



# 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



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)

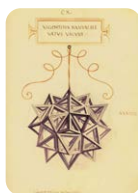
ISBN 978-88-5518-340-6 (PDF)

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](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](http://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](http://www.fupress.com)

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

## INDEX

<b>Preface</b>	<b>1</b>
----------------	----------

### Keynote Speakers

*Jean-Michel Loubes*

<b>Optimal transport methods for fairness in machine learning</b>	<b>5</b>
---	----------

*Peter Rousseeuw, Jakob Raymaekers and Mia Hubert*

<b>Class maps for visualizing classification results</b>	<b>6</b>
--	----------

*Robert Tibshirani, Stephen Bates and Trevor Hastie*

<b>Understanding cross-validation and prediction error</b>	<b>7</b>
--	----------

*Cinzia Viroli*

<b>Quantile-based classification</b>	<b>8</b>
--------------------------------------	----------

*Bin Yu*

<b>Veridical data science for responsible AI: characterizing V4 neurons through deepTune</b>	<b>9</b>
--	----------

### Plenary Session

*Daniel Diaz*

<b>A simple correction for COVID-19 sampling bias</b>	<b>14</b>
---	-----------

*Jeffrey S. Morris*

<b>A seat at the table: the key role of biostatistics and data science in the COVID-19 pandemic</b>	<b>15</b>
---	-----------

*Bhramar Mukherjee*

<b>Predictions, role of interventions and the crisis of virus in India: a data science call to arms</b>	<b>16</b>
---	-----------

*Danny Pfeffermann*

<b>Contributions of Israel's CBS to rout COVID-19</b>	<b>17</b>
---	-----------

### Invited Papers

*Claudio Agostinelli, Giovanni Saraceno and Luca Greco*

<b>Robust issues in estimating modes for multivariate torus data</b>	<b>21</b>
--	-----------

*Emanuele Aliverti*

<b>Bayesian nonparametric dynamic modeling of psychological traits</b>	<b>25</b>
--	-----------

<i>Andres M. Alonso, Carolina Gamboa and Daniel Peña</i> <b>Clustering financial time series using generalized cross correlations</b>	27
<i>Raffaele Argiento, Edoardo Filippi-Mazzola and Lucia Paci</i> <b>Model-based clustering for categorical data via Hamming distance</b>	31
<i>Antonio Balzanella, Antonio Irpino and Francisco de A.T. De Carvalho</i> <b>Mining multiple time sequences through co-clustering algorithms for distributional data</b>	32
<i>Francesco Bartolucci, Fulvia Pennoni and Federico Cortese</i> <b>Hidden Markov and regime switching copula models for state allocation in multiple time-series</b>	36
<i>Michela Battauz and Paolo Vidoni</i> <b>Boosting multidimensional IRT models</b>	40
<i>Matteo Bottai</i> <b>Understanding and estimating conditional parametric quantile models</b>	44
<i>Niklas Bussmann, Roman Enzmann, Paolo Giudici and Emanuela Raffinetti</i> <b>Shapley Lorenz methods for eXplainable artificial intelligence</b>	45
<i>Andrea Cappelozzo, Ludovic Duponchel, Francesca Greselin and Brendan Murphy</i> <b>Robust classification of spectroscopic data in agri-food: first analysis on the stability of results</b>	49
<i>Andrea Cerasa, Enrico Checchi, Domenico Perrotta and Francesca Torti</i> <b>Issues in monitoring the EU trade of critical COVID-19 commodities</b>	53
<i>Marcello Chiodi</i> <b>Smoothed non linear PCA for multivariate data</b>	54
<i>Roberto Colombi, Sabrina Giordano and Maria Kateri</i> <b>Accounting for response behavior in longitudinal rating data</b>	58
<i>Claudio Conversano, Giulia Contu, Luca Frigau and Carmela Cappelli</i> <b>Network-based semi-supervised clustering of time series data</b>	62
<i>Federica Cugnata, Chiara Brombin, Pietro Cippà, Alessandro Ceschi, Paolo Ferrari and Clelia Di Serio</i> <b>Characterising longitudinal trajectories of COVID-19 biomarkers within a latent class framework</b>	64
<i>Silvia D'Angelo</i> <b>Sender and receiver effects in latent space models for multiplex data</b>	68
<i>Anna Denkowska and Stanisław Wanat</i> <b>DTW-based assessment of the predictive power of the copula-DCC-GARCH-MST model developed for European insurance institutions</b>	71
<i>Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha</i> <b>Two-step estimation of multilevel latent class models with covariates</b>	75
<i>Marie Du Roy de Chaumaray and Matthieu Marbac</i> <b>Clustering data with non-ignorable missingness using semi-parametric mixture models</b>	79

