Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

Università Politecnica delle Marche

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | Business Analytics and Data Science Hub

Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM

ISBN 9788891935618AAVV

Contents

Preface **XXII**

1 Plenary Sessions **1**

Inequality indices: accurate simulation-based inference 2
Maria-Pia Victoria-Feser

Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications 7
Christopher K. Wikle, Likun Zhang, Myungsoo Yoo and Xiaoyu Ma

Demographic change and sustainability: novel approaches from digital and computational demography n.a.
Emilio Zagheni

2 Invited Sessions **14**

Machine learning in the design, analysis and integration of sample surveys

Causal Discovery for complex survey data 15
Paola Vicard

Data Integration without conditional independence: a Bayesian Networks approach 21
Pier Luigi Conti, Paola Vicard and Vincenzina Vitale

Mass imputation through Machine Learning techniques in presence of multi-source data 27
Fabrizio De Fausti, Marco Di Zio, Romina Filippini and Simona Toti

Machine learning: different uses and perspectives

Evaluation of pollution containment policies in the US and the role of machine learning algorithms 32
Marco Di Cataldo, Margherita Gerolimetto, Stefano Magrini and Alessandro Spiganti

Machine Learning for Official Statistics: An Application on External Trade n.a.
Mauro Bruno, Maria Serena Causo, Alessio Guandalini, Francesco Ortame and Silvia Russo

Machine learning, data quality and official statistics: challenges and opportunities n.a.
Stefano Menghinello

Statistical Machine Learning for environmental applications

Gaussian Processes and Deep Neural Networks for Spatial Prediction 38
Alex Cucco, Luigi Ippoliti, Nicola Pronello, Pasquale Valentini and Carlo Zaccardi

How can we explain Random Forests in a spatial framework? 42
Natalia Golini, Luca Patelli and Xavier Barber

Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns 48
Jorge Mateu and Abdollah Jalilian

Statistical Process Monitoring for Complex Data in Industry 4.0

A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts 53
Konstantinos Bourazas, Konstantinos Fokianos, Christos Panayiotou and Marios Polycarpou

An Approach for Profile Monitoring via Mixture Regression Models 58
Davide Forcina, Antonio Lepore and Biagio Palumbo

Anomaly Detection in Circular Data 63
Houyem Demni and Giovanni C. Porzio

Advances in Data Science and Statistical Learning [IMS Invited Session]

Empirical Bayes approximation of Bayesian learning: understanding a common practice n.a.
Sonia Petrone

Generalized Fiducial Inference on Differentiable Manifolds - a geometric perspective n.a.
Jan Hannig

Model-free bootstrap and conformal prediction in regression n.a.
Dimitris Politis

ENBIS Session: System Maintenance, Boosting algorithms for regression, and Research Excellence

Boosting Diversity in Regression Ensembles 69
Mathias Bourel, Jairo Cugliari, Yannig Goude and Jean-Michel Poggi

How ENBIS has contributed to the UK Universities Research Excellence Framework 71
Shirley Coleman

Maintenance of degrading systems by dynamic programming or reinforcement learning 75
Antonio Pievatolo

Population Dynamics, Climate Change and Sustainability

Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data n.a.
Côme Cheritel, Roman Hoffmann and Raya Muttarak

Environmental Exposures and Under-5 Mortality in India: A Survival Analysis of DHS data 79
Vinod Joseph Kannankeril Joseph

The impact of temperature on expressed sentiment by migration status: Evidence from geo-located Twitter data 84
Risto Conte Keivabu and Jisu Kim

Statistical Learning for health research and omics data

An alternative to the Dirichlet-multinomial regression model for microbiome data analysis 95
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Modelling ordinal response to treatment in a real-world cohort study 101
Marco Alfò, Maria Francesca Marino and Silvia D’Elia

On the application of the symmetric graphical lasso for paired data 105
Saverio Ranciati and Alberto Roverato

The Economic behaviour of Sustainability

Airports performances and sustainable practices. An empirical study on Italian data 110
Riccardo Gianluigi Serio, Maria Michela Dickson, Diego Giuliani and Giuseppe Espa

Sustainability: still an undefined concept for Italians 116
Raffaele Angelone and Andrea Marletta

Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability 122
Beatrice Biondi and Mario Mazzocchi

Advances in statistical methods for complex problems

Inferring multiple treatment effects from observational studies using confounder importance learning n.a.
Omiros Papaspiliopoulos

Path analysis in Ising models: an application to cyber-security risk assessment 127
Monia Lupparelli and Giovanni M. Marchetti

Causal Regularization n.a.
Lucas Kania and Ernst Wit

Explainable machine learning models

Enhancing Markowitz model: inspection of correlations and tail covariances 133
Gloria Polinesi

Objective and subjective dimension of economic well-being: an approach based on statistical matching 139
Daniela Marella, Vincenzina Vitale and Pierpaolo D'Urso

Sustainable, Accurate, Fair and Explainable Machine Learning Models n.a.
Paolo Giudici and Emanuela Raffinetti

Flexible Learning for Environmental Sustainability

Comparison of traffic flow data sources for air pollution modelling 145
Theresa Smith and Nick McCullen

Data analysis of photogrammetry-based mapping: the sea cucumbers in the Giglio Island as a case-study 150
Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio and Alessio Pollice

Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive 156
Nicole Augustin, Heike Puhlmann and Simon Trust

Inequalities in higher education outcomes: learning from data

Inequalities in international students mobility 163
Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino

Uncovering the interplay of territorial, socioeconomic, and demographic factors in high school to university transition 169
Vincenzo Giuseppe Genova, Andrea Priulla and Martina Vittorietti

Statistical Learning of demographic and health dynamics

Estimating the impact of a vaccine mandate: the case of measles in Italy n.a.
Chiara Chiavenna

Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth n.a.
Andrea Nigri

Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data n.a.
Claire Dooley, Ridhi Kashyap, Douglas Leasure and Francesco Rampazzo

Challenges towards Fairness and Transparency for Data Processes, Algorithms and Decision-Support Models

Challenges on Ethics, and Privacy in AI Applications to Fintech 175
Catarina Silva, Joana Matos Dias and Bernardete Ribeiro

Uncertainty and fairness metrics 180
Anna Gottard

Educational Data mining: methods for complex data in students' assessment

Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring 186
Michela Battauz

Predicting high schools' students performances with registry's data: a machine learning approach 191
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Using response times to identify cheaters in CAT: A simulation study 195
Luca Bungaro, Bernard P. Veldkamp and Mariagiulia Matteucci

Spatial and Spatio-Temporal Modeling: Theory and Applications

A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy 200
Paolo Maranzano, Kelly McConville , Philipp Otto and Felicetta Carillo

Bayesian multi-species N-mixture models for large scale spatial data in community ecology 206
Michele Peruzzi

Minimum contrast for point processes' first-order intensity estimation 212
Nicoletta D'Angelo and Giada Adelfio

Statistical Framework for Measuring the Sustainability of Tourism

Data validity and statistical conformity with Benford's Law: the case of tourism in Sicily 217
Roy Cerqueti and Davide Provenzano

Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model 228
Claudia Cappello, Sabrina Maggio and Sandra De Iaco

Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data 232
Ida Camminatiello, Rosaria Lombardo, Jean-Francois Durand and Leonardo S. Alaimo

Statistical learning for well-being analysis

Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach 238
Mariateresa Ciommi, Chiara Gigliarano, Francesca Mariani and Gloria Polinesi

The fuzzy set approach as statistical learning for the analysis of multidimensional well-being 244
Gianni Betti, Federico Crescenzi, Antonella D'Agostino and Laura Neri

What Makes a Satisfying Life? Prediction and Interpretation with Machine-Learning Algorithms n.a.
Conchita D'Ambrosio

Bayesian contributions to Statistical Learning

A Bayesian framework for early cancer screening 249
Sally Paganin and Jeff Miller

Imputing Synthetic Pseudo Data from Aggregate Data: Development and Validation for Precision Medicine n.a.
Cecilia Balocchi

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach 254
Filippo Ascolani and Valentina Ghidini

Data Visualization for Smart Insights and Advanced Predictive Analytics

Applications of data visualization for industry 259
Martina Dossi, Stefano Sangaletti, Marilena Di Bari and Federica Bruschini

Some Notes on the Use of the Circular Boxplot n.a.
Giovanni Camillo Porzio and Davide Buttarazzi

TERRA: a smart visualization tool for international trade in goods statistics 265
Francesco Amato, Mauro Bruno and Maria Serena Causo

Methods for the analysis of distributional data

Clustering of Distributional Data based on LDQ transformation 271
Gianmarco Borrata and Rosanna Verde

Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis 276
Francesca Fortuna, Fabrizio Maturo and Tonio Di Battista

Multivariate Parametric Analysis of Distributional Data n.a.
Paula Brito

Migrants and Refugees in Europe: social, economic and health-related issues

Labor Market Return to Refugees' Human Capital Investment: A Natural Experiment in Sweden n.a.
Eleonora Mussino

Social networks and loneliness among older migrants in Italy 282
Viviana Amati, Eralba Cela and Elisa Barbiano di Belgiojoso

The Italian Decree on Security: An Analysis of the Impact on Asylum Applications 287
Giorgio Piccitto

Modelling and Forecasting High-dimensional time series

Adaptive combinations of tail-risk forecasts 293
Alessandra Amendola, Vincenzo Candila, Antonio Naimoli and Giuseppe Storti

Are Monetary Policy Announcements related to Volatility Jumps? 299
Giampiero Gallo, Demetrio Lacava and Edoardo Otranto

Regularized Estimation and Prediction of the El Nino/Southern Oscillation Cycle n.a.
Alessandro Giovannelli and Tommaso Proietti

3 Contributed Sessions 305

Bayesian nonparametric methods

Bayesian density estimation for modeling age-at-death distribution 306
Davide Agnoletto, Tommaso Rigon and Bruno Scarpa

Bayesian mixing distribution estimation in the Gaussian-smoothed 1-Wasserstein distance 311
Catia Scricciolo

Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models 316
Francesco Denti, Antonio Di Noia and Antonietta Mira

Bayesian nonparametric multiple change point detection for time series of compositional data 322
Edoardo Marchionni and Riccardo Corradin

Galton-Watson process: a non parametric prior for the offspring distribution 328
Massimo Cannas, Michele Guindani and Nicola Piras

Hierarchical processes in survival analysis 333
Riccardo Cogo, Federico Camerlenghi and Tommaso Rigon

Economics and Statistics

A regression analysis for count data to investigate the effectiveness of incentives on the adoption of 4.0 technologies 339
Stefano Bonnini and Michela Borghesi

Statistical analysis on SDGs indicators related to environmental sustainability 344
Najada Firza, Anisa Bakiu and Dante Mazzitelli

Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach 349
Yuri Calleo, Simone Di Zio and Francesco Pilla

Stocks price forecasts using Stochastic Differential Equations: an empirical assessment 355
Dario Frisardi and Matteo Spuri

The Added-Worker Effect within Italian Households 361
Donata Favaro and Anna Giraldo

Health statistics 1

A model for the natural history of breast cancer: application to a Norwegian screening dataset 365
Laura Bondi, Marco Bonetti and Solveig Hofvind

Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data 370
Elena Ballante

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach 375
Chiara Masci, Marta Spreafico and Francesca Ieva

Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data 381
Ilaria Bombelli and Daniele De Rocchi

Tracking attitudes towards COVID vaccines: A text mining analysis 387
Leonardo Scarso, Marco Novelli and Francesco Saverio Violante

