



SIS - CLADAG



CLADAG 2015

10° Scientific Meeting of the Classification and Data Analysis
Group of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula,
October 8-10, 2015

BOOK OF ABSTRACTS

Editors:

Francesco Mola, Claudio Conversano
CUEC Editrice, Cagliari



CUEC
editrice

ISBN: 978 88 8467 749 9



Univeristà degli Studi
di Cagliari



Fondazione
Banco di Sardegna

CUEC EDITRICE

by Sardegna Novamedia Soc. Coop.

Via Basilicata n. 57/59

09127 Cagliari,

ITALY

Tel. & Fax +39 070 271573

www.cuec.eu

info@cuec.eu

ISBN: 978-88-8467-749-9

First Edition CUEC © 2015

PREFACE

CLADAG 2015, the 10th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society (SIS), will be held in Santa Margherita di Pula, Cagliari, Italy, from October 8th to October 10th 2015. The local organizer is the Department of Business and Economics of the University of Cagliari.

CLADAG 2015 will take place under the auspices of the International Federation of Classification Societies (IFCS) and of the Italian Statistical Society (SIS). It promotes advanced methodological research in multivariate statistics with a special vocation in Data Analysis and Classification. CLADAG supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results. It will also benefit of the support of Fondazione Banco di Sardegna.

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial scientific meeting, schools related to classification and data analysis, publishes a newsletter, and cooperates with other member societies of the IFCS to the organization of their conferences.

The scientific program comprises three Keynote Lectures, an Invited Session, 10 Specialized Sessions, 15 Solicited Sessions and 15 Contributed Sessions. All the Specialized and Solicited Sessions have been promoted by the members of the Scientific Program Committee. The organizers wish to thank them for their cooperation in contributing to the success of CLADAG 2015.

The Book of Abstracts contains short papers of all the presentations scheduled in the conference program. It is organized according to type of session/lecture: Keynote Lectures, Specialized Sessions, Solicited Sessions and Contributed Sessions.

The editors would like to express their gratitude to the Rector of the University of Cagliari, the Director of the Department of Business and Economics and to all the statisticians working in the Department of Business and Economics for their enthusiasm in supporting the organization of this event from the very beginning, as well as to all people who worked hard to make it a success. Special thanks go to Dr. Massimo Cannas, Dr. Luca Frigau and Dr. Farideh Tavazoei for their editorial support

Last but not least, we thank all authors and participants, without whom the conference would not have been possible.

Cagliari, October 8 2015.

*Francesco Mola,
Claudio Conversano*

CONFERENCE THEMES

The 10th Meeting is orientated towards all topics related to data analysis, classification, multivariate and computational statistics. Submission of papers addressing these topics in both methodological and practical perspective has been encouraged by the members of the Scientific Program Committee.

The list of topics includes, but is not limited to, the following:

A Classification Theory

Bayesian Classification Biplots Clustering models Consensus of Classifications Correspondence Analysis Discrimination and Classification Factor Analysis and Dimension Reduction Methods Fuzzy Methods Genetic Algorithms Hierarchical Classification Multidimensional Scaling Multiway Scaling Multiway Methods Neural Networks for Classification Non Hierarchical Classification Similarities and Dissimilarities Software algorithms for classification Unfolding and Related Scaling Methods

B Data Analysis

Bayesian data Analysis Big data analysis- Categorical Data Analysis Covariance Structure Analysis Data Mining Data Science Data Visualization Decision Trees Functional data analysis Mixture and Latent Class Models Multilevel data Analysis Non Linear Data Analysis Nonparametric and Semiparametric Regression Partial Least Squares Pattern recognition Robustness and Data Diagnostics Social networks- Software algorithms for multivariate analysis Spatial Data Analysis Symbolic Data Analysis.

COMMITTEES

Scientific Program Committee

Chair: Paolo Giudici (*University of Pavia*)

Members

Giuseppe Bove (*University of Roma Tre*)
Daniela Calo (*University of Bologna*)
Agostino Di Ciaccio (*University of Roma La Sapienza*)
Vincenzo Esposito Vinzi (*ESSEC, France*)
Francesca Greselin (*University of Milano Bicocca*)
Francesco Mola (*University of Cagliari*)
Francesco Palumbo (*University of Naples Federico II*)
Carla Rampichini (*University of Firenze*)
Giancarlo Ragozini (*University of Naples Federico II*)
Fabrizio Ruggeri (*CNR IMATI, Milan*)
Silvia Salini (*University of Milan*)
Adalbert F.X. Wilhelm (*Jacobs University Bremen, Germany*)

Local Organizing Committee

Chair: Francesco Mola (*University of Cagliari, Italy*).

Members: Stefano Cabras, Massimo Cannas, Claudio Conversano, Luca Frigau, Monica Musio, Mariano Porcu, Luisa Salaris, Isabella Sulis, Nicola Tedesco.