<i>Pierpaolo D'Urso, Livia De Giovanni and Vincenzina Vitale</i> <b>Spatial-temporal clustering based on B-splines: robust models with applications to COVID-19 pandemic</b>	83
<i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli and Nicola Torelli</i> <b>PIVMET: pivotal methods for Bayesian relabelling in finite mixture models</b>	87
<i>Tahir Ekin and Claudio Conversano</i> <b>Cluster validity by random forests</b>	91
<i>Luis Angel García-Escudero, Agustín Mayo-Isacar and Marco Riani</i> <b>Robust estimation of parsimonious finite mixture of Gaussian models</b>	92
<i>Silvia Facchinetti and Silvia Angela Osmetti</i> <b>A risk indicator for categorical data</b>	93
<i>Matteo Fasiolo</i> <b>Additive quantile regression via the qgam R package</b>	97
<i>Michael Fop, Dimitris Karlis, Ioannis Kosmidis, Adrian O'Hagan, Cairiona Ryan and Isobel Claire Gormley</i> <b>Gaussian mixture models for high dimensional data using composite likelihood</b>	98
<i>Carlo Gaetan, Paolo Girardi and Victor Muthama Musau</i> <b>On model-based clustering using quantile regression</b>	102
<i>Carlotta Galeone</i> <b>Socioeconomic inequalities and cancer risk: myth or reality?</b>	106
<i>Michael Gallagher, Christophe Biernacki and Paul McNicholas</i> <b>Parameter-wise co-clustering for high dimensional data</b>	107
<i>Francesca Greselin and Alina Jędrzejczak</i> <b>Quantifying the impact of covariates on the gender gap measurement: an analysis based on EU-SILC data from Poland and Italy</b>	108
<i>Alessandra Guglielmi, Mario Beraha, Matteo Giannella, Matteo Pegoraro and Riccardo Peli</i> <b>A transdimensional MCMC sampler for spatially dependent mixture models</b>	112
<i>Christian Hennig and Pietro Coretto</i> <b>Non-parametric consistency for the Gaussian mixture maximum likelihood estimator</b>	116
<i>Yinxuan Huang and Natalie Shlomo</i> <b>Improving the reliability of a nonprobability web survey</b>	120
<i>Maria Iannario and Claudia Tarantola</i> <b>A semi-Bayesian approach for the analysis of scale effects in ordinal regression models</b>	124
<i>Jayant Jha</i> <b>Best approach direction for spherical random variables</b>	128



<i>Maria Kateri</i>	
<b>Simple effect measures for interpreting generalized binary regression models</b>	<b>129</b>
<i>Shogo Kato, Kota Nagasaki and Wataru Nakanishi</i>	
<b>Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data</b>	<b>133</b>
<i>John Kent</i>	
<b>How to design a directional distribution</b>	<b>137</b>
<i>Simona Korenjak-Černe and Nataša Kejžar</i>	
<b>Identifying mortality patterns of main causes of death among young EU population using SDA approaches</b>	<b>141</b>
<i>Fabrizio Laurini and Gianluca Morelli</i>	
<b>Robust supervised clustering: some practical issues</b>	<b>142</b>
<i>Daniela Marella and Danny Pfeffermann</i>	
<b>A nonparametric approach for statistical matching under informative sampling and nonresponse</b>	<b>146</b>
<i>Mariagiulia Matteucci and Stefania Mignani</i>	
<b>Investigating model fit in item response models with the Hellinger distance</b>	<b>150</b>
<i>Matteo Mazziotta and Adriano Pareto</i>	
<b>PCA-based composite indices and measurement model</b>	<b>154</b>
<i>Marcella Mazzoleni, Angiola Pollastri and Vanda Tulli</i>	
<b>Gender inequalities from an income perspective</b>	<b>158</b>
<i>Yana Melnykov, Xuwen Zhu and Volodymyr Melnykov</i>	
<b>Transformation mixture modeling for skewed data groups with heavy tails and scatter</b>	<b>162</b>
<i>Luca Merlo, Lea Petrella and Nikos Tzavidis</i>	
<b>Unconditional M-quantile regression</b>	<b>163</b>
<i>Jesper Møller, Mario Beraha, Raffaele Argiento and Alessandra Guglielmi</i>	
<b>MCMC computations for Bayesian mixture models using repulsive point processes</b>	<b>167</b>
<i>Keefe Murphy, Cinzia Viroli and Isobel Claire Gormley</i>	
<b>Infinite mixtures of infinite factor analysers</b>	<b>168</b>
<i>Stanislav Nagy, Petra Laketa and Rainer Dyckerhoff</i>	
<b>Angular halfspace depth: computation</b>	<b>169</b>
<i>Yarema Okhrin, Gazi Salah Uddin and Muhammad Yahya</i>	
<b>Nonlinear Interconnectedness of crude oil and financial markets</b>	<b>173</b>
<i>M. Rosário Oliveira, Ana Subtil and Lina Oliveira</i>	
<b>Detection of internet attacks with histogram principal component analysis</b>	<b>174</b>
<i>Sally Paganin</i>	
<b>Semiparametric IRT models for non-normal latent traits</b>	<b>178</b>

