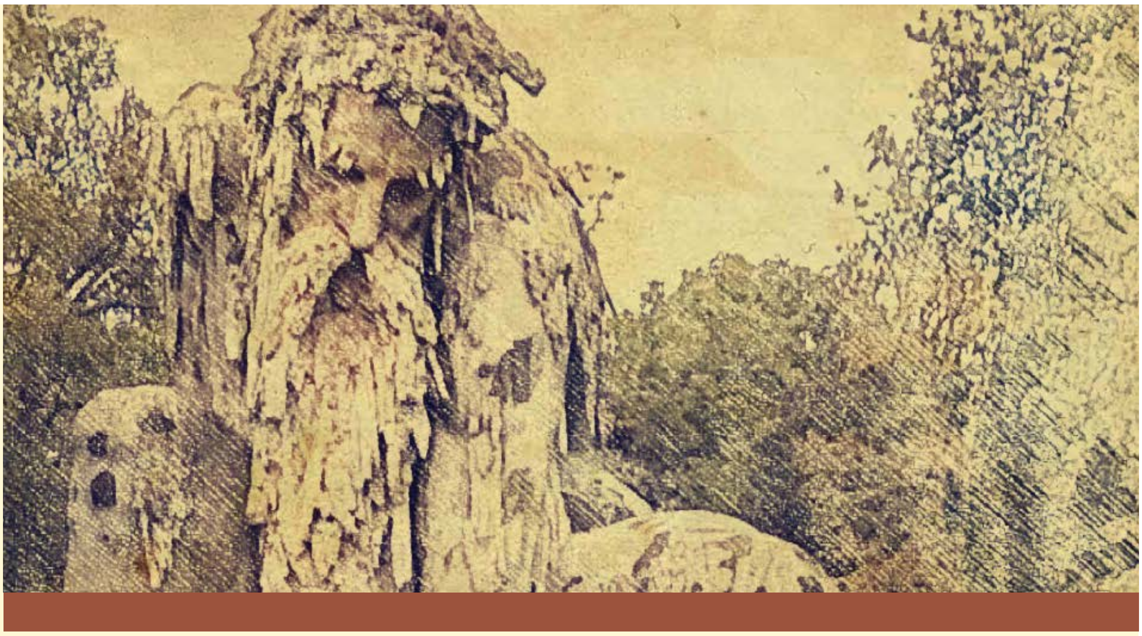


CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS

PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

– 128 –

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS
AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group

Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

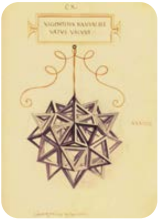
FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard



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

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

Firenze University Press Editorial Board
M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

The online digital edition is published in Open Access on www.fupress.com.

Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: http://creativecommons.org/licenses/by/4.0/legalcode). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.

Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: https://creativecommons.org/publicdomain/zero/1.0/legalcode).

© 2021 Author(s)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

INDEX

Preface 1

Keynote Speakers

Jean-Michel Loubes
Optimal transport methods for fairness in machine learning 5

Peter Rousseeuw, Jakob Raymaekers and Mia Hubert
Class maps for visualizing classification results 6

Robert Tibshirani, Stephen Bates and Trevor Hastie
Understanding cross-validation and prediction error 7

Cinzia Viroli
Quantile-based classification 8

Bin Yu
Veridical data science for responsible AI: characterizing V4 neurons through deepTune 9

Plenary Session

Daniel Diaz
A simple correction for COVID-19 sampling bias 14

Jeffrey S. Morris
A seat at the table: the key role of biostatistics and data science in the COVID-19 pandemic 15

Bhramar Mukherjee
Predictions, role of interventions and the crisis of virus in India: a data science call to arms 16

Danny Pfeffermann
Contributions of Israel's CBS to rout COVID-19 17

Invited Papers

Claudio Agostinelli, Giovanni Saraceno and Luca Greco
Robust issues in estimating modes for multivariate torus data 21

Emanuele Aliverti
Bayesian nonparametric dynamic modeling of psychological traits 25