PARTICIPATING ORGANIZATIONS



International Federation of Classification Societies (IFCS)



Società Italiana di Statistica (SIS)



SIS - CLADAG (Classification and Data Analysis Group of the Italian Statistical Society)



Università degli Studi di Cagliari



Fondazione Banco di Sardegna

Table of Contents

Keynote Lectures

MINING KEY NETWORKS	21
<i>David Banks</i>	
VARIABLE SELECTION FOR MODEL-BASED CLUSTERING OF CATEGORICAL DATA	22
<i>Brendan Murphy</i>	
EIGENVALUES IN MIXTURE MODELING: GEOMETRIC, ROBUSTNESS AND COMPUTATIONAL ISSUES	23
<i>Salvatore Ingrassia</i>	

Specialized sessions

Robust methods for the analysis of Economic (Big) data

Organizer and Chair: Silvia Salini

FAST AND ROBUST SEEMINGLY UNRELATED REGRESSION	25
<i>Mia Hubert, Tim Verdonck and Ozlem Yorulmaz</i>	
APPLICATION TO THE DETECTION OF CUSTOMS FRAUD OF THE GOODNESS-OF-FIT TESTING FOR THE NEWCOMB-BENFORD LAW	30
<i>Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta</i>	
MONITORING THE ROBUST ANALYSIS OF A SINGLE MULTIVARIATE SAMPLE	34
<i>Marco Riani, Anthony C. Atkinson and Andrea Cerioli</i>	

Bayesian nonparametric clustering

Organizer: Fabrizio Ruggeri; Chair: Renata Rotondi

A BAYSIAN NONPARAMETRIC APPROACH TO MODEL ASSOCIATION BETWEEN CLUSTERS OF SNPS AND DISEASE RESPONSES	39
<i>Raffaele Argiento, Alessandra Guglielmi, Chuhsing Kate Hsiao, Fabrizio Ruggeri and Charlotte Wang</i>	
A BAYESIAN NONPARAMETRIC MODEL FOR CLUSTERING AND BORROWING INFORMATION	43
<i>Antonio Lijoi, Bernardo Nipoti and Igor Prünster</i>	
SEQUENTIAL CLUSTERING BASED ON DIRICHLET PROCESS PRIORS	47
<i>Roberto Casarin, Andrea Pastore and Stefano F. Tonellato</i>	

Causal Inference with Complex Data Structures

Organizer and Chair: Alessandra Mattei

SHORT TERM IMPACT OF PM10 EXPOSURE ON MORTALITY: A PROPENSITY SCORE APPROACH	52
--	----

Michela Baccini, Alessandra Mattei and Fabrizia Mealli

IDENTIFICATION AND ESTIMATION OF CAUSAL MECHANISMS IN CLUSTERED ENCOURAGEMENT DESIGNS: DISENTANGLING BED NETS USING BAYESIAN PRINCIPAL STRATIFICATION	54
---	----

Laura Forastiere, Fabrizia Mealli and Tyler VanderWeele

THE EFFECTS OF A DROPOUT PREVENTION PROGRAM ON SECONDARY STUDENTS' OUTCOMES	56
---	----

Enrico Conti, Silvia Duranti, Alessandra Mattei, Fabrizia Mealli and Nicola Sciclone

Clustering in Time Series

Organizer and Chair: Michele La Rocca

PROBABILISTIC BOOSTED-ORIENTED CLUSTERING OF TIME SERIES	61
--	----

Antonio D'Ambrosio, Gianluca Frasso, Carmela Iorio and Roberta Siciliano

COPULA-BASED FUZZY CLUSTERING OF TIME SERIES	65
--	----

Pierpaolo D'Urso, Marta Disegna and Fabrizio Durante

COMPARING MULTI-STEP AHEAD FORECASTING FUNCTIONS FOR TIME SERIES CLUSTERING	69
---	----

Marcella Corduas and Giancarlo Ragozini

Multiway Analysis

Organizer and Chair: Giuseppe Bove

(INTERACTIVE) VISUALISATION OF THREEWAY DATA	74
--	----

Casper J. Albers and John C. Gower

ROBUST FUZZY CLUSTERING OF MULTIVARIATE TIME TRAJECTORIES	78
---	----

Pierpaolo D'Urso and Riccardo Massari

ESTIMATION PROCEDURES FOR AVOIDING DEGENERATE SOLUTIONS IN CANDECOMP/PARAFAC	82
--	----

Paolo Giordani

Big Data Analysis

Organizer and Chair: Donato Malerba

TOWARDS A STATISTICAL FRAMEWORK FOR ATTRIBUTE COMPARISON IN VERY
LARGE RELATIONAL DATABASES 83

Cesare Alippi, Elisa Quintarelli, Manuel Roveri and Letizia Tanca

MINING BIG DATA WITH HIGH PERFORMANCE COMPUTING SOLUTIONS 91

Fabrizio Angiulli, Stefano Basta, Stefano Lodi, Gianluca Moro and Claudio Sartori

