CLADAG 2021

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

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Conference Program

The schedule is given in Central European Summer Time (CEST - UTC+02)

Abstract and short papers can be found on the *CLADAG 2021 Book of Abstracts and Short papers* (**https://www.fupress.com/isbn/9788855183406**)

DAY 1- Thursday September 9th 2021

10:15 - 10:45 (CEST) **Opening Session**

OPENING SESSION

11:00 - 12:15 (CEST) **Invited Session #1**

INV 1.1 Time series clustering - Organizers and Chair: *Michele La Rocca, Pietro Coretto*

- *Andres M. Alonso, Carolina Gamboa and Daniel Peña* - Clustering financial time series using generalized cross correlations (p.27)
- *Claudio Conversano, Giulia Contu, Luca Frigau and Carmela Cappelli* - Network-based semi-supervised clustering of time series data (p.62)
- *Pierpaolo D'Urso, Livia De Giovanni and Vincenzina Vitale* - Spatial-temporal clustering based on B-splines: robust models with applications to COVID-19 pandemic (p.83)

INV 1.2 Modern likelihood methods for model based-clustering - Organizer: *Monia Ranalli* - Discussant: *Roberto Rocci*

- *Francesco Bartolucci, Fulvia Pennoni and Federico Cortese* - Hidden Markov and regime switching copula models for state allocation in multiple time-series (p.36)
- *Marie Du Roy de Chaumaray and Matthieu Marbac* - Clustering data with non-ignorable missingness using semi-parametric mixture models (p.79)
- *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 (p.98)

INV 1.3 Robust classification in action - Chair: *Marco Riani*

- *Claudio Agostinelli, Giovanni Saraceno and Luca Greco* - Robust issues in estimating modes for multivariate torus data (p.21)
- *Luis Angel García-Escudero, Agustín Mayo-Iscar and Marco Riani* - Robust estimation of parsimonious finite mixture of Gaussian models (p.92)
- *Fabrizio Laurini and Gianluca Morelli* - Robust supervised clustering: some practical issues (p.142)

INV 1.4 Flexible Bayesian mixture models for complex data - Chair: *Alessandra Guglielmi*

- *Isadora Antoniano Villalobos, Simone Padoan and Boris Beranger* - Prediction of large observations via Bayesian inference for extreme-value theory (p.231)
- *Raffaele Argiento, Edoardo Filippi-Mazzola and Lucia Paci* - Model-based clustering for categorical data via Hamming distance (p.31)
- *Jesper Møller , Mario Beraha, Raffaele Argiento and Alessandra Guglielmi* - MCMC computations for Bayesian mixture models using repulsive point processes (p.167)

12:30 - 13:30 (CEST) **Keynote #1 - Optimal transport methods for fairness in machine learning**

KEY 1

Jean-Michel Loubes - *CLADAG 2021 Book of Abstract and Short papers*, p. 5
Chair: *Francesca Grieselin*

14:00 - 15:00 (CEST) **Contributed Session #1**

CON 1.A Models for clustering

- *Andrea Cappozzo, Luis Angel Garcìa Escudero, Francesca Greselin and Agustìn Mayo-Iscar* - Exploring solutions via monitoring for cluster weighted robust models (p.284)
- *Michele La Rocca, Francesco Giordano and Cira Perna* - Clustering production indexes for construction with forecast distributions (p.360)
- *Roberto Rocci and Monia Ranalli* - Semi-constrained model-based clustering of mixed-type data using a composite likelihood approach (p.408)
- *Donatella Vicari and Paolo Giordani* - Clustering models for three-way data (p.432)

CON 1.B Nonparametric and semiparametric methods

- *Marco Berrettini, Giuliano Galimberti and Saverio Ranciati* - Semiparametric finite mixture of regression models with Bayesian P-splines (p.268)
- *Houyem Demni, Davide Buttarazzi, Stanislav Nagy and Giovanni Camillo Porzio* - Angular halfspace depth: classification using spherical bagdistances (p.316)
- *Francesca Fortuna, Alessia Naccarato and Silvia Terzi* - Functional cluster analysis of HDI evolution in European countries (p.336)
- *Giovanna Menardi and Federico Ferraccioli* - A nonparametric test for mode significance (p.388)

CON 1.C Data analysis in biomedical science

- *Giuseppe Bove* - A subject-specific measure of interrater agreement based on the homogeneity index (p.272)
- *Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli* - Clustering longitudinal data with category theory for diabetic kidney disease (p.364)
- *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 (p.412)
- *Theresa Scharl and Bettina Grün* - Modelling three-way RNA sequencing data with mixture of multivariate Poisson-lognormal distribution (p.416)

