^X}, [e_{g2}]_{m_{\text{opt}}^X}, ..., [e_{gp}]_{m_{\text{opt}}^X}\right)$ and m_{opt}^X minimizing the real valued function

$$m \mapsto \sum_{g=1}^{G} \pi_g^{(l+1)} \sum_{l=1}^{d} \left(\log\left([e_{gl}]_m\right) + \frac{e_{gl}}{[e_{gl}]_m} \right).$$

Analogously, we introduce the truncated variances $[s_g^2]_m = \min(c_\varepsilon \cdot m, (\max(s_g^2, m))$. The variance of the error terms are finally updated as $\sigma_g^{2(l+1)} = [s_g^2]_{m_{\text{opt}}^\varepsilon}$, with $m_{\text{opt}}^\varepsilon$ minimizing the real valued function

$$m \mapsto \sum_{g=1}^{G} \pi_g^{(l+1)} \left(\log\left([s_g^2]_m\right) + \frac{s_g^2}{[s_g^2]_m} \right).$$

Proposition 3.2 in Fritz *et al.* (2013) shows that m_{opt}^X and $m_{\text{opt}}^\varepsilon$ can be obtained, respectively, by evaluating $2pG+1$ times (respectively $2G+1$ times) the corresponding real valued function.

Finally, at convergence, the set of parameters yielding the highest value of the target function and the associated set I_α of untrimmed observations are returned as the final algorithm output.

In this work, we have presented a methodology based on trimming and constraints to robustify and control variabilities in a linear Gaussian CWM, moving the likelihood maximization to a well-posed setting. An algorithm, with an affordable increase in computing time, has been also given for its practical implementation. We have seen that the proposed methodology drives the estimation procedure to identify and discard sparse outliers, and even strongly concentrated contaminating observations, acting as leverage points, which are so harmful in the framework of regression mixtures. At the same time, the constraints serve to avoid the likelihood singularities and reduce the detection of spurious solutions.

Further research is needed to tune the choice of the involved parameters, and this is not an easy task, as these parameters are clearly interrelated. First attempts to extract such information from the observed sample are currently under study.

References

Fritz, H., García-Escudero, L., and Mayo-Iscar, A. (2013). A fast algorithm for robust constrained clustering. *Computational Statistics & Data Analysis*, (61), 124–136.

Gallegos, M. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhya (Ser. A)*, (71), 164–220.

García-Escudero, L., Gordaliza, A., and Mayo-Iscar, A. (2013). A constrained robust proposal for mixture modeling avoiding spurious solutions. Accepted for publication in *Advances in Data Analysis and Classification*.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

Gershenfeld, N. (1997). Nonlinear inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

Greselin, F. and Ingrassia, S. (2010). Constrained monotone EM algorithms for mixtures of multivariate *t* distributions. *Statistics and computing*, **20**(1), 9–22.

Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

Ingrassia, S. and Rocci, R. (2007). Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Comput. Stat. Data Anal.*, (51), 5339–5351.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via the Cluster-Weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, **52**(1), 299–308.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, pages 195–239.