ENHANCING BIG DATA EXPLORATION WITH FACETED BROWSING 95

Sonia Bergamaschi, Giovanni Simonini and Song Zhu

New Methodologies for Composite Indicators

Organizer and Chair: Agostino Di Ciaccio

ADVANCES IN COMPOSITE-BASED PATH MODELING FOR SYNTHETIC INDICATORS 100

Vincenzo Esposito Vinzi, Laura Trinchera and Giorgio Russolillo

COMPOSITE INDICATORS MODELING 102

Maurizio Vichi

MEASURING THE IMPORTANCE OF VARIABLES IN COMPOSITE INDICATORS . . . 104

William Becker, Michaela Saisana, Paolo Paruolo and Andrea Saltelli

Cluster analysis software and validation

Organizer and Chair: Christian Hennig

ADAPTIVE CHOICE OF INPUT PARAMETERS IN ROBUST CLUSTERING 109

Luis A. García-Escudero and Augustin Mayo-Iscar

ROBUST MODEL-BASED CLUSTERING WITH COVARIANCE MATRIX CONSTRAINTS 113

Pietro Coretto and Christian Hennig

FLEXIBLE IMPLEMENTATION OF RESAMPLING SCHEMES FOR CLUSTER VALIDATION 117

Friedrich Leisch

Selecting a mixture model with a clustering focus

Organizer and Chair: Gilles Celeux

CLUSTERING IN FINITE MIXTURES USING AN INTEGRATED COMPLETED LIKELI-
HOOD CRITERION 122

Marco Bertoletti, Nial Friel and Riccardo Rastelli

ESTIMATION AND MODEL SELECTION FOR MODEL-BASED CLUSTERING WITH THE
CONDITIONAL CLASSIFICATION LIKELIHOOD 126

Jean-Patrick Baudry

ON THE DIFFERENT WAYS TO COMPUTE THE INTEGRATED COMPLETED LIKELIHOOD CRITERION	130
<i>Gilles Celeux</i>	
Exploring relationships between blocks of variables	
<i>Organizer and Chair: Giorgio Russolillo</i>	
WEIGHTED MULTIBLOCK CLUSTERING	135
<i>Ndéye Niang and Mory Ouattara</i>	
THEMATIC MODEL EXPLORATION THROUGH MULTIPLE CO-STRUCTURE MAXIMISATION: METHOD AND SOFTWARE	139
<i>Xavier Bry and Thomas Verron</i>	
A NEW COMPONENT-BASED APPROACH OF REGULARISATION FOR MULTIVARIATE GENERALISED LINEAR REGRESSION	144
<i>Catherine Trottier, Xavier Bry, Frederic Mortier and Guillaume Cornu</i>	
Solicited Sessions	
Advances in Density-based clustering	
<i>Organizer and Chair: Francesca Greselin</i>	
A NONPARAMETRIC CLUSTERING METHOD FOR IMAGE SEGMENTATION	150
<i>Giovanna Menardi</i>	
ROBUST CLUSTERING FOR HETEROGENOUS SKEW DATA	154
<i>Luis A. García-Escudero, Francesca Greselin and Agustin Mayo-Iscar</i>	
REGULARIZING FINITE MIXTURES OF GAUSSIAN DISTRIBUTIONS	154
<i>Bettina Grün and Gertraud Malsiner-Walli</i>	
Latent variable models for longitudinal data - Part I	
<i>Organizer and Chair: Silvia Bacci</i>	
A JOINT MODEL FOR LONGITUDINAL AND SURVIVAL DATA BASED ON AN AR(1) LATENT PROCESS	163
<i>Silvia Bacci, Francesco Bartolucci and Silvia Pandolfi</i>	
FINITE MIXTURE MODELS FOR MIXED DATA: EM ALGORITHMS AND PARAFAC REPRESENTATIONS	167
<i>Marco Alfó and Paolo Giordani</i>	
ON THE USE OF THE CONTAMINATED GAUSSIAN DISTRIBUTION IN HIDDEN MARKOV MODELS FOR LONGITUDINAL DATA	171
<i>Antonio Punzo and Antonello Maruotti</i>	
Latent variable models for longitudinal data - Part II	
<i>Organizer and Chair: Francesco Bartolucci</i>	
A HIDDEN MARKOV APPROACH TO THE ANALYSIS OF INCOMPLETE MULTIVARIATE LONGITUDINAL DATA	177
<i>Francesco Lagona</i>	
LATENT MARKOV AND GROWTH MIXTURE MODELS: A COMPARISON	181

Fulvia Pennoni and Isabella Romeo

LATENT WORTHS AND LONGITUDINAL PAIRED COMPARISONS - A MARKOV MODEL OF DEPENDENCE	185
<i>Brian Francis, Alexandra Grand and Regina Dittrich</i>	

Multivariate data analysis in environmental sciences

Organizer: Fabrizio Ruggeri; Chair: Raffaele Argiento

MULTIVARIATE DOWNSCALING FOR NON-GAUSSIAN DATA	191
<i>Daniela Cocchi, Lucia Paci and Carlo Trivisano</i>	

PRELIMINARY RESULTS ON TAPERING MULTIVARIATE SPATIO TEMPORAL MODELS FOR EXPOSURE TO AIRBORNE MULTIPOLLUTANTS IN EUROPE	195
<i>Alessandro Fassó, Francesco Finazzi and Ferdinand Ndongo</i>	

CLUSTERING MACROSEISMIC FIELDS BY STATISTICAL DATA DEPTH FUNCTIONS	199
<i>Claudio Agostinelli, Renata Rotondi and Elisa Varini</i>	

Advanced models for tourism analysis

Organizer and Chair: Stefania Mignani

ANALYSING TERRITORIAL HETEROGENEITY IN TOURISTS'SATISFACTION TOWARDS ITALIAN DESTINATIONS	204
<i>Cristina Bernini, Augusto Cerqua and Guido Pellegrini</i>	

MICRO-ECONOMIC DETERMINANTS OF TOURIST EXPENDITURE: A QUANTILE REGRESSION APPROACH	208
<i>Emanuela Marrocu, Raffaele Paci and Andrea Zara</i>	