Treatment effect assessment in observational studies with multi-level treatment and outcome 393
Federica Cugnata, Paola Vicard, Paola M.V. Rancoita, Fulvia Mecatti, Clelia Di Serio and Pier Luigi Conti

Indicators: composition, uses and limitations

Are European consumers willing to pay the true price for sustainable food? 399
Luca Secondi and Mengting Yu

Can the reliability of composite indexes be impacted by uncertainty of individual indicators? 406
Caterina Giusti, Stefano Marchetti and Vincenzo Mauro

Initial Coin Offerings and ESG: allies or enemies? 411
Alessandro Bitetto and Paola Cerchiello

On the impact of intraclass correlation in the ANVUR evaluation of academic departments 417
Giorgio Edoardo Montanari and Marco Doretti

Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes 422
Luigi Biggeri, Stefano Marchetti, Caterina Giusti, Monica Pratesi, Francesco Schirripa Spagnolo and Gaia Bertarelli

Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis 428
Camilla Salvatore, Annamaria Bianchi and Silvia Biffignandi

Multivariate data analysis 1

A note on most powerful tests for right censored survival data 434
Maria Veronica Vinattieri and Marco Bonetti

Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data 439
Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data 445
Farah Naz and Elena Ballante

ROBOUT: a multi-step methodology for conditional outlier detection 450
Matteo Farnè and Angelos Vouldis

Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments 456
Rosamarie Frieri, Alessandro Baldi Antognini, Maroussa Zagoraiou, and Marco Novelli

Separation scores: a new statistical tool for scoring and ranking partially ordered data 462
Marco Fattore

Statistics in Society 1

Community detection analysis with robin on hashtag network 468
Valeria Policastro, Francesco Santelli and Giancarlo Ragozini

Film Tourism Motivation through the lens of Trip Advisor data 474
Nicolò Biasetton, Marta Disegna, Girish Prayag and Elena Barzizza

Life satisfaction and social activities in later life in Italy: a focus on the Internet use 480
Claudia Furlan and Silvia Meggiolaro

Social capital endowment's role in the intergenerational transmission of education 485
Alessandra Trimarchi, Maria Gabriella Campolo and Antonino Di Pino Incognito

Streaming Data from Social Networks to Track Political Trends 490
Emiliano del Gobbo and Barbara Cafarelli

The scientific production on gender dysphoria: a bibliometric analysis 495
Maria Gabriella Grassia, Marina Marino, Massimo Aria, Rocco Mazza,
Luca D'Aniello and Agostino Stavolo

Assessment and Education

A hierarchical modelling approach to explain differential functioning of mathematics items by student's gender 500
Clelia Cascella

A latent variable approach to Millennials' knowledge of green finance 506
Maria Iannario, Alessandra Tanda and Claudia Tarantola

Archetypal analysis and latent Markov models: A step-wise approach 512
Lucio Palazzo, Rosa Fabbricatore and Francesco Palumbo

From high school to university: academic intentions and enrolment of foreign students in Italy 518
Francesca Di Patrizio, Eleonora Trappolini and Cristina Giudici

Growth models for the progress test in Italian dentistry degree program 523
Giulio Biscardi, Leonardo Grilli, Carla Rampichini, Laura Antonucci and Corrado Crocetta

The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions 527
Francesco Santelli, Teresa Gentile, Davide Bizjak and Lorenzo Fattori

Working Students and job market outcomes: Insights from the University of Florence 532
Gabriele Lombardi, Valentina Tocchioni and Alessandra Petrucci

Bayesian methods and applications 1

Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation 538
Elena Sabbioni, Enrico Bibbona, Gianluca Mastrantonio and Guido Sanguinetti

Approximate Bayesian Computation for Probabilistic Damage Identification 544
Cecilia Viscardi, Silvia Monchetti, Luisa Collodi, Gianni Bartoli, Michele Betti, Michele Boreale and Fabio Corradi

Estimation of scientific productivity with a hierarchical Bayesian model 550
Maura Mezzetti and Ilia Negri

Heat waves and free-knots splines 555
Gioia Di Credico and Francesco Pauli

The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector 560
Marco Dalla Pria and Silvia Montagna

Health and mortality

A novel definition of comorbidity based on the Global Burden of Diseases project weights 566
Angela Andreella, Lorenzo Monasta and Stefano Campostrini

An Age-Period-Cohort model of gender gap in youth mortality 572
Giacomo Lanfiuti Baldi and Andrea Nigri

Kinlessness in adult and old age across Europe 578
Marta Pittavino, Bruno Arpino and Elena Pirani

Parameter orthogonalization for Siler mortality model 584
Claudia Di Caterina and Lucia Zanotto

Pseudo-observations in survival analysis 590
Marta Cipriani, Alfonso Piciocchi, Valentina Arena and Marco Alfò

Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity 595
Alessandro Feraldi, Cristina Giudici and Nicolas Brouard

Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence 599
Micaela Arcaio and Anna Maria Parroco

Mixture Models

An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks 605
Dalila Failli, Maria Francesca Marino and Francesca Martella

Constrained Mixtures of Generalized Normal Distributions 611
Pierdomenico Duttilo, Alfred Kume and Stefano Antonio Gattone

Mixture-based clustering with covariates for ordinal responses 617
Kemmawadee Preedalikit, Daniel Fernàndez, Ivy Liuc, Louise McMillan, Marta Nai Ruscone and Roy Costilla

Partial membership models for soft clustering of multivariate count data 623
Emiliano Seri, Thomas Brendan Murphy and Roberto Rocci

Regression for mixture models for extremes 629
Viviana Carcaiso, Ilaria Prosdocimi and Isadora Antoniano-Villalobos

Robust matrix-variate mixtures of regressions 635
Salvatore Daniele Tomarchio and Michael P. B. Gallaugher