<i>Giuseppe Pandolfo</i>	
<b>A graphical depth-based aid to detect deviation from unimodality on hyperspheres</b>	<b>182</b>
<i>Panos Pardalos</i>	
<b>Networks of networks</b>	<b>186</b>
<i>Xanthi Pedeli and Cristiano Varin</i>	
<b>Pairwise likelihood estimation of latent autoregressive count models</b>	<b>187</b>
<i>Mark Reiser and Maduranga Dassanayake</i>	
<b>A study of lack-of-fit diagnostics for models fit to cross-classified binary variables</b>	<b>191</b>
<i>Giorgia Riveccio, Jean-Paul Chavas, Giovanni De Luca, Salvatore Di Falco and Fabian Capitanio</i>	
<b>Assessing food security issues in Italy: a quantile copula approach</b>	<b>195</b>
<i>Nicoleta Rogovschi</i>	
<b>Co-clustering for high dimensional sparse data</b>	<b>199</b>
<i>Massimiliano Russo</i>	
<b>Malaria risk detection via mixed membership models</b>	<b>203</b>
<i>Paula Saavedra-Nieves and Rosa M. Crujeiras</i>	
<b>Nonparametric estimation of the number of clusters for directional data</b>	<b>207</b>
<i>Shuchismita Sarkar, Volodymyr Melnykov and Xuwen Zhu</i>	
<b>Tensor-variate finite mixture model for the analysis of university professor remuneration</b>	<b>208</b>
<i>Florian Schuberth</i>	
<b>Specifying composites in structural equation modeling: the Henseler-Ogasawara specification</b>	<b>209</b>
<i>Jarod Smith, Mohammad Arashi and Andriette Bekker</i>	
<b>Network analysis implementing a mixture distribution from Bayesian viewpoint</b>	<b>210</b>
<i>Paul Smith, Peter van der Heijden and Maarten Cruyff</i>	
<b>Measurement errors in multiple systems estimation</b>	<b>211</b>
<i>Valentin Todorov and Peter Filzmoser</i>	
<b>Robust classification in high dimensions using regularized covariance estimates</b>	<b>215</b>
<i>Salvatore Daniele Tomarchio, Luca Bagnato and Antonio Punzo</i>	
<b>Clustering via new parsimonious mixtures of heavy tailed distributions</b>	<b>216</b>
<i>Agostino Torti, Marta Galvani, Alessandra Menafoglio, Piercesare Secchi and Simone Vantini</i>	
<b>A general bi-clustering technique for functional data</b>	<b>217</b>
<i>Laura Trinchera</i>	
<b>Developing a multidimensional and hierarchical index following a composite-based approach</b>	<b>220</b>

<i>Rosanna Verde, Francisco T. de A. De Carvalho and Antonio Balzanella</i> <b>A generalised clusterwise regression for distributional data</b>	223
<i>Marika Vezzoli, Francesco Doglietto, Stefano Renzetti, Marco Fontanella and Stefano Calza</i> <b>A machine learning approach for evaluating anxiety in neurosurgical patients during the COVID-19 pandemic</b>	227
<i>Isadora Antoniano Villalobos, Simone Padoan and Boris Beranger</i> <b>Prediction of large observations via Bayesian inference for extreme-value theory</b>	231
<i>Maria Prosperina Vitale, Vincenzo Giuseppe Genova, Giuseppe Giordano and Giancarlo Ragozini</i> <b>Community detection in tripartite networks of university student mobility flows</b>	232
<i>Ernst Wit and Lucas Kania</i> <b>Causal regularization</b>	236
<i>Qiuyi Wu and David Banks</i> <b>Minimizing conflicts of interest: optimizing the JSM program</b>	240