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Andres M. Alonso, Carolina Gamboa and Daniel Peña
Clustering financial time series using generalized cross correlations **27**

Raffaele Argiento, Edoardo Filippi-Mazzola and Lucia Paci
Model-based clustering for categorical data via Hamming distance **31**

Antonio Balzanella, Antonio Irpino and Francisco de A.T. De Carvalho
Mining multiple time sequences through co-clustering algorithms for distributional data **32**

Francesco Bartolucci, Fulvia Pennoni and Federico Cortese
Hidden Markov and regime switching copula models for state allocation in multiple time-series **36**

Michela Battauz and Paolo Vidoni
Boosting multidimensional IRT models **40**

Matteo Bottai
Understanding and estimating conditional parametric quantile models **44**

Niklas Bussmann, Roman Enzmann, Paolo Giudici and Emanuela Raffinetti
Shapley Lorenz methods for eXplainable artificial intelligence **45**

Andrea Cappozzo, Ludovic Duponchel, Francesca Greselin and Brendan Murphy
Robust classification of spectroscopic data in agri-food: first analysis on the stability of results **49**

Andrea Cerasa, Enrico Checchi, Domenico Perrotta and Francesca Torti
Issues in monitoring the EU trade of critical COVID-19 commodities **53**

Marcello Chiodi
Smoothed non linear PCA for multivariate data **54**

Roberto Colombi, Sabrina Giordano and Maria Kateri
Accounting for response behavior in longitudinal rating data **58**

Claudio Conversano, Giulia Contu, Luca Frigau and Carmela Cappelli
Netwok-based semi-supervised clustering of time series data **62**

Federica Cugnata, Chiara Brombin, Pietro Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia Di Serio
Characterising longitudinal trajectories of COVID-19 biomarkers within a latent class framework **64**

Silvia D'Angelo
Sender and receiver effects in latent space models for multiplex data **68**

Anna Denkowska and Stanisław Wanat
DTW-based assessment of the predictive power of the copula-DCC-GARCH-MST model developed for European insurance institutions **71**

Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha
Two-step estimation of multilevel latent class models with covariates **75**

Marie Du Roy de Chaumaray and Matthieu Marbac
Clustering data with non-ignorable missingness using semi-parametric mixture models **79**

Pierpaolo D'Urso, Livia De Giovanni and Vincenzina Vitale
Spatial-temporal clustering based on B-splines: robust models with applications to COVID-19 pandemic **83**

Leonardo Egidi, Roberta Pappadà, Francesco Pauli and Nicola Torelli
PIVMET: pivotal methods for Bayesian relabelling in finite mixture models **87**

Tahir Ekin and Claudio Conversano
Cluster validity by random forests **91**

Luis Angel García-Escudero, Agustín Mayo-Iscar and Marco Riani
Robust estimation of parsimonious finite mixture of Gaussian models **92**

Silvia Facchinetti and Silvia Angela Osmetti
A risk indicator for categorical data **93**

Matteo Fasiolo
Additive quantile regression via the qgam R package **97**

Michael Fop, Dimitris Karlis, Ioannis Kosmidis, Adrian O'Hagan, Caitriona Ryan and Isobel Claire Gormley
Gaussian mixture models for high dimensional data using composite likelihood **98**

Carlo Gaetan, Paolo Girardi and Victor Muthama Musau
On model-based clustering using quantile regression **102**

Carlotta Galeone
Socioeconomic inequalities and cancer risk: myth or reality? **106**

Michael Gallaugher, Christophe Biernacki and Paul McNicholas
Parameter-wise co-clustering for high dimensional data **107**

Francesca Greselin and Alina Jędrzejczak
Quantifying the impact of covariates on the gender gap measurement: an analysis based on EU-SILC data from Poland and Italy **108**

Alessandra Guglielmi, Mario Beraha, Matteo Giannella, Matteo Pegoraro and Riccardo Peli
A transdimensional MCMC sampler for spatially dependent mixture models **112**

Christian Hennig and Pietro Coretto
Non-parametric consistency for the Gaussian mixture maximum likelihood estimator **116**