CON 1.D Modelling dependence structures

- *Lorenzo Focardi Olmi and Anna Gottard* - An alternative to joint graphical lasso for learning multiple Gaussian graphical models (p.332)
- *Chiara Galimberti, Federico Castelletti and Stefano Peluso* - A Bayesian framework for structural learning of mixed graphical models (p.344)
- *Antonino Abbruzzo, Maria Francesca Cracolici and Furio Urso* - Model selection procedure for mixture hidden Markov models (p.243)

CON 1.E Data analysis in economics and finance

- *Claudia Berloco, Raffaele Argiento and Silvia Montagna* - Predictive power of Bayesian CAR models on scale free networks: an application for credit risk (p.264)
- *Salvatore Cuomo, Federico Gatta, Fabio Giampaolo, Carmela Iorio and Francesco Piccialli* - A Machine Learning Approach in stock risk management (p.308)
- *Francesca Condino* - Clustering income data based on share densities (p.300)
- *Cristina Davino and Giuseppe Lamberti* - Pathmox segmentation trees to compare linear regression models (p.312)

15:15 - 16:30 (CEST) **Invited Session #2**

INV 2.1 Copulas in time series analysis - Chair: *Marta L. Di Lascio, Roberta Pappadà*

- *Anna Denkowska and Stanisław Wanat* - DTW-based assessment of the predictive power of the copula-DCC-GARCH-MST model developed for European insurance institutions (p.71)
- *Yarema Okhrin, Gazi Salah Uddin and Muhammad Yahya* - Nonlinear Interconnectedness of crude oil and financial markets (p.173)
- *Giorgia Rivieccio, Jean-Paul Chavas, Giovanni De Luca, Salvatore Di Falco and Fabian Capitanio* - Assessing food security issues in Italy: a quantile copula approach (p.195)

INV 2.2 Advances in mixture models for matrix-variate and tensor data - Chair: *Antonio Punzo*

- *Katharine Clark and Paul McNicholas* - Using Subset Log-Likelihoods to Trim Outliers in Gaussian Mixture Models
- *Shuchismita Sarkar, Volodymyr Melnykov and Xuwen Zhu* - Tensor-variate finite mixture model for the analysis of university professor remuneration (p.208)
- *Jarod Smith, Mohammad Arashi and Andriette Bekker* - Network analysis implementing a mixture distribution from Bayesian viewpoint (p.210)

INV 2.3 Methods for inference from innovative or multiple data sources - Chair: *Emilia Rocco, Chiara Bocci*

- *Yinxuan Huang and Natalie Shlomo* - Improving the reliability of a nonprobability web survey (p.120)
- *Daniela Marella and Danny Pfeffermann* - A nonparametric approach for statistical matching under informative sampling and nonresponse (p.146)
- *Paul Smith, Peter van der Heijden and Maarten Cruyff* - Measurement errors in multiple systems estimation (p.211)

INV 2.4 Social inequalities - Chair: *Mariangela Zenga*

- *Carlotta Galeone* - Socioeconomic inequalities and cancer risk: myth or reality? (p.106)
- *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 (p.108)
- *Marcella Mazzoleni, Angiola Pollastri and Vanda Tulli* - Gender inequalities from an income perspective (p.158)

16:45 – 17:45 (CEST) **KEY 2**

Keynote #2 - Veridical Data Science: the practice of responsible data analysis and decision-making

Bin Yu - *CLADAG 2021 Additional Abstracts*
Chair: *Maurizio Vichi*

18:00 - 19:15 (CEST)

Invited Session #3

INV 3.1 Recent advances in item response theory models - Chair: *Silvia Cagnone*

- *Mariagiulia Matteucci and Stefania Mignani - Investigating model fit in item response models with the Hellinger distance (p.150)*
- *Michela Battauz and Paolo Vidoni* - Boosting multidimensional IRT models (p.40)
- *Mark Reiser and Maduranga Dassanayake* - A study of lack-of-fit diagnostics for models fit to cross-classified binary variables (p.191)

INV 3.2 Advances in clustering - Chair: *Luca Frigau*

- *Tahir Ekin and Claudio Conversano* - Cluster validity by random forests (p.91)
- *Christian Hennig and Pietro Coretto* - Non-parametric consistency for the Gaussian mixture maximum likelihood estimator (p.116)
- *Qiuyi Wu and David Banks* - Minimizing conflicts of interest: optimizing the JSM program (p.240)