## Contributed Papers

<i>Antonino Abbruzzo, Maria Francesca Cracolici and Furio Urso</i> <b>Model selection procedure for mixture hidden Markov models</b>	243
<i>Roberto Ascari and Sonia Migliorati</i> <b>A full mixture of experts model to classify constrained data</b>	247
<i>Luigi Augugliaro, Gianluca Sottile and Angelo Mineo</i> <b>Sparse inference in covariate adjusted censored Gaussian graphical models</b>	251
<i>Simona Balzano, Mario Rosario Guarracino and Giovanni Camillo Porzio</i> <b>Semi-supervised learning through depth functions</b>	255
<i>Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta</i> <b>A combined test of the Benford hypothesis with anti-fraud applications</b>	256
<i>Chiara Bardelli</i> <b>Unbalanced classification of electronic invoicing</b>	260
<i>Claudia Berloco, Raffaele Argiento and Silvia Montagna</i> <b>Predictive power of Bayesian CAR models on scale free networks: an application for credit risk</b>	264
<i>Marco Berrettini, Giuliano Galimberti and Saverio Ranciati</i> <b>Semiparametric finite mixture of regression models with Bayesian P-splines</b>	268

<i>Giuseppe Bove</i>	
<b>A subject-specific measure of interrater agreement based on the homogeneity index</b>	<b>272</b>
<i>Antonio Calcagni</i>	
<b>Estimating latent linear correlations from fuzzy contingency tables</b>	<b>276</b>
<i>Andrea Cappozzo, Alessandro Casa and Michael Fop</i>	
<b>Model-based clustering with sparse matrix mixture models</b>	<b>280</b>
<i>Andrea Cappozzo, Luis Angel Garcia Escudero, Francesca Greselin and Agustín Mayo-Iscar</i>	
<b>Exploring solutions via monitoring for cluster weighted robust models</b>	<b>284</b>
<i>Maurizio Carpita and Silvia Golia</i>	
<b>Categorical classifiers in multi-class classification problems</b>	<b>288</b>
<i>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</i>	
<b>Model-based clustering for estimating cetaceans site-fidelity and abundance</b>	<b>292</b>
<i>Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria</i>	
<b>Model-based clustering with parsimonious covariance structure</b>	<b>296</b>
<i>Francesca Condino</i>	
<b>Clustering income data based on share densities</b>	<b>300</b>
<i>Paula Costa Fontichiarì, Miriam Giuliani, Raffaele Argiento and Lucia Paci</i>	
<b>Group-dependent finite mixture model</b>	<b>304</b>
<i>Salvatore Cuomo, Federico Gatta, Fabio Giampaolo, Carmela Iorio and Francesco Piccialli</i>	
<b>A machine learning approach in stock risk management</b>	<b>308</b>
<i>Cristina Davino and Giuseppe Lamberti</i>	
<b>Pathmix segmentation trees to compare linear regression models</b>	<b>312</b>
<i>Houyem Demni, Davide Buttarazzi, Stanislav Nagy and Giovanni Camillo Porzio</i>	
<b>Angular halfspace depth: classification using spherical bagdistances</b>	<b>316</b>
<i>Agostino Di Ciaccio</i>	
<b>Neural networks for high cardinality categorical data</b>	<b>320</b>
<i>F. Marta L. Di Lascio, Andrea Menapace and Roberta Pappadà</i>	
<b>Ali-Mikhail-Haq copula to detect low correlations in hierarchical clustering</b>	<b>324</b>
<i>Maria Veronica Dorgali, Silvia Bacci, Bruno Bertaccini and Alessandra Petrucci</i>	
<b>Higher education and employability: insights from the mandatory notices of the ministry of labour</b>	<b>328</b>
<i>Lorenzo Focardi Olmi and Anna Gottard</i>	
<b>An alternative to joint graphical lasso for learning multiple Gaussian graphical models</b>	<b>332</b>