Yinxuan Huang and Natalie Shlomo
Improving the reliability of a nonprobability web survey **120**

Maria Iannario and Claudia Tarantola
A semi-Bayesian approach for the analysis of scale effects in ordinal regression models **124**

Jayant Jha
Best approach direction for spherical random variables **128**

Maria Kateri
Simple effect measures for interpreting generalized binary regression models **129**

Shogo Kato, Kota Nagasaki and Wataru Nakanishi
Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data **133**

John Kent
How to design a directional distribution **137**

Simona Korenjak-Černe and Nataša Kejžar
Identifying mortality patterns of main causes of death among young EU population using SDA approaches **141**

Fabrizio Laurini and Gianluca Morelli
Robust supervised clustering: some practical issues **142**

Daniela Marella and Danny Pfeffermann
A nonparametric approach for statistical matching under informative sampling and nonresponse **146**

Mariagiulia Matteucci and Stefania Mignani
Investigating model fit in item response models with the Hellinger distance **150**

Matteo Mazziotta and Adriano Pareto
PCA-based composite indices and measurement model **154**

Marcella Mazzoleni, Angiola Pollastri and Vanda Tulli
Gender inequalities from an income perspective **158**

Yana Melnykov, Xuwen Zhu and Volodymyr Melnykov
Transformation mixture modeling for skewed data groups with heavy tails and scatter **162**

Luca Merlo, Lea Petrella and Nikos Tzavidis
Unconditional M-quantile regression **163**

Jesper Møller , Mario Beraha, Raffaele Argiento and Alessandra Guglielmi
MCMC computations for Bayesian mixture models using repulsive point processes **167**

Keefe Murphy, Cinzia Viroli and Isobel Claire Gormley
Infinite mixtures of infinite factor analysers **168**

Stanislav Nagy, Petra Laketa and Rainer Dyckerhoff
Angular halfspace depth: computation **169**

Yarema Okhrin, Gazi Salah Uddin and Muhammad Yahya
Nonlinear Interconnectedness of crude oil and financial markets **173**

M. Rosário Oliveira, Ana Subtil and Lina Oliveira
Detection of internet attacks with histogram principal component analysis **174**

Sally Paganin
Semiparametric IRT models for non-normal latent traits **178**

Giuseppe Pandolfo
A graphical depth-based aid to detect deviation from unimodality on hyperspheres 182

Panos Pardalos
Networks of networks 186

Xanthi Pedeli and Cristiano Varin
Pairwise likelihood estimation of latent autoregressive count models 187

Mark Reiser and Maduranga Dassanayake
A study of lack-of-fit diagnostics for models fit to cross-classified binary variables 191

Giorgia Rivieccio, Jean-Paul Chavas, Giovanni De Luca, Salvatore Di Falco and Fabian Capitanio
Assessing food security issues in Italy: a quantile copula approach 195

Nicoleta Rogovschi
Co-clustering for high dimensional sparse data 199

Massimiliano Russo
Malaria risk detection via mixed membership models 203

Paula Saavedra-Nieves and Rosa M. Crujeiras
Nonparametric estimation of the number of clusters for directional data 207

Shuchismita Sarkar, Volodymyr Melnykov and Xuwen Zhu
Tensor-variate finite mixture model for the analysis of university professor remuneration 208

Florian Schuberth
Specifying composites in structural equation modeling: the Henseler-Ogasawara specification 209

Jarod Smith, Mohammad Arashi and Andriette Bekker
Network analysis implementing a mixture distribution from Bayesian viewpoint 210

Paul Smith, Peter van der Heijden and Maarten Cruyff
Measurement errors in multiple systems estimation 211

Valentin Todorov and Peter Filzmoser
Robust classification in high dimensions using regularized covariance estimates 215

Salvatore Daniele Tomarchio, Luca Bagnato and Antonio Punzo
Clustering via new parsimonious mixtures of heavy tailed distributions 216

Agostino Torti, Marta Galvani, Alessandra Menafoglio, Piercesare Secchi and Simone Vantini
A general bi-clustering technique for functional data 217