Sampling methods and analysis of survey data

On the use of auxiliary information to define the sampling design for large-scale geospatial data 641
Chiara Bocci and Emilia Rocco

Optimal joint inclusion probabilities for spatial sampling n.a.
Giuseppe Arbia, Piero Demetrio Falorsi and Vincenzo Nardelli

Robustness and Balance of Sampling or Experimental Designs and Mixture of Designs 647
Yves Tillé and Ejub Talovic

Robustness Bounds for Sampling and Experimental Designs 654
Ejub Talovic and Yves Tillé

Statistical Matching: Hotdeck or Propensity Score? 661
Elena Dalla Chiara, Marcello D'Orazio and Federico Perali

The Italian experience on register-based statistics considering measurement, coverage and sampling errors 667
Marco Di Zio, Romina Filippini and Simona Toti

Space-time statistics

A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis 673
Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Daria Valente and Marco Gamba

An approach to cluster time series extremes with spatial constraints 679
Alessia Benevento, Fabrizio Durante and Roberta Pappadà

An integrated space-time model to evaluate the innovation drivers in Italy 685
Emma Bruno, Rosalia Castellano and Gennaro Punzo

Revealing the dynamic relations between traffic and crowding using big data from mobile phone network 691
Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

SMaC: Spatial Matrix Completion method 697
Giulio Grossi, Alessandra Mattei and Georgia Papadogeorgou

The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks 702
Andrea Gilardi and Riccardo Borgoni

Clustering and classification 1

A clustering model for flow data: an application to international student mobility 708
Cinzia Di Nuzzo and Donatella Vicari

Contingency tables with structural zeros and discrete copulas 713
Roberto Fontana, Elisa Perrone and Fabio Rapallo

Levels Merging in the Latent Class Model 719
Christophe Biernacki

Model-based clustering of count processes with multiple change 725
Shuchismita Sarkar and Xuwen Zhu

Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data 729
Jana Cibulková, Zdeněk Šulc, Hana Řezanková and Jaroslav Horníček

Spectral clustering of mixed data via association-based distance 735
Alfonso Iodice D'Enza, Francesco Palumbo and Cristina Tortora

Dynamic models and time series

A graph based convolution Neural Network approach for forecast reconciliation 741
Andrea Marcocchia and Pierpaolo Brutti

A multivariate hidden semi-Markov model for the analysis of multiple air pollutants 747
Marco Mingione, Pierfrancesco Alaimo Di Loro, Francesco Lagona and Antonello Maruotti

A smooth transition autoregressive model for matrix-variate time series 753
Andrea Bucci

Dynamic network models with time-varying nodes 759
Luca Gherardini, Mauro Bernardi and Monia Lupparelli

Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling 765
Ivan Sciascia, Andrea Crosino and Andrea Genre

Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models 770
Mirko Armillotta

Environmental learning and indicators

Assessing the performance of nuclear norm-based matrix completion methods on CO_2 emissions data 776
Rodolfo Metulini, Francesco Biancalani, Giorgio Gnecco and Massimo Riccaboni

Deep Learning for smart and sustainable agriculture 782
Amalia Vanacore, Armando Ciardiello, Annalisa Izzo, Pierdomenico Zaffino, Carolina Vecchio, Gennaro Pio Auricchio and Luigi Uccelli

Do green transition, environmental taxes and renew-able energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression 788
Aamir Javed, Agnese Rapposelli and Asif Javed

Doubly Robust DID for National Parks evaluation: “just” environmental benefits, or socioeconomics impacts as well? 795
Riccardo D’Alberto, Francesco Pagliacci and Matteo Zavalloni

On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us? 801
Lorenzo Mori and Maria Rosaria Ferrante

Small scale analysis of energy vulnerability in the municipality of Palermo 806
Giuliana La Mantia

Health statistics 2

A test for non-differential misclassification error in database epidemiological studies 812
Giorgio Limoncella, Leonardo Grilli, Emanuela Dreassi, Carla Rampichini, Robert Platt and Rosa Gini

Is the COVID-19 ‘color code’ of Italian regions subjected to political manipulation? 816
Giovanni Busetta and Fabio Fiorillo

Modelling multilevel ordinal response under endogeneity: application to DTC patients’ outcome 822
Silvia D’Elia

Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques 827
Nicole Fontana, Laura Savaré and Francesca Ieva

Optimal two-stage design based on error rates under a Bayesian perspective 833
Susanna Gentile and Valeria Sambucini

Migrants in Italy and return migration

Comparing migrant and “native” Italian adolescents in risky behaviours from FSS and SHARE Corona surveys n.a.
Daniela Foresta

EU-Border crisis on Twitter: sentiments and misinformation analysis 839
Elena Ambrosetti, Cecilia Fortunato and Sara Miccoli

Graduates' interregional migration in times of crisis: the Italian case 843
Thaís García-Pereiro, Ivano Dileo and Anna Paterno

Intentions to stay: The experience of return migrants in Albania 848
Maria Carella, Thaís García-Pereiro, Roberta Pace and Anna Paterno

Return migration to home country: a systematic literature review with text mining and topic modelling 853
Cecilia Fortunato, Andrea Iacobucci and Elena Ambrosetti

The allocation of time within native and foreign couples living in Italy 860
Giovanni Busetta, Maria Gabriella Campolo and Antonino Di Pino Incognito

Είλείϑυια comes from afar: The foreigners’ contribution to fertility by Italian provinces 866
Eleonora Miaci, Cristina Giudici, Eleonora Trappolini, Marina Attili, Cinzia Castagnaro and Antonella Guarneri

Sustainability assessment

ESG, sustainability and stock market risk 871
Michele Costa

Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment 875
Gloria Solano-Hermosilla, Jesus Barreiro-Hurle and Ilaria Amerise

Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: innovative approaches 881
Vito Santarcangelo, Angelo Lamacchia, Emilio Massa, Saverio Gianluca Crisafulli, Massimiliano Giacalone and Vincenzo Basile

Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques 887
Massimo Gastaldi, Ginevra Virginia Lombardi, Agnese Rapposelli and Giulia Romano

Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism 893
Francesca Bassi, Josè Fernando Vera and Juan Antonio Marmolejo Martìn

Sustainable tourism: a survey on the propensity towards eco-friendly accommodations 899
Claudia Furlan and Giovanni Finocchiaro

Bayesian methods and applications 2

A comparison of computational approaches for posterior inference in Bayesian Poisson regression 903
Laura D’Angelo

Bias-reduction methods for Poisson regression models 908
Luca Presicce, Tommaso Rigon and Emanuele Aliverti

Finite Mixture Model for Multiple Sample Data 913
Alessandro Colombi, Raffaele Argiento, Federico Camerlenghi and Lucia Paci

On Bayesian power analysis in reliability 918
Fulvio De Santis, Stefania Gubbiotti and Francesco Mariani

Power priors elicitation through Bayes factors 923
Roberto Macrì Demartino, Leonardo Egidi and Nicola Torelli

Predictive Bayes factors 929
Leonardo Egidi and Ioannis Ntzoufras

Clustering and classification 2

A Clusterwise Regression Method for Distributional-Valued Data 935
Antonio Balzanella, Rosanna Verde and Francisco de A.T. de Carvalho

A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels 939
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Introducing a novel directional distribution depth function for supervised classification 945
Edoardo Redivo and Cinzia Viroli

Clustering alternatives in the preference-approval context 950
Alessandro Albano, José Luis Garcia-Lapresta , Mariangela Sciandra and Antonella Plaia

Computational assessment of k-means clustering on a Structural Equation Model based index 955
Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Handling missing data in complex phenomena: an ultrametric model-based approach for clustering 961
Francesca Greselin and Giorgia Zaccaria

Economics and labour markets

A multivariate ranking analysis on the employability of young adults 967
Rosa Arboretti, Elena Barzizza, Nicolo Biasetton, Riccardo Ceccato, Monica Fedeli and Concetta Tino

Analysis of the Gender Pay Gap in the Italian Labour Market 973
Giulia Cappelletti and Daniele Toninelli

Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes 979
Lorenzo Giammei

Patterns of flexible employment careers. Does measurement error matter? 985
Mauricio Garnier-Villarreal, Dimitris Pavlopoulos and Roberta Varriale

Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention 991
Ulpiani Kocollari, Fabio Demaria and Maddalena Cavicchioli

The CAP instruments impact on GVA and employment: a multivalued treatment approach 997
Montezuma Dumangane and Marzia Freo

The determinants of leaving the parental home in Italy: 2012-18 1003
Ilaria Rocco and Gianpiero Dalla Zuanna

Environmental modeling

A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy 1109
Michela Frigeri, Alessandra Guglielmi and Giovanni Lonati

A preliminary study on shape descriptors for the characterization of microplastics ingested by fish 1015
Greta Panunzi, Tommaso Valente, Marco Matiddi and Giovanna Jona Lasinio

Artificial neural network in predicting odour concentrations: a case study 1021
Veronica Distefano and Gideon Mazuruse

Bayesian analysis of PM10 concentration by spatio-temporal ARIMA and STS models 1026
Michela Frigeri and Ilenia Epifani

Functional ANOVA to monitor yearly Adriatic sea temperature variations 1032
Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Nicola Di Deo

New perspectives in the measurement of biodiversity 1038
Linda Altieri, Daniela Cocchi and Massimo Ventrucci

Multivariate data analysis 2

Feature Selection via anomaly detection autoencoders in radiogenomics studies 1044
Alessia Mapelli, Michela Carlotta Massi, Nicola Rares Franco, Francesca Ieva, Catharine West, Petra Seibold, Jenny Chang-Claude and the REQUITE and RADprecise Consortia

Further considerations on the Spectral Information Criterion 1050
Luca Martino

How to increase the power of the test in sparse contingency tables: a simulation study 1057
Federica Nicolussi and Manuela Cazzaro

Latent event history models for quasi-reaction systems 1063
Matteo Framba, VeronicaVinciotti and Ernst Wit

Quantile-based graphical models for continuous and discrete variables 1069
Luca Merlo, Marco Geraci and Lea Petrella

The logratio Student t distribution 1075
Gianna Monti and Gloria Mateu-Figueras

Statistics in Society 2

A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period 1079
Mauro Mussini

Bayesian networks as a territorial gender impact assessment tool 1084
Flaminia Musella, Lorenzo Giammei, Fulvia Mecatti and Paola Vicar

Can statistics be helpful in detecting electoral fraud? 1088
Massimo Attanasio, Vincenzo G. Genova and Michele Tumminello

Companies' sustainability disclosure and contrast to hunger: the role of social inclusion 1093
Chiara Di Maria and Rodolfo Damiano

Passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017 1099
Riccardo Ievoli, Lucio Palazzo and Giancarlo Ragozini

Topic Modeling for the travel and tourism industry: classical and innovative methods compared 1105
Fabrizio Di Mari

Bayesian methods and applications 3

An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior 1111
Paolo Onorati and Brunero Liseo

Bayesian analysis of Amazon's best-selling books via finite nested mixture model 1117
Laura D'Angelo and Francesco Denti

Binomial Extended Stochastic Block Model for Brain Networks 1121
Valentina Ghidini, Sirio Legramanti and Raffaele Argiento

Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures 1127
Giulia Capitoli, Simone Colombara, Alessia Cotroneo, Francesco De Caro, Riccardo Morandi, Chiara Schembri, Alfredo G. Zapiola and Francesco Denti

Efficient expectation propagation for high-dimensional probit models 1133
Augusto Fasano, Niccolo Anceschi, Beatrice Franzolini and Giovanni Rebaudo