<i>Francesca Fortuna, Alessia Naccarato and Silvia Terzi</i>	
<b>Functional cluster analysis of HDI evolution in European countries</b>	<b>336</b>
<i>Sylvia Frühwirth-Schnatter, Bettina Grün and Gertraud Malsiner-Walli</i>	
<b>Estimating Bayesian mixtures of finite mixtures with telescoping sampling</b>	<b>340</b>
<i>Chiara Galimberti, Federico Castelletti and Stefano Peluso</i>	
<b>A Bayesian framework for structural learning of mixed graphical models</b>	<b>344</b>
<i>Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu</i>	
<b>Measurement error models on spatial network lattices: car crashes in Leeds</b>	<b>348</b>
<i>Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, Massimo Aria and Roberta Siciliano</i>	
<b>The L<sup>P</sup> data depth and its application to multivariate process control charts</b>	<b>352</b>
<i>Petra Laketa and Stanislav Nagy</i>	
<b>Angular halfspace depth: central regions</b>	<b>356</b>
<i>Michele La Rocca, Francesco Giordano and Cira Perna</i>	
<b>Clustering production indexes for construction with forecast distributions</b>	<b>360</b>
<i>Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli</i>	
<b>Clustering longitudinal data with category theory for diabetic kidney disease</b>	<b>364</b>
<i>Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore</i>	
<b>A redundancy analysis with multivariate random-coefficients linear models</b>	<b>368</b>
<i>Paolo Mariani, Andrea Marletta and Matteo Locci</i>	
<b>The use of multiple imputation techniques for social media data</b>	<b>372</b>
<i>Federico Marotta, Paolo Provero and Silvia Montagna</i>	
<b>Prediction of gene expression from transcription factors affinities: an application of Bayesian non-linear modelling</b>	<b>376</b>
<i>Francesca Martella, Fabio Attorre, Michele De Sanctis and Giuliano Fanelli</i>	
<b>High dimensional model-based clustering of European georeferenced vegetation plots</b>	<b>380</b>
<i>Ana Martins, Paula Brito, Sónia Dias and Peter Filzmoser</i>	
<b>Multivariate outlier detection for histogram-valued variables</b>	<b>384</b>
<i>Giovanna Menardi and Federico Ferraccioli</i>	
<b>A nonparametric test for mode significance</b>	<b>388</b>
<i>Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia and Arnaud Sallaberry</i>	
<b>Visualizing cluster of words: a graphical approach to grammar acquisition</b>	<b>392</b>

<i>Marta Nai Ruscone and Dimitris Karlis</i> <b>Robustness methods for modelling count data with general dependence structures</b>	396
<i>Roberta Paroli, Luigi Spezia, Marc Stutter and Andy Vinten</i> <b>Bayesian analysis of a water quality high-frequency time series through Markov switching autoregressive models</b>	400
<i>Mariano Porcu, Isabella Sulis and Cristian Usala</i> <b>Detecting the effect of secondary school in higher education university choices</b>	404
<i>Roberto Rocci and Monia Ranalli</i> <b>Semi-constrained model-based clustering of mixed-type data using a composite likelihood approach</b>	408
<i>Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Damiana Pieragostino</i> <b>Antibodies to SARS-CoV-2: an exploratory analysis carried out through the Bayesian profile regression</b>	412
<i>Theresa Scharl and Bettina Grün</i> <b>Modelling three-way RNA sequencing data with mixture of multivariate Poisson-lognormal distribution</b>	416
<i>Luca Scrucca</i> <b>Stacking ensemble of Gaussian mixtures</b>	420
<i>Rosaria Simone, Cristina Davino, Domenico Vistocco and Gerhard Tutz</i> <b>A robust quantile approach to ordinal trees</b>	424
<i>Venera Tomaselli, Giulio Giacomo Cantone and Valeria Mazzeo</i> <b>The detection of spam behaviour in review bomb</b>	428
<i>Donatella Vicari and Paolo Giordani</i> <b>Clustering models for three-way data</b>	432
<i>Gianpaolo Zammarchi and Jaromir Antoch</i> <b>Using eye-tracking data to create a weighted dictionary for sentiment analysis: the eye dictionary</b>	436

# UNCONDITIONAL M-QUANTILE REGRESSION

Luca Merlo<sup>1</sup>, Lea Petrella<sup>2</sup> and Nikos Tzavidis<sup>3</sup>

<sup>1</sup> Department of Statistics, Sapienza University of Rome,  
(e-mail: luca.merlo@uniroma1.it)

<sup>2</sup> MEMOTEF Department, Sapienza University of Rome,  
(e-mail: lea.petrella@uniroma1.it)

<sup>3</sup> Department of Social Statistics and Demography and Southampton Statistical Sciences  
Research Institute, University of Southampton,  
(e-mail: N.TZAVIDIS@soton.ac.uk)

**ABSTRACT:** In this paper we develop the unconditional M-quantile regression for modeling unconditional M-quantiles in the presence of covariates. Extending the paper by Firpo *et al.* (2009), we assess the impact of small changes in the explanatory variables on the M-quantile of the unconditional distribution of the dependent variable by running a mean regression of the recentered influence function of the unconditional M-quantile on the covariates. The proposed methodology is applied on the Survey of Household Income and Wealth (SHIW) 2016 conducted by the Bank of Italy.