Laura Trinchera
Developing a multidimensional and hierarchical index following a composite-based approach 220

Rosanna Verde, Francisco T. de A. De Carvalho and Antonio Balzanella
A generalised clusteriwise regression for distributional data 223

Marika Vezzoli, Francesco Doglietto, Stefano Renzetti, Marco Fontanella and Stefano Calza
A machine learning approach for evaluating anxiety in neurosurgical patients during the COVID-19 pandemic 227

Isadora Antoniano Villalobos, Simone Padoan and Boris Beranger
Prediction of large observations via Bayesian inference for extreme-value theory 231

Maria Prosperina Vitale, Vincenzo Giuseppe Genova, Giuseppe Giordano and Giancarlo Ragozini
Community detection in tripartite networks of university student mobility flows 232

Ernst Wit and Lucas Kania
Causal regularization 236

Qiuyi Wu and David Banks
Minimizing conflicts of interest: optimizing the JSM program 240

Contributed Papers

Antonino Abbruzzo, Maria Francesca Cracolici and Furio Urso
Model selection procedure for mixture hidden Markov models 243

Roberto Ascari and Sonia Migliorati
A full mixture of experts model to classify constrained data 247

Luigi Augugliaro, Gianluca Sottile and Angelo Mineo
Sparse inference in covariate adjusted censored Gaussian graphical models 251

Simona Balzano, Mario Rosario Guarracino and Giovanni Camillo Porzio
Semi-supervised learning through depth functions 255

Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta
A combined test of the Benford hypothesis with anti-fraud applications 256

Chiara Bardelli
Unabalanced classfication of electronic invoicing 260

Claudia Berloco, Raffaele Argiento and Silvia Montagna
Predictive power of Bayesian CAR models on scale free networks: an application for credit risk 264

Marco Berrettini, Giuliano Galimberti and Saverio Ranciati
Semiparametric finite mixture of regression models with Bayesian P-splines 268

Giuseppe Bove
A subject-specific measure of interrater agreement based on the homogeneity index **272**

Antonio Calcagni
Estimating latent linear correlations from fuzzy contingency tables **276**

Andrea Cappozzo, Alessandro Casa and Michael Fop
Model-based clustering with sparse matrix mixture models **280**

Andrea Cappozzo, Luis Angel Garcìa Escudero, Francesca Greselin and Agustìn Mayo-Iscar
Exploring solutions via monitoring for cluster weighted robust models **284**

Maurizio Carpita and Silvia Golia
Categorical classifiers in multi-class classification problems **288**

Gianmarco Caruso, Greta Panunzi, Marco Mingione, Pierfrancesco Alaimo Di Loro, Stefano Moro, Edoardo Bompiani, Caterina Lanfredi, Daniela Silvia Pace, Luca Tardella and Giovanna Jona Lasinio
Model-based clustering for estimating cetaceans site-fidelity and abundance **292**

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria
Model-based clustering with parsimonious covariance structure **296**

Francesca Condino
Clustering income data based on share densities **300**

Paula Costa Fontichiari, Miriam Giuliani, Raffaele Argiento and Lucia Paci
Group-dependent finite mixture model **304**

Salvatore Cuomo, Federico Gatta, Fabio Giampaolo, Carmela Iorio and Francesco Piccialli
A machine learning approach in stock risk management **308**

Cristina Davino and Giuseppe Lamberti
Pathmox segmentation trees to compare linear regression models **312**

Houyem Demni, Davide Buttarazzi, Stanislav Nagy and Giovanni Camillo Porzio
Angular halfspace depth: classification using spherical bagdistances **316**

Agostino Di Ciaccio
Neural networks for high cardinality categorical data **320**

F. Marta L. Di Lascio, Andrea Menapace and Roberta Pappadà
Ali-Mikhail-Haq copula to detect low correlations in hierarchical clustering **324**

Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci
Higher education and employability: insights from the mandatory notices of the ministry of labour **328**

Lorenzo Focardi Olmi and Anna Gottard
An alternative to joint graphical lasso for learning multiple Gaussian graphical models **332**