INEQUALITIES AND TOURISM CONSUMPTION BEHAVIOUR: A MIXTURE MODEL ANALYSIS	212
<i>Cristina Bernini, Maria Francesca Cracolici and Cinzia Viroli</i>	

Bayesian Networks and Graphical Models in Socio-Economic Sciences

Organizer and Chair: Paola Vicard

BAYESIAN NETWORKS FOR FIRM PERFORMANCE EVALUATION	217
<i>Maria E. De Giuli, Pietro Gottardo, Anna M. Moisello and Claudia Tarantola</i>	

GRAPHICAL MODEL USING COPULAS FOR MEASUREMENT ERROR MODELING . .	221
<i>Daniela Marella and Paola Vicard</i>	

Time Series in Clustering

Organizer and Chair: Michele La Rocca

PARSIMONIOUS CLUSTERING OF TIME SERIES	226
<i>Carmela Iorio, Antonio D'Ambrosio, Gianluca Frasso and Roberta Siciliano</i>	

DYNAMIC TIME WARPING-BASED FUZZY CLUSTERING FOR SPATIAL TIME SERIES	230
<i>Pierpaolo D'Urso, Marta Disegna and Riccardo Massari</i>	

PERIODICAL FEATURE BASED TIME SERIES CLUSTERING	234
<i>Francesco Giordano, Michele La Rocca and Maria Lucia Parrella</i>	

Big Data Analysis

Organizer and Chair: Donato Malerba

INTERACTIVE MACHINE LEARNING WITH R 239

Giorgio Maria Di Nunzio

WORKLOAD ESTIMATION FOR A CALL CENTER 243

Pierluigi Riva and Ruggiero Scommegna

PREDICTION IN OLIVE OIL TRADE USING REGRESSION MODELS ON TEMPORAL
DATA NETWORK 245

Corrado Loglisci , Umberto Medicamento and Arturo Casieri

Advances in Ordinal and Preference Data

Organizer and Chair: Antonio D'Ambrosio

MEASURING CONSENSUS IN THE SETTING OF NON-UNIFORM QUALITATIVE SCALES 250

José L. García-Lapresta and David Pérez-Román

ACCURATE ALGORITHMS FOR CONSENSUS RANKING DETECTION 255

Giulio Mazzeo, Antonio D'Ambrosio and Roberta Siciliano

LOGISTIC REGRESSION TREES FOR ORDINAL AND PREFERENCE DATA 259

Thomas Rusch, Achim Zeileis and Kurt Hornik

Case studies in data science from Ligurian companies

Organizer and Chair: Delio Panaro

STATISTICAL METHODS FOR THE ANALYSIS OF OSTREOPSIS OVATA BLOOM EVENTS
FROM METEO-MARINE DATA 262

*Ennio Ottaviani, Valentina Asnaghi, Mariachiara Chiantore, Andrea Pedroncini
and Rosella Bertolotto*

DATA MINING FOR OPTIMAL GAMBLING 266

Gabriele Torre and Fabrizio Malfanti

A FRAUD DETECTION ALGORITHM FOR ONLINE BANKING 271

Delio Panaro, Eva Riccomagno and Fabrizio Malfanti

DOES DIRECTORS' BACKGROUND MATTER? FIRM PERFORMANCE, BOARD FEAT-
URES AND FINANCIAL REPORTING RELIABILITY 275

Delio Panaro, Silvia Ferramosca and Sara Trucco

Modeling ordinal data

Organizer and Chair: Maurizio Carpita

POSTERIOR PREDICTIVE MODEL CHECKS FOR ASSESSING THE GOODNESS OF FIT
OF BAYESIAN MULTIDIMENSIONAL IRT MODELS 280

Mariagiulia Matteucci and Stefania Mignani

INTERNATIONAL TOURISM IN ITALY: A BAYESIAN NETWORK APPROACH 284

Federica Cugnata and Giovanni Perucca

CLUSTERING UPPER LEVEL UNITS IN MULTILEVEL MODELS FOR ORDINAL DATA 288

Leonardo Grilli, Agnese Panzera and Carla Rampichini

Functional data analysis for environmental data

Organizer and Chair: Tonio Di Battista

CLUSTERING SPATIALLY DEPENDENT FUNCTIONAL DATA: A METHOD BASED ON
THE CONCEPT OF SPATIAL DISPERSION FUNCTION OF A CURVE 292

Elvira Romano, Antonio Balzanella and Rosanna Verde

TWO CASE STUDIES ON OBJECT ORIENTED SPATIAL STATISTICS 296

Piercesare Secchi, Simone Vantini and Valeria Vitelli

INFERENCE ON FUNCTIONAL BIODIVERSITY TOOLS 298

Tonio Di Battista, Francesca Fortuna and Fabrizio Maturo

Advances in quantile regression

Organizer and Chair: Cristina Davino

M-QUANTILE REGRESSION: DIAGNOSTICS AND PARAMETRIC REPRESENTATION
OF THE MODEL 303

Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati and Nikos Tzavidis

QUANTILE REGRESSION: A BAYESIAN ROBUST APPROACH 307

Marco Bottone, Mauro Bernardi and Lea Petrella

A COMPARISON AMONG ESTIMATORS FOR LINEAR REGRESSION METHODS 311

Marilena Furno and Domenico Vistocco

HANDLING HETEROGENEITY AMONG UNITS IN QUANTILE REGRESSION 315

Cristina Davino and Domenico Vistocco

Directional Data

Organizer and Chair: Giovanni C. Porzio

SMALL BIASED CIRCULAR DENSITY ESTIMATION 320

Marco Di Marzio, Stefania Fensore, Agnese Panzera, and Charles C. Taylor

A DEPTH-BASED CLASSIFIER FOR CIRCULAR DATA 324

Giuseppe Pandolfo

NONPARAMETRIC ESTIMATES OF THE MODE FOR DIRECTIONAL DATA	328
<i>Thomas Kirschstein, Steffen Liebscher, Giovanni C. Porzio and Giancarlo Ragozini</i>	