**KEYWORDS:** Influence function, M-estimation, RIF regression, Robust method

## 1 Introduction

Quantile Regression (QR), as proposed by Koenker & Bassett Jr (1978), has proven to be a powerful tool to explore conditional distributions in many empirical applications. However, if one is interested in how the whole unconditional distribution of the dependent variable responds to changes in the covariates, using the well-known QR would yield misleading inferences (see Firpo *et al.* 2009 and Borah & Basu 2013). Motivated by this interest, Firpo *et al.* (2009) proposed the Unconditional Quantile Regression (UQR) approach for modeling unconditional quantiles of a dependent variable as a function of the explanatory variables. This method builds upon the concept of Recentered Influence Function (RIF) which originates from a widely used tool in robust statistics, namely the Influence Function (IF) discussed in Hampel *et al.* (2011). The RIF of a distributional statistic  $v$  is obtained by adding back the statistic to the IF and it can be thought of as the contribution of an individual observation on  $v$ . In the regression framework where covariates are available, Firpo *et al.* (2009) proposed to replace the dependent variable with the RIF to model the

unconditional quantiles of the response and evaluate the effect of changes in the law of the covariates on unconditional quantiles. When the interest of the research is concentrated on the entire distribution of a response variable, in addition to the classical QR, a possible alternative is represented by the M-quantile regression (MQR) approach proposed by Breckling & Chambers (1988). This method provides a “quantile-like” generalization of the mean regression based on influence functions, combining in a common framework the robustness and efficiency properties of quantiles and expectiles (Newey & Powell 1987), respectively.

In this article, we extend the UQR of Firpo *et al.* (2009) to the M-quantile regression framework. We develop the Unconditional M-quantile Regression (UMQR) to model the M-quantiles of the unconditional distribution of the response variable. In order to analyze how the entire unconditional distribution of the outcome is affected by changes in the distribution of explanatory variables, we regress the RIF of the unconditional M-quantile on the covariates and denote such effect as Unconditional M-Quantile Partial Effect (UMQPE).

## 2 Methodology

Let  $Y$  denote a scalar random variable with absolutely continuous distribution function  $F_Y$ . The M-quantile of order  $\tau \in (0, 1)$  of  $Y$  is defined as the solution,  $\theta_\tau \in \mathbb{R}$ , of the following estimating equation:

$$\int \psi_\tau(y - \theta_\tau) dF_Y(y) = 0, \quad (1)$$

where  $\psi_\tau(u) = |\tau - \mathbf{1}_{(u < 0)}| \psi(u/\sigma_\tau)$ , with  $\psi$  being the first derivative of a convex loss function  $\rho$  and  $\sigma_\tau$  is a suitable scale parameter. In this work, we consider the well-known Huber influence function (Huber (1964)):

$$\psi(u) = u\mathbf{1}_{(|u| \leq c)} + c \text{sign}(u)\mathbf{1}_{(|u| > c)}, \quad (2)$$

where  $c$  denotes a tuning constant bounded away from zero that can be used to trade robustness for efficiency in the model fit. In particular, M-quantiles nicely include quantiles when  $c \rightarrow 0$ ,  $\psi(u) = \text{sign}(u)$ , and expectiles when  $c \rightarrow \infty$ ,  $\psi(u) = u$ .

To build the UMQR model, it follows from Firpo *et al.* (2009) and Hampel *et al.* (2011) that the RIF of the M-quantile  $\theta_\tau$  is defined as:

$$RIF(y; \theta_\tau) = \theta_\tau + IF(y; \theta_\tau) = \theta_\tau + \frac{\psi_\tau(y - \theta_\tau)}{\int \psi'_\tau(y - \theta_\tau) dF_Y(y)}, \quad (3)$$



where  $IF(y; \theta_\tau)$  is the IF of  $\theta_\tau$  and  $\psi'(u) = \mathbf{1}_{(|u| < c)}$  is the derivative of  $\psi$  in (2). In a regression framework when covariates  $\mathbf{X} \subset \mathbb{R}^k$  are available, from (3) we define the UMQR model as follows:

$$\mathbb{E}[RIF(Y; \theta_\tau) | \mathbf{X} = \mathbf{x}] = \theta_\tau + \mathbb{E}\left[\frac{\Psi_\tau(y - \theta_\tau)}{\int \Psi'_\tau(y - \theta_\tau) dF_Y(y)} \middle| \mathbf{X} = \mathbf{x}\right]. \quad (4)$$