Francesca Fortuna, Alessia Naccarato and Silvia Terzi
Functional cluster analysis of HDI evolution in European countries 336

Sylvia Frühwirth-Schnatter, Bettina Grün and Gertraud Malsiner-Walli
Estimating Bayesian mixtures of finite mixtures with telescoping sampling 340

Chiara Galimberti, Federico Castelletti and Stefano Peluso
A Bayesian framework for structural learning of mixed graphical models 344

Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu
Measurement error models on spatial network lattices: car crashes in Leeds 348

Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, Massimo Aria and Roberta Siciliano
The L^p data depth and its application to multivariate process control charts 352

Petra Laketa and Stanislav Nagy
Angular halfspace depth: central regions 356

Michele La Rocca, Francesco Giordano and Cira Perna
Clustering production indexes for construction with forecast distributions 360

Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli
Clustering longitudinal data with category theory for diabetic kidney disease 364

Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore
A redundancy analysis with multivariate random-coefficients linear models 368

Paolo Mariani, Andrea Marletta and Matteo Locci
The use of multiple imputation techniques for social media data 372

Federico Marotta, Paolo Provero and Silvia Montagna
Prediction of gene expression from transcription factors affinities: an application of Bayesian non-linear modelling 376

Francesca Martella, Fabio Attorre, Michele De Sanctis and Giuliano Fanelli
High dimensional model-based clustering of European georeferenced vegetation plots 380

Ana Martins, Paula Brito, Sónia Dias and Peter Filzmoser
Multivariate outlier detection for histogram-valued variables 384

Giovanna Menardi and Federico Ferraccioli
A nonparametric test for mode significance 388

Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia and Arnaud Sallaberry
Visualizing cluster of words: a graphical approach to grammar acquisition 392

Marta Nai Ruscone and Dimitris Karlis
Robustness methods for modelling count data with general dependence structures **396**

Roberta Paroli, Luigi Spezia, Marc Stutter and Andy Vinten
Bayesian analysis of a water quality high-frequency time series through Markov switching autoregressive models **400**

Mariano Porcu, Isabella Sulis and Cristian Usala
Detecting the effect of secondary school in higher education university choices **404**

Roberto Rocci and Monia Ranalli
Semi-constrained model-based clustering of mixed-type data using a composite likelihood approach **408**

Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Damiana Pieragostino
Antibodies to SARS-CoV-2: an exploratory analysis carried out through the Bayesian profile regression **412**

Theresa Scharl and Bettina Grün
Modelling three-way RNA sequencing data with mixture of multivariate Poisson-lognormal distribution **416**

Luca Scrucca
Stacking ensemble of Gaussian mixtures **420**

Rosaria Simone, Cristina Davino, Domenico Vistocco and Gerhard Tutz
A robust quantile approach to ordinal trees **424**

Venera Tomaselli, Giulio Giacomo Cantone and Valeria Mazzeo
The detection of spam behaviour in review bomb **428**

Donatella Vicari and Paolo Giordani
Clustering models for three-way data **432**

Gianpaolo Zammarchi and Jaromir Antoch
Using eye-traking data to create a weighted dictionary for sentiment analysis: the eye dictionary **436**

Preface

This book collects the abstracts and short papers presented at CLADAG 2021, the 13th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Statistics, Computer Science, Applications 'G. Parenti' of the University of Florence, under the auspices of the University of Florence, the SIS and the International Federation of Classification Societies (IFCS).

CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research.

Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings. CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became an important meeting point for people interested in classification and data analysis. A selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG 2021 conceived the Plenary and Invited Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2021 is particularly rich. All in all, it comprises 5 Keynote Lectures, 26 Invited Sessions promoted by the members of the Scientific Program Committee, 10 Contributed Sessions, and a Plenary Session on *Statistical Issues in the COVID-19 Pandemic*. We thank all the session organizers for inviting renowned speakers, coming from many different countries. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts and short papers collected in this book. The Conference was planned as an in presence event; unfortunately due to the persistent uncertainty of the COVID-19 epidemic condition, CLADAG 2021 will be completely online.

Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2021. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Giovanni Camillo Porzio
Carla Rampichini
Chiara Bocci

Florence, September 2021

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

ROBUSTNESS METHODS FOR MODELLING COUNT DATA WITH GENERAL DEPENDENCE STRUCTURES

Marta Nai Ruscone [1]and Dimitris Karlis[2]

[1] DIMA - Department of Mathematics, University of Genoa, (e-mail: `marta.nairuscone@unige.it`)

[2] Department of Statistics, Athens University of Economics and Business, (e-mail: `karlis@aueb.it`)

ABSTRACT: Bivariate Poisson models are appropriate for modelling paired count data. However, the bivariate Poisson model does not allow for a negative dependence structure. Therefore, it is necessary to consider alternatives. A natural way is to consider copulas to generate various bivariate discrete distributions. While such models exist in the literature, the issue of choosing a suitable copula has been overlooked so far. Different copulas lead to different structures and any copula misspecification can render the inference useless. In this work, we consider bivariate Poisson models generated with a copula and investigate its robustness under outliers contamination and model misspecification. Particular focus is on the robustness of copula related parameters. English Premier League data are used to demonstrate the effectiveness of our approach.

KEYWORDS: copula, dependence, outliers, robustness.

1 Introduction

Bivariate Poisson models are appropriate for modelling paired count data exhibiting correlation. Paired count data arise in a wide context including, for example, sports (e.g. the number of goals scored by each one of the two opponent teams in soccer). Several models are available that can incorporate different structures and marginal properties, see for example Karlis & Ntzoufras, 2003. See also the work in Nikoloulopoulos, 2013 for defining models with copulas. While several extensions and models have been proposed, up to our knowledge, issues of robustness have been overlooked. Following da Fonseca & Fieller, 2006, there are two kinds of achieved robustness that one should consider. The first one refers to contamination from outlier observations or, better, from observations that are unexpected under a certain model. The second one refers to model deviation, i.e. a researcher would like to fit the model with such a method that even if the model is not correct the method would protect from deriving inconsistent results.

In this work, we consider a copula based bivariate Poisson distribution. We apply a minimum distance estimation methodology using Hellinger distance. We investigate its robustness under outliers contamination and model misspecification. Particular focus is given on the robustness of copula related parameters that measure the association exhibited by paired count data. The effectiveness of this methodology is examined on data from English Premier League 2013-2014.

2 Copulas

Copula are functions that join multivariate distributions to their marginal distributions (Nelsen, 2007). They describe the dependence structure existing across marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate Gaussian distribution.

A bivariate copula $C : I^2 \to I$, with $I = [0,1]$, is the cumulative bivariate distribution function of the random variables (U,V) with uniform marginal distributions in $[0,1]$. It is define as:

$$C(u,v;\theta) = P(U \leq u, V \leq v;\theta), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U and V.

Let (Y_1,Y_2) be a bivariate random vector with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1,y_2;\theta)$. There always exists a copula function $C(\cdot,\cdot;\theta)$ such that

$$F_{Y_1,Y_2}(y_1,y_2;\theta) = C\big(F_{Y_1}(y_1),F_{Y_2}(y_2);\theta\big), \quad y_1,y_2 \in \mathbb{R}. \tag{2}$$

This result states that each joint distribution can be expressed in terms of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function $C(\cdot,\cdot;\theta)$.

When Y_1 and Y_2 are discrete random variables taking values on some lattice, Ω, the copula C is unique in $(y_1,y_2) \in \Omega$ but not elsewhere. Thus, in the discrete case the mapping from two marginals and a copula $\{F_1,F_2,C\}$ to a bivariate distribution $F(Y_1,Y_2)$ is not one-to-one. However, this is not-uniqueness is of no consequence as the region outside Ω is not of interest in the discrete case (Nelsen, 2007).