Recent developments in statistical analysis of network data

Organizer and Chair: Domenico De Stefano

GAME THEORY AND NETWORK MODELS FOR THE RECONSTRUCTION OF AR- CHAEOLOGICAL NETWORKS	331
<i>Viviana Amati and Ulrik Brandes</i>	

A MODEL FOR CLUSTERING A SPATIAL NETWORK WITH APPLICATION TO LOCAL LABOUR SYSTEM IDENTIFICATION	335
<i>Francesco Pauli, Nicola Torelli and Susanna Zaccarin</i>	

ON THE SAMPLING DISTRIBUTIONS OF THE ML ESTIMATORS IN NETWORK EF- FECT MODELS	339
<i>Michele La Rocca, Giovanni C. Porzio, Maria Prosperina Vitale and Patrick Dor- eian</i>	

CORRESPONDENCE ANALYSIS WITH DOUBLING FOR TWO-MODE VALUED NET- WORKS	343
<i>Giancarlo Ragozini, Domenico De Stefano and Daniela D'Ambrosio</i>	

Current challenges in clustering and classification of biomedical data

Organizer and Chair: Adalbert F.X. Wilhelm

SEMANTIC MULTI CLASSIFIER SYSTEMS FOR THE DETECTION OF AGING RELATED PROCESSES	348
<i>Hans A. Kestler, Ludwig Lausser, Lyn-Rouven Schirra, Florian Schmid</i>	

EMOTION RECOGNITION IN HUMAN COMPUTER INTERACTION USING MULTIPLE CLASSIFIER SYSTEMS	349
<i>Friedhelm Schwenker</i>	

ENSEMBLE OF SELECTED CLASSIFIERS	352
<i>Berthold Lausen, Asma Gul, Zardad Khan and Osama Mahmoud</i>	

Contributed papers

A GENERALIZED DISTANCE FOR INFERENCE IN FUNCTIONAL DATA	354
<i>Andrea Ghiglietti and Anna M. Paganoni</i>	

LONG GAPS IN MULTIVARIATE SPATIO-TEMPORAL DATA: AN APPROACH BASED ON FUNCTIONAL DATA ANALYSIS	359
<i>Mariantonietta Ruggieri, Antonella Plaia and Francesca Di Salvo</i>	

EFFECTS ON CURVE CLUSTERING OF DIFFERENT TRANSFORMATIONS OF CHRONO- LOGICAL TEXTUAL DATA	363
<i>Matilde Trevisani and Arjuna Tuzzi</i>	

A NOTE ON THE RELIABILITY OF A CLASSIFIER	366
<i>Luca Frigau</i>	

ROBUSTIFIED CLASSIFICATION OF MULTIVARIATE FUNCTIONAL DATA	370
<i>Francesca Ieva and Anna M. Paganoni</i>	
SIZE CONTROL OF ROBUST REGRESSION ESTIMATORS	374
<i>Silvia Salini, Andrea Cerioli, Fabrizio Laurini and Marco Riani</i>	
THE MOVEMENTS OF EMOTIONS: AN EXPLORATORY CLASSIFICATION ON AF- FECTIVE MOVEMENT DATA	378
<i>Pasquale Dente, Arvid Kappas and Adalbert F. X. Wilhelm</i>	
ELECTRE TRI-MACHINE LEARNING APPROACH TO THE RECORD LINKAGE PROBLEM	382
<i>Valentina Minnetti and Renato De Leone</i>	
QUALITY OF CLASSIFICATION APPROACHES FOR THE QUANTITATIVE ANALYSIS OF INTERNATIONAL CONFLICT	387
<i>Adalbert F.X. Wilhelm</i>	
THE RTCLUST PROCEDURE FOR ROBUST CLUSTERING	391
<i>Francesco Dotto, Alessio Farcomeni, Luis Angel García-Escudero and Agustín Mayo- Iscar</i>	
WHAT ARE THE TRUE CLUSTERS?	396
<i>Christian Hennig</i>	
A NOVEL MODEL-BASED CLUSTERING APPROACH FOR MASSIVE DATASETS OF SPATIALLY REGISTERED TIME SERIES. WITH APPLICATION TO SEA SURFACE TEMPERATURE REMOTE SENSING DATA	399
<i>Francesco Finazzi and Marian Scott</i>	
BIG DATA CLASSIFICATION: SIMULATIONS IN THE MANY FEATURES CASE	403
<i>Claus Weihs</i>	
FROM BIG DATA TO INFORMATION: STATISTICAL ISSUES THROUGH EXAMPLES .	407
<i>Silvia Biffignandi and Serena Signorelli</i>	
BIG DATA MEET PHARMACEUTICAL INDUSTRY: AN APPLICATION ON SOCIAL MEDIA DATA	411
<i>Caterina Liberati and Paolo Mariani</i>	
DEFINING THE SUBJECTS DISTANCE IN HIERARCHICAL CLUSTER ANALYSIS BY COPULA APPROACH	416
<i>Andrea Bonanomi, Marta Nai Ruscone and Silvia Angela Osmetti</i>	