Model-based clustering of non-stationary time series with common historical change times 1139
Riccardo Corradin, Luca Danese, Wasiur KhudaBukhsh and Andrea Ongaro

Functional Data Analysis

A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves 1145
Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio

Conditional Gaussian Graphical Models for Functional Variables whit Partial Separable Operators 1149
Rita Fici, Gianluca Sottile and Luigi Augugliaro

Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot 1155
Annachiara Rossi, Andrea Cappozzo and Francesca Ieva

Functional Graphical Models to map Brexit debate on Twitter 1160
Nicola Pronello, Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Sara Fontanella

Measuring Dependence in Multivariate Functional Datasets 1166
Francesca Ieva, Michael Ronzulli and Anna Maria Paganoni

Robust Statistical Process Monitoring of Multivariate Functional Data 1173
Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo

The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021 1179
Veronica Mazzola, Giovanni Bonaccorsi, Piercesare Secchi and Francesca Ieva

Machine Learning and text mining

A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement 1185
Giulio Giacomo Cantone, Simone Del Sarto and Michela Gnaldi

Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study 1191
Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti, Andrea R. Beccari, Filippo Lunghini, Carmine Talarico and Carmen Gratteri

Hedging global currency risk with factorial machine learning models 1197
Paolo Pagnottoni and Alessandro Spelta

InstanceSHAP: An instance-based estimation approach for Shapley values 1203
Golnoosh Babaei and Paolo Giudici

Networks & Nature Based Solutions: an application for Milan hydric resources 1209
Alessia Forciniti and Emma Zavarrone

The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization 1215
Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Agostino Stavolo

Multivariate data analysis 3

A comparison of different techniques for handling missing covariate values in propensity score methods 1219
Anna Zanovello, Alessandra R. Brazzale and Omar Paccagnella

A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach 1224
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

A tool for assessing weak identifiability of statistical models 1230
Antonio Di Noia, Francesco Denti and Antonietta Mira

Computing Highest Density Regions with Copulae 1235
Nina Deliu and Brunero Liseo

Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions 1241
Francesca Labanca and Anna Gottard

Sequential marginal likelihood selection for the estimation of sparse correlation matrices 1246
Claudia Di Caterina and Davide Ferrari

Nonparametric statistical methods

A Comparison of Distribution-Free Control Charts 1252
Michele Scagliarini

Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida 1257
Dafne Zorzetto

Comparing three robust procedures for CANDECOMP/PARAFAC estimation 1262
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov

How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods 1268
Anna Vesely and Angela Andreella

Non Parametric Combination methodology: a literature review on recent developments 1274
Elena Barzizza, Nicolò Biasetton and Riccardo Ceccato

Regression modeling

A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law 1280
Carlo Cusatelli, Massimiliano Giacalone and Eugenia Nissi

A variable selection procedure based on predictive ability: a preliminary study on logistic regression 1285
Rosaria Simone and Mariarosaria Coppola

Comparison of binary regressions with asymmetric link function for imbalanced data 1291
Michele La Rocca, Marcella Niglio and Marialuisa Restaino

New advances in Regression Forests 1297
Mila Andreani, Lea Petrella and Nicola Salvati

On the Optimal Non-Convexity of Penalty in Sparse Regression Models 1303
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

Using expectile regression with latent variables for digital assets 1309
Beatrice Foroni, Luca Merlo and Lea Petrella

4 Program 1315

Computing Highest Density Regions with Copulae

Nina Deliu[a] and Brunero Liseo[b]

[a]MEMOTEF, Sapienza Università di Roma, `nina.deliu@uniroma1.it`
[b]MEMOTEF, Sapienza Università di Roma, `brunero.liseo@uniroma1.it`

Abstract

We investigate the problem of deriving highest density regions (HDRs) from multivariate data samples. We are interested in estimating minimum volume sets that contain a given probability. In the case of unknown distribution probabilities f, the problem involves their estimation, which may be challenging in multidimensional settings. Motivated by the ubiquitous role of copula modelling in modern statistics, we explore their use in the context of HDR estimation. Rather than directly estimating the multivariate f, we propose to estimate the marginals and their dependence structure, i.e., the copula structure, separately. We evaluate this new method, using both a parametric and a nonparametric approach, in a number of synthetic experiments and considering a real dataset.

Keywords: Highest density regions, Copula modelling, Kernel density estimation

1. Introduction

A ubiquitous problem in statistics is to derive statistical intervals or *regions* (in the multivariate setting) for population parameters or other unknown quantities. Given a random sample of data, they provide a way to quantify the uncertainty about a quantity of interest, or simply a way to summarize the information contained in a distribution. In this work, we are interested in statistical *regions* for summarizing probability distributions and we focus on one approach to addressing this problem: *highest density regions* (HDRs, 1). As the name suggests, an HDR specifies the set of points of highest density: the density for every point inside the region is greater than that for every point outside it. More specifically, as we will better discuss in Section 2, the concrete problem is to estimate minimum volume sets of the form $R(f_\alpha) = \{\mathbf{x}\colon f_{\mathbf{X}}(\mathbf{x}) \geq f_\alpha\}$, such that $P(\mathbf{X} \in R(f_\alpha)) \geq 1 - \alpha$, where $f_{\mathbf{X}}$ is the probability distribution of the variable of interest $\mathbf{X} \in \mathbb{R}^d$ and $1 - \alpha$, with $\alpha \in (0, 1)$, a prespecified coverage probability. In principle, HDRs can be derived for any probability distribution and their scope can be widely different. The following are possible applications of HDRs.