Our objective is to identify how small changes in the distribution of  $\mathbf{X}$  affect the M-quantile of the unconditional distribution of  $Y$ . From (4) and Firpo *et al.* (2009), the unconditional effect of the  $\tau$ -th M-quantile, that we denote Unconditional M-quantile Partial Effect,  $\alpha_\tau$ , is formally defined as:

$$\alpha_\tau = \int \frac{d\mathbb{E}[RIF(Y; \theta_\tau) | \mathbf{X} = \mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}) = \frac{1}{s_\tau} \int \frac{d\mathbb{E}[\Psi_\tau(Y - \theta_\tau) | \mathbf{X} = \mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}), \quad (5)$$

where  $F_{\mathbf{X}}$  is the distribution function of  $\mathbf{X}$  and  $s_\tau = \int \Psi'_\tau(y - \theta_\tau) dF_Y(y)$ . As suggested by Firpo *et al.* (2009), we can estimate  $\alpha_\tau$  in (5) via a mean regression of the  $RIF(Y; \theta_\tau)$  as dependent variable onto  $\mathbf{X}$  by using a two-step procedure. Specifically, an estimate  $\hat{\theta}_\tau$  of  $\theta_\tau$  is obtained by solving (1) via Iterative Reweighted Least Squares, substitute  $\hat{\theta}_\tau$  in (3) and then regress the  $RIF(Y; \hat{\theta}_\tau)$  on  $\mathbf{X}$ .

### 3 Application

We investigate the effect of economic and socio-demographic characteristics on Italian households' log-consumption using data from the SHIW 2016. We fit the UMQR at different points of the unconditional distribution of the response and compare the results with standard conditional M-quantile regressions. The tuning constant  $c$  in (2) has been set to 1.345 and 100. In the second case, we obtain the Unconditional Expectile Regression (UER). The results in Table 1 highlight that the impact of income, gender, age and education is very different on the conditional and unconditional distributions of consumption, especially in the tails. This demonstrates the ability of the UMQR to extend mean regression for estimating the effect of covariates, not only at the center, but also at different parts of the unconditional distribution of interest.

### References

BORAH, BIJAN J, & BASU, ANIRBAN. 2013. Highlighting differences between conditional and unconditional quantile regression approaches through an