SUPERVISED CLASSIFICATION OF DEFECTIVE CRANKSHAFTS BY IMAGE ANALYSIS	420
<i>Beatriz Remeseiro, Javier Tarrío-Saavedra, Mario Francisco-Fernández, Manuel G. Penedo, Salvador Naya and Ricardo Cao</i>	
ARCHETYPAL ANALYSIS FOR DATA-DRIVEN PROTOTYPE IDENTIFICATION	424
<i>Giancarlo Ragozini, Francesco Palumbo and Maria R. D'Esposito</i>	
PRINCIPAL COMPONENT ANALYSIS OF COMPLEX DATA AND APPLICATION TO CLIMATOLOGY	428
<i>Sergio Camiz and Silvia Creta</i>	
SPARSE EXPLORATORY MULTIDIMENSIONAL IRT MODELS	432
<i>Lara Fontanella, Sara Fontanella, Pasquale Valentini, Nickolay Trendafilov</i>	
ITERATIVE FACTOR CLUSTERING FOR CATEGORICAL DATA RECONSIDERED	437
<i>Alfonso Iodice D'Enza, Angelos Markos and Francesco Palumbo</i>	
TESTING ANTIPODAL SYMMETRY OF CIRCULAR DATA	442
<i>Giovanni Casale, Giuseppe Pandolfo and Giovanni C. Porzio</i>	
HOW TO DEFINE DEVIANCE RESIDUALS IN MULTINOMIAL REGRESSION	446
<i>Giovanni Romeo, Mariangela Sciandra and Marcello Chiodi</i>	
DIAGNOSTIC TOOLS FOR GAMLSS FITTED OBJECTS	451
<i>Andrea Marletta and Mariangela Sciandra</i>	
BAYESIAN REGRESSION ANALYSIS WITH LINKED AND DUPLICATED DATA	455
<i>Andrea Tancredi, Rebecca Steorts and Brunero Liseo</i>	
A SEMI-PARAMETRIC FAY-HERRIOT-TYPE MODEL WITH UNKNOWN SAMPLING VARIANCES	460
<i>Silvia Polettini</i>	
POSTERIOR DISTRIBUTIONS FROM OPTIMALLY B-ROBUST ESTIMATING FUNCTIONS AND APPROXIMATE BAYESIAN COMPUTATION	464
<i>Ivan Luciano Danesi, Fabio Piacenza, Erlis Ruli and Laura Ventura</i>	
MCA BASED COMMUNITY DETECTION	468
<i>Carlo Drago</i>	
CLASSIFYING SOCIAL ROLES BY NETWORK STRUCTURES	472
<i>Simona Gozzo and Venera Tomaselli</i>	
A MULTILEVEL HECKMAN MODEL TO INVESTIGATE FINANCIAL ASSETS AMONG OLD PEOPLE IN EUROPE	476
<i>Omar Paccagnella and Chiara Dal Bianco</i>	
OPTIMAL PRICING USING BAYESIAN SEMIPARAMETRIC PRICE RESPONSE MODELS	480
<i>Winfried J. Steiner, Anett Weber, Stefan Lang and Peter Wechselberger</i>	

MONETARY TRANSMISSION MODELS FOR BANKING INTEREST RATES	484
<i>Laura Parisi, Paolo Giudici, Igor Gianfrancesco and Camillo Giliberto</i>	
ESTIMATING THE EFFECT OF PRENATAL CARE ON BIRTH OUTCOMES	490
<i>Emiliano Sironi and Massimo Cannas</i>	
RECURSIVE PARTITIONING: AN APPROACH BASED ON THE WEIGHTED KEMENY DISTANCE	494
<i>Mariangela Sciandra, Antonella Plaia and Veronica Picone</i>	
WHY TO STUDY ABROAD? AN EXAMPLE OF CLUSTERING	498
<i>Valeria Caviezel and Anna M. Falzoni</i>	
A GRAPHICAL COPULA-BASED TOOL FOR DETECTING TAIL DEPENDENCE	502
<i>Roberta Pappadà, Fabrizio Durante and Nicola Torelli</i>	
CLASSIFICATION MODELS AS TOOLS OF BANKRUPTCY PREDICTION - POLISH EX- PERIENCE	506
<i>Józef Pocięcha, Barbara Pawełek, Mateusz Baryła and Sabina Augustyn</i>	
THE RELATIONSHIP BETWEEN INDIVIDUAL PRICE RESPONSE OF BEER CON- SUMERS AND THEIR DEMOGRAPHIC/PSYCHOGRAPHIC CHARACTERISTICS	510
<i>Friederike Paetz</i>	
THE ENSEMBLE CONCEPTUAL CLUSTERING OF SYMBOLIC DATA FOR CUSTOMER LOYALTY ANALYSIS	514
<i>Marcin Pełka</i>	
INSERT HERE CONSUMERS' PERCEPTIONS OF CORPORATE SOCIAL RESPONSIBIL- ITIES AND WILLINGNESS TO PAY: A PARTIAL LEAST SQUARES	519
<i>Karsten Lübke, Christian Hose and Thomas Obermeier</i>	
INSPECTING THE QUALITY OF ITALIAN WINE THROUGH CAUSAL REASONING	521
<i>Eugenio Brentari, Maurizio Carpita and Silvia Golia</i>	
EXPLORING SOCIO-ECONOMIC FACTORS ASSOCIATED WITH ADHERENCE TO THE MEDITERRANEAN DIET: A MULTILEVEL APPROACH	525
<i>Tiziana Laureti and Luca Secondi</i>	
BIG DATA AND 'SOCIAL' REPUTATION: A FINANCIAL EXAMPLE	529
<i>Paola Cerchiello</i>	
BAYESIAN NETWORKS FOR STOCK PICKING	533
<i>Alessandro Greppi, Maria Elena De Giuli and Claudia Tarantola</i>	
PORTFOLIO SELECTION WITH LASSO ALGORITHM	537
<i>Riccardo Bramante, Silvia Facchinetti and Diego Zappa</i>	
SUNSPOT IN ECONOMIC MODELS WITH EXTERNALITIES	540
<i>Beatrice Venturi and Alessandro Pirisinu</i>	