Forecasting The goal is to obtain a "prediction region" for a set of observable variables in order to inform any required action (for illustrative examples, see e.g., 1; 2);

Anomaly detection The goal is to detect abnormal observations from a sample: if a data point does not belong to a region of normal or concentrated data, then it is regarded as anomalous (see e.g., 3);

Unsupervised or semi-supervised classification Identify areas or clusters with a relatively high concentration of a given phenomenon, e.g., areas with remarkably high coronavirus incidence (4).

Such regions are of interest in Bayesian analysis as well, in the formulation of *highest posterior density regions* and *credibility regions* (5; 6). In that context, they are based on a posterior distribution.

Because of their flexibility "to convey both multimodularity and asymmetry in the forecast density", HDRs are argued to be a "more effective summary of the forecast distribution than other common forecast regions" (1). However, to build an accurate HDR, one needs to know (or accurately estimate) the underlying probability distribution. Methods for estimating f such as the kernel density estimator (KDE, 7) or the local likelihood approach (8) work very well for unidimensional problems, but they may be inefficient for multidimensional problems (9). For example, the bandwidth selection in KDE, recognized as the most crucial and difficult step (see e.g., Chapter 2 in 10), has no definite and unique solution. Further, high-dimensional data pose challenges also from the algorithmic/computational perspective. Altogether, these aspects hamper the ability to derive an appropriate HDR.

The scope of this work is to propose an alternative approach to build HDRs using *copulae* so as to overcome the direct estimation of the multivariate f. Specifically, copulae allow to relax the estimation of multivariate random vectors, by separately estimating the marginals and their dependence structure, i.e., the copula model (11). Motivated by the ubiquitous role of copula modelling in modern multivariate statistics, specifically multivariate density estimation, we explore its use the context of HDR estimation.

2. Problem Setting: Highest Density Regions

Assume we have access to a sample $\mathbf{s}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of independent and identically distributed (iid) observations, drawn from a probability measure $\mathbb{P}$. Each data point can be multidimensional, that is $\mathbf{x}_i \in \mathbb{R}^d$, with $d \geq 1$. We denote by $x_i^{(j)}$ the j-th coordinate of $\mathbf{x}_i$, for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. For simplicity, we restrict our analysis to bivariate data points with $d = 2$, and we focus on continuous random variables $\mathbf{X} = (X^{(1)}, X^{(2)}) \in \mathbb{R}^2$. Let $f_{\mathbf{X}}$ denote the probability density function (PDF) of $\mathbf{X}$ and $F_{\mathbf{X}}$ its cumulative density function (CDF). Then, given a coverage probability $1-\alpha$, with $\alpha \in (0, 1)$, the $100(1-\alpha)\%$ HDR is defined as the subset $R(f_\alpha)$ of the sample space of $\mathbf{X}$ such that:

$$R(f_\alpha) \doteq \{\mathbf{x} \colon f_{\mathbf{X}}(\mathbf{x}) \geq f_\alpha\}, \tag{1}$$

where f_α is the largest constant such that $P(\mathbf{X} \in R(f_\alpha)) \geq 1 - \alpha$.

It follows from the definition that the boundary of an HDR consists of those values of the sample space with equal density. Hence a plot of a bivariate HDR is a form of contour plot. One of the most distinctive properties of HDRs is that, of all regions of probability coverage $100(1-\alpha)\%$, the HDR has the smallest region possible in the sample space. Clearly, an HDR always contains the global mode, and in the case of multimodal distributions, it often consists of several disjoint subregions, each containing a local mode. This provides useful information which is "masked" by other types of statistical regions.

To estimate an HDR for $\mathbf{X}$ according to Eq. (1), one needs to know the density function $f_{\mathbf{X}}$. If this is known, the typical way to compute HDRs is the density quantile approach (1), which is based on the following rationale. Let $\mathbf{Y} = f_{\mathbf{X}}(\mathbf{X})$ be the random variable obtained by transforming $\mathbf{X}$ by $f_{\mathbf{X}}$ (bounded and continuous in $\mathbf{x}$). Consider a set of independent observations from the distribution of $\mathbf{X}$, say $\mathbf{s}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. It follows that independent observations from the distribution of $\mathbf{Y}$ can be obtained as $\{f_{\mathbf{X}}(\mathbf{x}_1), \ldots, f_{\mathbf{X}}(\mathbf{x}_m)\}$. Consider now the ordered sample $\{f_{(1)}, \ldots, f_{(m)}\}$ with $f_{(j)}$ the j-th largest of $f_{\mathbf{X}}(\mathbf{x}_i)$, $i = 1, \ldots, m$, so that $f_{(j)}$ is the (j/m) sample quantile of $\mathbf{Y}$. Then, given a constant $\alpha \in [0, 1]$, and denoted with $\lfloor \cdot \rfloor$ the floor operator, we have that:

$$\hat{f}_\alpha \doteq f_{\lfloor \alpha m \rfloor} \to f_\alpha, \quad \text{and} \quad R_m(\hat{f}_\alpha) \doteq \{\mathbf{x} \colon f_{\mathbf{X}}(\mathbf{x}) > \hat{f}_\alpha\} \to R(f_\alpha), \quad \text{as } m \to \infty.$$

Basically, the HDR can be derived based on the sample quantile of $\mathbf{Y} = f_{\mathbf{X}}(\mathbf{X})$.

However, in most real-world scenarios, the density function is unknown. If we have access to a sample of iid observations $\mathbf{s}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, then we can estimate it, and subsequently obtain an estimate of the $100(1-\alpha)\%$ HDR by using the again the density quantile approach with

$$\hat{R}_n(\hat{f}_\alpha) \doteq \{\mathbf{x} \colon \hat{f}(\mathbf{x}) > f_{\lfloor \alpha n \rfloor}\}, \tag{2}$$

where $\hat{f}$ is a possibly consistent estimator of f.

Note that for small n it may not be possible to get a reasonable density estimate. Besides, with few observations and no prior knowledge of the underlying density function, there seems little point in attempting to summarize the sample space. In higher dimensions, the difficulty of selecting an appropriate region is even greater due to the density estimation challenges (see e.g., 9; 10).