INV 3.3 Advances in parsimonious mixture modelling - Organizer: *Volodymyr Melnykov* - Chair: *Xuwen Zhu*

- *Michael Gallaugher, Christophe Biernacki and Paul McNicholas* - Parameter-wise co-clustering for high dimensional data (p.107)
- *Yana Melnykov, Xuwen Zhu and Volodymyr Melnykov* - Transformation mixture modeling for skewed data groups with heavy tails and scatter (p.162)
- *Salvatore Daniele Tomarchio, Luca Bagnato and Antonio Punzo* - Clustering via new parsimonious mixtures of heavy tailed distributions (p.216)

INV 3.4 Bayesian non parametrics methods for classification - Chair: *Bruno Scarpa*

- *Emanuele Aliverti* - Bayesian nonparametric dynamic modeling of psychological traits (p.25)
- *Sally Paganin* - Semiparametric IRT models for non-normal latent traits (p.178)
- *Massimiliano Russo* - Malaria risk detection via mixed membership models (p.203)

DAY 2- Friday, September 10th 2021

9:15-10:30 (CEST) **ASSEMBLY**

CLADAG Members' Assembly

10:45 – 12:00 (CEST)

Invited Session #4

INV 4.1 Recent developments in symbolic data analysis - Chair: *Paula Brito*

- *Rosanna Verde, Francisco T. de A. De Carvalho and Antonio Balzanella* - A generalised clusteriwise regression for distributional data (p.223)
- *M. Rosário Oliveira, Ana Subtil and Lina Oliveira* - Detection of internet attacks with histogram principal component analysis (p.174)
- *Simona Korenjak-Černe and Nataša Kejžar* - Identifying mortality patterns of main causes of death among young EU population using SDA approaches (p.141)

INV 4.2 Recent advances in dynamic clustering: Markov models and extensions - Chair: *Daniele Tomarchio*

- *Roberto Colombi, Sabrina Giordano and Maria Kateri* - Accounting for response behavior in longitudinal rating data (p.58)
- *Bård Støve, Geir D. Berentsen, Jan Bulla and Antonello Maruotti* - Modeling clusters of corporate defaults: regime-switching models significantly reduce the contagion source
- *Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha* - Two-step estimation of multilevel latent class models with covariates (p.75)

INV 4.3 Networks data analysis and applications - Chair: *Mario R. Guarracino*

- *Silvia D'Angelo* - Sender and receiver effects in latent space models for multiplex data (p.68)
- *Panos Pardalos* - Networks of networks (p.186)
- *Maria Prosperina Vitale, Vincenzo Giuseppe Genova, Giuseppe Giordano and Giancarlo Ragozini* - Community detection in tripartite networks of university student mobility flows (p.232)

INV 4.4 Recent developments in the statistical analysis of categorical data - Chair: *Claudia Tarantola*

- *Silvia Facchinetti and Silvia Angela Osmetti* - A risk indicator for categorical data (p.93)
- *Maria Iannario and Claudia Tarantola* - A semi-Bayesian approach for the analysis of scale effects in ordinal regression models (p.124)
- *Maria Kateri* - Simple effect measures for interpreting generalized binary regression models (p.129)

12:15 – 13:30 (CEST)

Invited Session #5

INV 5.1 New issues in univariate and multivariate quantile regression - Chair: *Lea Petrella*

- *Matteo Bottai* - Understanding and estimating conditional parametric quantile models (p.44)
- *Carlo Gaetan, Paolo Girardi and Victor Muthama Musau* - On model-based clustering using quantile regression (p.102)
- *Luca Merlo, Lea Petrella and Nikos Tzavidis* - Unconditional M-quantile regression (p.163)

INV 5.2 Penalized techniques for data analysis - Chair: *Gianluca Sottile*

- *Niklas Bussmann, Roman Enzmann, Paolo Giudici and Emanuela Raffinetti* - Shapley Lorenz methods for eXplainable artificial intelligence (p.45)
- *Marcello Chiodi* - Smoothed non linear PCA for multivariate data (p.54)
- *Ernst Wit and Lucas Kania* - Causal regularization (p.236)

INV 5.3 Latent variable mixture modeling in epidemiology - Chair: *Maria Iannario*

- *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 (p.64)
- *Xanthi Pedeli and Cristiano Varin* - Pairwise likelihood estimation of latent autoregressive count models (p.187)
- *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 (p.227)

INV 5.4 Co-clustering for temporal sequences and distributional data - Chair: *Rosanna Verde*