3 Bivariate count models based on copulas

For count data, a common starting point is to use the Poisson distribution for the marginals:

$$f(y;\mu_j) = \mu_j^y e^{-\mu_j}/y!, \qquad j=1,2 \quad y=0,1,\dots \tag{3}$$

where $\mu_j > 0$. Models based on copulas in the case of bivariate counts offer the advantage of allowing easy generalization to several different models which is not easy in general. Take, for instance, the Frank copula:

$$C(u,v;\gamma) = -\gamma^{-1}\log\left[1+\frac{(\exp^{-\gamma u}-1)(\exp^{-\gamma v}-1)}{\exp(-\gamma)-1}\right], \quad \gamma\in R-\{0\},\ u,v\in[0,1]. \tag{4}$$

Then

$$F(y_1,y_2;\mu_1,\mu_2,\gamma) \equiv C(F(y_1;\mu_1),F(y_2;\mu_2);\gamma), \tag{5}$$

is a well defined distribution function with a dependence structure. It's probability mass function is

$$\begin{aligned} P(Y_1=y_1,Y_2=y_2;\mu_1,\mu_2,\gamma) &= F(y_1,y_2;\mu_1,\mu_2,\gamma)-F(y_1-1,y_2;\mu_1,\mu_2,\gamma) \\ &\quad -F(y_1,y_2-1;\mu_1,\mu_2,\gamma)+F(y_1-1,y_2-1;\mu_1,\mu_2,\gamma) \end{aligned} \tag{6}$$

In the present work we focus on bivariate models. For a review of discrete valued models based on copulas see Nikoloulopoulos, 2013.

4 Minimum distance estimation

In discrete data, model robustness and efficiency can be achieved almost at the same time, by defining distances that downweight some observations Lindsay, 1994. The minimum distance estimators can be interpreted as weighted likelihood estimators, the weights are determined by some kind of distance between observed and expected frequencies. For example, consider Minimum Hellinger distance estimators based on minimizing

$$\sum_x \left(d(x)^{1/2} - m_\beta(x)^{1/2}\right)^2$$

where $d(x)$ is the observed relative frequency and $m_\beta(x)$ is the probability mass at x with the assumed model with parameters of interest β. It turns out that this quantity leads to estimating equations of the form

$$\sum_x \left(\frac{d(x)}{m_\beta(x)}\right)^{1/2} \frac{\partial m_\beta(x)}{\partial\beta} = 0$$

directly comparable to the ML estimating equations

$$\sum_x \frac{d(x)}{m_\beta(x)} \frac{\partial m_\beta(x)}{\partial \beta} = 0$$

which actually implies that we weight the observations differently (see Lindsay, 1994).

In this work we extend the approach for bivariate count models defined by copulas aiming at deriving robust estimators for both the marginal and the copula parameters. Now x implies a pair of observations. Also, in our case the parameters β to estimate are those of the marginal distribution plus the copula parameter(s).We have also developed an iterative algorithm that facilitates the estimation. In the bivariate case we are interested in the relative frequencies are still reasonable estimators of the underlying probabilities but we need larger sample sizes for that. As we move on higher dimensions, problems similar to that of the regression setting may occur.

5 Application

Bivariate count models are widely used for modelling the outcome of a football game. The two counts refer to the number of goals scored by each team. It seems natural to assume some dependence between the goals to represent the competitive nature of soccer. Our data refer to all scores from English Premier League 2013-2014 where a series of unexpectedly large scores have occurred. We apply a robust approach to estimate the parameters of the model to reduce the effect of the large scores.

References

DA FONSECA, V GRUNERT, & FIELLER, NRJ. 2006. Distortion in statistical inference: the distinction between data contamination and model deviation. *Metrika*, **63**(2), 169–190.

KARLIS, D., & NTZOUFRAS, I. 2003. Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc.*, **52**(3), 381–393.

LINDSAY, B. G. 1994. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.*, **22**(2), 1081–1114.

NELSEN, R B. 2007. *An introduction to copulas*. New York: Springer.

NIKOLOULOPOULOS, A.K. 2013. Copula-based models for multivariate discrete response data. *Pages 231–249 of: Copulae in Mathematical and Quantitative Finance*.