ROBUST CLUSTERING FOR HETEROGENOUS SKEW DATA

Luis Angel García-Escudero,¹ Francesca Greselin² and Agustin Mayo-Iscar¹

¹ Department of Statistics and Operations Research and IMUVA, University of Valladolid (e-mail: lagarcia@eio.uva.es, agustinm@eio.uva.es)

² Department of Statistics and Quantitative Methods, Milano-Bicocca University (e-mail: francesca.greselin@unimib.it)

ABSTRACT: The existing robust methods for model-based classification and clustering deal with elliptically contoured components. Here we introduce robust estimation for mixtures of skew-normal, by the joint usage of trimming and constraints. The model allows to fit heterogeneous skew data with great flexibility.

KEYWORDS: Clustering, robust estimation, skew data.

1 Introduction

In recent years, empirical evidences of asymmetric departures from normality in some real subpopulations suggested the introduction of skewed components in the classical mixture model approach. Therefore, increasing attention has been devoted to mixtures of multivariate skew normal and skew t , as well as to mixtures of normal-inverse-Gaussian distributions, shifted asymmetric Laplace distributions, generalized hyperbolic distributions and hierarchical mixtures (mixtures of mixtures), to capture asymmetric shapes in components. Most of these parametric families of skew distributions are clearly related, as it has been described in a careful review by Lee & McLachlan (2013a). In this paper we address robust estimation of mixtures of skew normal distributions. Our interest for the Finite Mixtures of Canonical Fundamental Skew Normal (FM CFUSN) has been motivated by the aim of adapting to the Gaussian case the methodology developed for Finite Mixtures of Canonical Fundamental Skew t (FM CFUST) in Lee & McLachlan (2014), based on skew distributions originated in Arellano-Valle & Genton (2005). By mimicking the CFUST properties, CFUSN has the special appealing of including as particular cases the restricted and unrestricted Skew Normal, respectively rMSN and uMSN. The great flexibility of CFUSN is due to its parameters, which allow to separately govern location, scale, correlation, and skewness. We will show,

herein, that the FM CFUSN offers a very tractable model for many situations of departures from symmetry, being analytically and computationally easier than the FM CFUST. Taking advantage of its robust estimation, the model becomes even more adaptable to real phenomena. Indeed, robustness is often required by many datasets available nowadays, where populations are noisy, have some non-symmetric features, and could contain outliers asymmetrically disposed around the data clusters.

The first step to achieve robustness and obtain good breakdown properties for the estimators, is based on a *trimming procedure* incorporated along the iterations of the EM algorithm. The key idea is that a small portion of observations, which are highly unlikely to occur under the current fitted model assumption, is discarded from contributing to the parameter estimates. The same methodology has been proven to be very effective when addressing robust clustering for Gaussian and t mixtures models (see Neykov *et al.*, 2007; Gallegos & Ritter, 2009; García-Escudero *et al.*, 2014). Furthermore - and this is the second step - we implement a *constrained ML estimation* for the component covariances, aiming at reducing spurious solutions and preventing singularities of the likelihood, along the lines of Ingrassia & Rocci (2007).

Monte Carlo experiments show that bias and MSE of the estimator in several cases of contaminated data are dramatically inflated, while they return to be comparable to results obtained on skew data without noise, when the combined effect of trimming and constrained estimation is applied. Further, it can be shown that the estimator resists the influence of all classes of contaminating observations, as far as the outliers appear in a proportion lower than the trimming level α . As a final remark, we want to stress that the joint usage of trimming and constrained estimation allow to set the underlying mathematical and statistical problems as well posed ones.

2 Methodology

We consider the ML estimation for a g -component mixture model in which a random sample $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ follows a mixture of CFUSNs. The probability density function can be written as

$$\mathbf{Y}_j \sim \sum_{i=1}^g \pi_i f(\mathbf{y}_j | \mu_i, \Sigma_i, \Delta_i), \quad \pi_i \geq 0, \quad \sum_{i=1}^g \pi_i = 1,$$

where $f(\cdot)$ denotes a CFUSN density, $\Theta = (\theta_1, \dots, \theta_g)$ with $\theta_i = (\pi_i, \mu_i, \Sigma_i, \Delta_i)$ are the unknown parameters of component i , and π_i are the weights of the groups.