- *Antonio Balzanella, Antonio Irpino and Francisco de A.T. De Carvalho* - Mining multiple time sequences through co-clustering algorithms for distributional data (p.32)
- *Nicoleta Rogovschi* - Co-clustering for high dimensional sparse data (p.199)
- *Agostino Torti, Marta Galvani, Alessandra Menafoglio, Piercesare Secchi and Simone Vantini* - A general bi-clustering technique for functional data (p.217)

14:00 – 15:15 (CEST) **Invited Session #6**

INV 6.1 Advances in robust cluster analysis - Organizers: *Luis Angel García-Escudero, Agustín Mayo-Iscar, Marco Riani* - Chair: *Agustín Mayo-Iscar*

- *Andrea Cappozzo, Ludovic Duponchel, Francesca Greselin and Brendan Murphy* - Robust classification of spectroscopic data in agri-food: first analysis on the stability of results (p.49)
- *Valentin Todorov and Peter Filzmoser* - Robust classification in high dimensions using regularized covariance estimates (p.215)
- *Andrea Cerasa, Enrico Checchi, Domenico Perrotta and Francesca Torti* - Issues in monitoring the EU trade of critical COVID-19 commodities (p.53)

INV 6.2 Bayesian analysis of finite and infinite mixtures - Chair: *Bettina Grün*

- *Leonardo Egidi, Roberta Pappadà, Francesco Pauli and Nicola Torelli* - PIVMET: pivotal methods for Bayesian relabelling in finite mixture models (p.87)
- *Alessandra Guglielmi, Mario Beraha, Matteo Giannella, Matteo Pegoraro and Riccardo Peli* - A transdimensional MCMC sampler for spatially dependent mixture models (p.112)
- *Keefe Murphy, Cinzia Viroli and Isobel Claire Gormley* - Infinite mixtures of infinite factor analysers (p.168)

INV 6.3 Latent variable models for constructing composite indices - Chair: *Rosaria Romano*

- *Matteo Mazziotta and Adriano Pareto* - PCA-based composite indices and measurement model (p.154)
- *Florian Schuberth* - Specifying composites in structural equation modeling: the Henseler-Ogasawara specification (p.209)
- *Laura Trinchera* - Developing a multidimensional and hierarchical index following a composite-based approach (p.220)

INV 6.4 Issues in directional data analysis - Chair: *Giovanni Camillo Porzio*

- *Jayant Jha* - Best approach direction for spherical random variables (p.128)
- *Stanislav Nagy, Petra Laketa and Rainer Dyckerhoff* - Angular halfspace depth: computation (p.169)
- *Paula Saavedra-Nieves and Rosa M. Crujeiras* - Nonparametric estimation of the number of clusters for directional data (p.207)

15:30 – 16:30 (CEST) **Keynote #3 - Understanding cross-validation and prediction error**

KEY 3 Robert Tibshirani, Stephen Bates and Trevor Hastie - *CLADAG 2021 Book of Abstract and Short papers*, p. 7
Chair: *Anna Gottard*

16:45 – 17:45 (CEST) **Keynote #4 - Class maps for visualizing classification results**

KEY 4 Peter Rousseeuw, Jakob Raymaekers and Mia Hubert - *CLADAG 2021 Book of Abstract and Short papers*, p. 6
Chair: *Marco Riani*

18:00 - 20:00 (CEST) **Plenary Session - Statistical Issues in the COVID-19 Pandemic**

PLENARY Organizer and Chair: *J. Sunil Rao*

- *Daniel Diaz* - A simple correction for COVID-19 sampling bias (p.14)
- *Jeffrey S. Morris* - A seat at the table: the key role of biostatistics and data science in the COVID-19 pandemic (p.15)
- *Bhramar Mukherjee* - Predictions, role of interventions and the crisis of virus in India: a data science call to arms (p.16)
- *Danny Pfeffermann* - Contributions of Israel's CBS to rout COVID-19 (p.17)

DAY 3 - Saturday, September 11th 2021

10:00 - 11:15 (CEST) **Invited Session #7 and Contributed #2**

INV 7.1 Recent advances in directional statistics - Chair: *Stefania Fensore, Agnese Panzera*

- *Shogo Kato, Kota Nagasaki and Wataru Nakanishi* - Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data (p.133)
- *John Kent*- How to design a directional distribution (p.137)
- *Giuseppe Pandolfo* - A graphical depth-based aid to detect deviation from unimodality on hyperspheres (p.182)

INV 7.2 Recent developments in flexible regression – methods and software - Chair: *Marco Geraci*