Variable	MQR			UMQR			ER			UER			
	$\tau$	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Log-Income		<b>0.570</b> (0.011)	<b>0.595</b> (0.007)	<b>0.442</b> (0.010)	<b>0.447</b> (0.038)	<b>0.391</b> (0.032)	<b>0.429</b> (0.038)	<b>0.483</b> (0.011)	<b>0.413</b> (0.008)	<b>0.263</b> (0.011)	<b>0.450</b> (0.038)	<b>0.413</b> (0.033)	<b>0.436</b> (0.038)
Gender		-0.019 (0.016)	-0.011 (0.009)	<b>-0.043</b> (0.014)	-0.011 (0.018)	<b>-0.024</b> (0.012)	<b>-0.038</b> (0.018)	-0.023 (0.016)	<b>-0.026</b> (0.011)	<b>-0.046</b> (0.016)	-0.010 (0.017)	<b>-0.026</b> (0.012)	-0.035 (0.018)
Age		-0.002 (0.003)	0.001 (0.002)	0.004 (0.003)	<b>-0.013</b> (0.003)	<b>0.006</b> (0.002)	<b>0.013</b> (0.003)	-0.001 (0.003)	0.004 (0.002)	<b>0.008</b> (0.003)	<b>-0.011</b> (0.003)	0.004 (0.002)	<b>0.011</b> (0.003)
Marital status													
never married		<b>-0.062</b> (0.020)	<b>-0.084</b> (0.012)	<b>-0.164</b> (0.018)	<b>-0.094</b> (0.025)	<b>-0.141</b> (0.017)	<b>-0.187</b> (0.022)	<b>-0.095</b> (0.020)	<b>-0.138</b> (0.014)	<b>-0.201</b> (0.020)	<b>-0.101</b> (0.024)	<b>-0.138</b> (0.017)	<b>-0.176</b> (0.022)
separated		<b>-0.066</b> (0.025)	<b>-0.056</b> (0.015)	<b>-0.127</b> (0.022)	<b>-0.102</b> (0.034)	<b>-0.151</b> (0.024)	<b>-0.155</b> (0.030)	<b>-0.111</b> (0.025)	<b>-0.137</b> (0.017)	<b>-0.207</b> (0.026)	<b>-0.105</b> (0.033)	<b>-0.137</b> (0.024)	<b>-0.141</b> (0.030)
widowed		-0.040 (0.022)	<b>-0.063</b> (0.013)	<b>-0.119</b> (0.020)	<b>-0.116</b> (0.029)	<b>-0.136</b> (0.019)	<b>-0.111</b> (0.025)	<b>-0.074</b> (0.022)	<b>-0.123</b> (0.015)	<b>-0.193</b> (0.022)	<b>-0.110</b> (0.028)	<b>-0.123</b> (0.019)	<b>-0.107</b> (0.025)
Education level													
elementary school		<b>0.175</b> (0.039)	<b>0.120</b> (0.023)	<b>0.151</b> (0.035)	<b>0.488</b> (0.069)	<b>0.125</b> (0.024)	-0.037 (0.022)	<b>0.188</b> (0.039)	<b>0.161</b> (0.027)	<b>0.187</b> (0.040)	<b>0.446</b> (0.066)	<b>0.161</b> (0.027)	-0.000 (0.022)
middle school		<b>0.240</b> (0.041)	<b>0.203</b> (0.024)	<b>0.316</b> (0.037)	<b>0.645</b> (0.070)	<b>0.269</b> (0.028)	<b>0.060</b> (0.029)	<b>0.281</b> (0.041)	<b>0.294</b> (0.028)	<b>0.398</b> (0.042)	<b>0.590</b> (0.067)	<b>0.294</b> (0.030)	<b>0.094</b> (0.028)
high school		<b>0.248</b> (0.042)	<b>0.235</b> (0.025)	<b>0.383</b> (0.038)	<b>0.652</b> (0.072)	<b>0.355</b> (0.033)	<b>0.147</b> (0.037)	<b>0.313</b> (0.042)	<b>0.363</b> (0.029)	<b>0.500</b> (0.043)	<b>0.598</b> (0.069)	<b>0.363</b> (0.034)	<b>0.168</b> (0.036)
university		<b>0.298</b> (0.045)	<b>0.297</b> (0.027)	<b>0.521</b> (0.040)	<b>0.631</b> (0.076)	<b>0.440</b> (0.040)	<b>0.506</b> (0.053)	<b>0.391</b> (0.045)	<b>0.484</b> (0.031)	<b>0.705</b> (0.046)	<b>0.608</b> (0.073)	<b>0.484</b> (0.042)	<b>0.515</b> (0.052)
Employment status													
self-employed		<b>-0.087</b> (0.024)	0.010 (0.014)	<b>0.083</b> (0.022)	<b>-0.060</b> (0.021)	0.021 (0.019)	<b>0.121</b> (0.038)	<b>-0.058</b> (0.024)	0.023 (0.017)	<b>0.081</b> (0.025)	<b>-0.046</b> (0.020)	0.023 (0.018)	<b>0.107</b> (0.037)
not-employed		0.008 (0.021)	<b>0.027</b> (0.013)	0.035 (0.019)	-0.046 (0.025)	<b>0.037</b> (0.016)	0.037 (0.025)	-0.002 (0.021)	0.014 (0.015)	0.017 (0.022)	<b>-0.052</b> (0.024)	0.014 (0.015)	0.031 (0.024)

**Table 1.** *M*-quantile and Expectile regression results at  $\tau = (0.1, 0.5, 0.9)$ . Parameter estimates are displayed in boldface when significant at the 5% level.

application to assess medication adherence. *Health Economics*, **22**(9), 1052–1070.

BRECKLING, JENS, & CHAMBERS, RAY. 1988. M-quantiles. *Biometrika*, **75**(4), 761–771.

FIRPO, SERGIO, FORTIN, NICOLE M, & LEMIEUX, THOMAS. 2009. Unconditional quantile regressions. *Econometrica: Journal of the Econometric Society*, **77**(3), 953–973.

HAMPEL, FRANK R, RONCHETTI, ELVEZIO M, ROUSSEEUW, PETER J, & STAHEL, WERNER A. 2011. *Robust statistics: the approach based on influence functions*. Vol. 196. John Wiley & Sons.

HUBER, PETER J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, **35**(1), 73–101.

KOENKER, ROGER, & BASSETT JR, GILBERT. 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

NEWBY, WHITNEY K, & POWELL, JAMES L. 1987. Asymmetric least squares estimation and testing. *Econometrica*, 819–847.