To define the location-scale variant of the CFUSN distribution, let $(\mathbf{Y}_0, \mathbf{Y}_1)$ be a $p + q$ multivariate normal r.v., such that

$$\begin{bmatrix} \mathbf{Y}_0 \\ \mathbf{Y}_1 \end{bmatrix} \sim \mathcal{N}_{q+p} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \Sigma \end{bmatrix} \right) \quad (1)$$

where Σ is a positive definite scale matrix and $\mathbf{0}$ is a vector of zeros with appropriate dimension. Then, given a $p \times q$ matrix Δ , we obtain the stochastic representation for \mathbf{Y} , i.e. $\mathbf{Y} = \mu + \Delta|\mathbf{Y}_0| + \mathbf{Y}_1$, that follows the CFUSN distribution with density given by

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

where $\Omega = \Sigma + \Delta\Delta^T$, $\Lambda = \mathbf{I}_p - \Delta^T \Omega^{-1} \Delta$, and 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.

In the ML approach, we consider the following *trimmed* log-likelihood function (see Neykov *et al.*, 2007; Gallegos & Ritter, 2009; García-Escudero *et al.*, 2014)

$$\mathcal{L}_{trim} = \sum_{i=1}^n z(\mathbf{y}_i) \log \left[\sum_{g=1}^G \phi_p(\mathbf{y}_i; \mu_g, \Omega_g) \Phi_q(\Delta_g^T \Omega_g^{-1}(\mathbf{y}_i - \mu_g); \mathbf{0}, \Lambda_g) \pi_g \right]. \quad (2)$$

By $z(\cdot)$ we denote a 0-1 trimming indicator function that indicates whether observation \mathbf{y}_i is trimmed off: $z(\mathbf{y}_i)=0$, or not: $z(\mathbf{y}_i)=1$. A fixed fraction α of observations, whose contributions to the likelihood are lower than their α -quantile, can be unassigned by setting $\sum_{i=1}^n z(\mathbf{y}_i) = [n(1 - \alpha)]$ at each E-step, hence they do not contribute to the parameter estimation.

Further, we want to deal with the unboundedness of the target function \mathcal{L}_{trim} when no constraints are imposed on the scatter parameters. In this case, the defining problem is ill-posed because the log-likelihood tends to ∞ when either $\mu_g = \mathbf{y}_i$ and $|\Sigma_g| \rightarrow 0$. As a trivial consequence, the EM algorithm can be trapped into non-interesting local maximizers, called “spurious” solutions, and the result of the EM algorithm strongly depends on its initialization. For this reason, we set a constraint on the maximization in (2), concerning the eigenvalues $\{\lambda_l(\Sigma_g)\}_{l=1, \dots, d}$ of the scatter matrices Σ_g , by imposing

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

as in Ingrassia & Rocci (2007); García-Escudero *et al.* (2008); Gallegos & Ritter (2009). It is important to remark that the parameters related with location

and skewness are not affected by the constrained maximization applied to the covariance matrices.

As usual, among many runs of the EM algorithm, we select the parameters given by the best final likelihood. To complete a synthetic description of the algorithm, our proposal for the initialization of the EM is to take a small random subsample for each component and to draw, for each observation in it, a random variate \mathbf{Y}_0 from $\mathcal{N}_p(0, \mathbf{I}_p)$. By applying multivariate regression on the selected observations in each component, and using the corresponding vectors $|\mathbf{Y}_0|$, we get an estimation of the model parameters (i.e. μ_g , Σ_g and Δ_g) for the g -th component. Finally, initial estimations for the group weights are drawn from a multinomial random variable.

References

- ARELLANO-VALLE, R.B., & GENTON, M.G. 2005. On fundamental skew distributions. *Journal of Multivariate Analysis*, **96**(1), 93–116.
- FRITZ, H., GARCIA-ESCUADERO, L.A., & MAYO-ISCAR, A. 2012. tclust: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, **47**(12).
- GALLEGOS, M.T., & RITTER, G. 2009. Trimmed ML estimation of contaminated mixtures. *Sankhya (Ser. A)*, 164–220.
- GARCÍA-ESCUADERO, L.A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 1324–1345.
- GARCÍA-ESCUADERO, L.A., GORDALIZA, A., & MAYO-ISCAR, A. 2014. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**(1), 27–43.
- INGRASSIA, S., & ROCCI, R. 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, **51**, 5339–5351.
- LEE, S., & MCLACHLAN, G. 2013a. Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications*, **22**(4), 427–454.
- LEE, S., & MCLACHLAN, G. 2013b. On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, **7**(3), 241–266.
- LEE, S., & MCLACHLAN, G. 2014. Finite Mixtures of Canonical Fundamental Skew t-Distributions. *arXiv preprint arXiv:1405.0685*.
- NEYKOV, N., FILZMOSER, P., DIMOVA, R., & NEYTCHEV, P. 2007. Robust Fitting of Mixtures Using the Trimmed Likelihood Estimator. *Computational Statistics & Data Analysis*, **52**(1), 299–308.