- *Matteo Fasiolo* - Additive quantile regression via the qgam R package (p.97)
- *Javier Rubio* - Additive Bayesian variable selection under censoring and misspecification

CON 2.A Robust methods and data diagnostic

- *Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta* - A combined test of the Benford hypothesis with anti-fraud applications (p.256)
- *Ana Martins, Paula Brito, Sónia Dias and Peter Filzmoser* - Multivariate outlier detection for histogram-valued variables (p.384)
- *Petra Laketa and Stanislav Nagy* - Angular halfspace depth: Central regions (p.356)
- *Marta Nai Ruscone and Dimitris Karlis* - Robustness methods for modelling count data with general dependence structures (p.396)
- *Rosaria Simone, Cristina Davino, Domenico Vistocco and Gerhard Tutz* - A robust quantile approach to ordinal trees (p.424)

CON 2.B Web mining and textual data analysis

- *Chiara Bardelli* - Unabalanced classfication of electronic invoicing (p.260)
- *Paolo Mariani, Andrea Marletta and Matteo Locci* - The use of multiple imputation techniques for social media data (p.372)
- *Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia and Arnaud Sallaberry* - Visualizing cluster of words: a graphical approach to grammar acquisition (p.392)
- *Gianpaolo Zammarchi and Jaromir Antoch* - Using eye-traking data to create a weighted dictionary for sentiment analysis: the eye dictionary (p.436)

11:30 - 12:30 (CEST) **Contributed Session #3**

CON 3.A Data analysis in biology and environmental sciences

- *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 (p.292)
- *Federico Marotta, Paolo Provero and Silvia Montagna* - Prediction of gene expression from transcription factors affinities: an application of Bayesian non-linear modelling (p.376)
- *Francesca Martella, Fabio Attorre, Michele De Sanctis and Giuliano Fanelli* - High dimensional model-based clustering of European georeferenced vegetation plots (p.380)
- *Roberta Paroli, Luigi Spezia, Marc Stutter and Andy Vinten* - Bayesian analysis of a water quality high-frequency time series through Markov switching autoregressive models (p.400)

CON 3.B Process and service quality evaluation

- *Antonio Calcagni* - Estimating latent linear correlations from fuzzy contingency tables (p.276)
- *Carmela Iorio, Giuseppe Pandolfo, Michele Staiano, Massimo Aria and Roberta Siciliano* - The LP data depth and its application to multivariate process control charts (p.352)
- *Mariano Porcu, Isabella Sulis and Cristian Usala* - Detecting the effect of secondary school in higher education university choices (p.404)

CON 3.C Mixture models and random effect models

- *Sylvia Frühwirth-Schnatter, Bettina Grün and Gertraud Malsiner-Walli* - Estimating Bayesian mixtures of finite mixtures with telescoping sampling (p.340)
- *Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu* - Measurement error models on spatial network lattices: car crashes in Leeds (p.348)
- *Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore* - A redundancy analysis with multivariate random-coefficients linear models (p.368)
- *Paula Costa Fontichiari, Miriam Giuliani, Raffaele Argiento and Lucia Paci* - Group-dependent finite mixture model (p.304)

CON 3.D Machine learning and statistical learning

- *Simona Balzano, Mario Rosario Guarracino and Giovanni Camillo Porzio* - Semi-supervised Learning through depth functions (p.255)
- *Andrea Cappozzo, Alessandro Casa and Michael Fop* - Model-based clustering with sparse matrix mixture models (p.280)
- *Agostino Di Ciaccio* - Neural networks for high cardinality categorical data (p.320)
- *Luca Scrucca* - Stacking ensemble of Gaussian mixtures (p.420)

CON 3.E Hierarchical clustering and classification methods

- *Roberto Ascari and Sonia Migliorati* - A full mixture of experts model to classify constrained data (p.247)
- *Maurizio Carpita and Silvia Golia* - Categorical classifiers in multi-class classification problems (p.288)
- *F. Marta L. Di Lascio, Andrea Menapace and Roberta Pappadà* - Ali-Mikhail-Haq copula to detect low correlations in hierarchical clustering (p.324)
- *Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria* - Model-based clustering with parsimonious covariance structure (p.296)

12:45 – 13:45 (CEST) **Keynote #5 - Quantile-based classification**

KEY 5

Cinzia Viroli - *CLADAG 2021 Book of Abstract and Short papers*, p. 8

Chair: *Brendan Murphy*

13:45 – 14:00 (CEST) **Closing Statements**

KEY 5