3. Using Copulae for Deriving HDRs

A general d-dimensional copula $C\colon [0,1]^d \to [0,1]$ is a joint cumulative density function whose d marginals are uniform over $[0,1]$. Consider, without major loss of generality, the bivariate case $d = 2$, with $F_{\mathbf{X}}$ the joint CDF of the random vector $\mathbf{X} = (X^{(1)}, X^{(2)})$, and $F_{X^{(1)}} = F_1$ and $F_{X^{(2)}} = F_2$ its marginals. Then, it follows from the *probability-integral transform* (12) that the joint distribution of (F_1, F_2) is a copula, say $C_{\mathbf{X}}$, and its expression can be derived by noting that

$$C_{\mathbf{X}}(u^{(1)}, u^{(2)}) = \mathbb{P}(F_1(X^{(1)}) \leq u^{(1)}, F_2(X^{(2)}) \leq u^{(2)}) = F_{\mathbf{X}}(F_1^{-1}(u^{(1)}), F_2^{-1}(u^{(2)})).$$

Letting $u^{(j)} \doteq F_j(x^{(j)})$, $j = 1, 2$, this yields the following result due to Sklar (13):

$$F_{\mathbf{X}}(x^{(1)}, x^{(2)}) = C_{\mathbf{X}}(F_1(x^{(1)}), F_2(x^{(2)})), \quad \forall \mathbf{x} = (x^{(1)}, x^{(2)}) \in \mathbb{R}^2.$$

In summary, we can decompose the bivariate CDF $F_{\mathbf{X}}$ into a composition of the two marginal distribution functions and a two-dimensional copula $C_{\mathbf{X}}$. $C_{\mathbf{X}}$ is the copula of $F_{\mathbf{X}}$ and describes the dependence structure of F_1 and F_2. We refer to (11) for book-length treatment of the foregoing ideas.

In case the bivariate distribution has density f, and if this is available, it holds further that

$$f_{\mathbf{X}}(x^{(1)}, x^{(2)}) = c_{\mathbf{X}}(F_1(x^{(1)}), F_2(x^{(2)}))f_1(x^{(1)})f_2(x^{(2)}),$$

with c being the copula density and f_1 and f_2 the marginal densities. The main advantage of this representation over the one involving the joint PDF is that an estimate of $f_{\mathbf{X}}$ can be obtained by estimating the marginals and the copula density separately, evading potential high-dimensional data challenges (see e.g., 14). Furthermore, copulae offer a flexible framework that can capture complex dependency structures.

If $\hat{c}$ is an estimate of the copula density, we propose to estimate the $100(1-\alpha)\%$ HDR as

$$\hat{R}_n(\hat{f}_\alpha) = \{\mathbf{x}\colon \hat{c}_{\mathbf{X}}(\hat{F}_1(x^{(1)}), \hat{F}_2(x^{(2)}))\hat{f}_1(x^{(1)})\hat{f}_2(x^{(2)}) > f_{\lfloor \alpha n \rfloor}\},$$

with $\hat{f}_j$ and $\hat{F}_j$ consistent estimators of the marginals f_j and F_j, $j = 1, 2$. While here, for the sake of space, we focus on the bivariate case, we emphasize that the approach can be easily extended to higher dimensions, as copulae naturally apply to multidimensional contexts (see e.g., *vine copula* methods; 14).

4. Empirical Evaluation

Simulation studies We start with simulation studies, considering the following four data-generation scenarios, with constant parameters fixed at $\mu_1 = 0, \mu_2 = 1, \sigma_1 = \sigma_2 = 2$, and $w_1 = 1 - w_2 = 0.7$ over an increasing number of sample sizes (from 50 to $10,000$). For copula specifications, we refer to (11).

SC1: Gaussian marginals $\mathcal{N}$ – Gaussian copula C^{Gauss}

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = \mathcal{N}(\mu_2, \sigma_2), \quad C = C^{\text{Gauss}}_{\rho=0.7}$$

SC2: Gaussian $\mathcal{N}$ & Student t marginals – Clayton copula C^{Clay}

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = t_{\nu=10}, \quad C = C^{\text{Clay}}_{\alpha=2}$$

SC3: Gaussian $\mathcal{N}$ & Gaussian mixture marginals – Student t copula C^{t}

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = w_1\mathcal{N}(\mu_2, \sigma_2) + w_2\mathcal{N}(\mu_2 + 10, \sigma_2), \quad C = C^{\text{t}}_{\rho=0.4,\nu=6}$$

SC4: Gaussian $\mathcal{N}$ mixture marginals – Gaussian copula C^{Gauss}

$$f_1 = w_1\mathcal{N}(\mu_1, \sigma_1) + w_2\mathcal{N}(\mu_1 + 10, \sigma_1), \quad f_2 = w_1\mathcal{N}(\mu_2, \sigma_2) + w_3\mathcal{N}(\mu_2 + 10, \sigma_2), \quad C = C^{\text{Gauss}}_{\rho=0.7}.$$

For each scenario, we evaluate the following three methods.

Method1: Direct estimation of the bivariate density We use the nonparametric KDE (7), and consider the asympotically optimal solution proposed in (15) for the bandwidths selection.

Method2: Indirect fully-parametric copula-based estimation of the bivariate density For the estimation of the marginals, we consider the true data-generation processes models (with no misspecification) and maximum likelihood fitting. For the copula model, we perform both model selection (with the AIC criterion) and parameter estimation (with maximum likelihood estimation).

Method3: Indirect fully-nonparametric copula-based estimation of the bivariate density For the marginal densities, we use the standard KDE. For the copula model, we use a KDE approach with the *transformation local likelihood estimator* of (16). We use the R KDECOPULA package, adopting the method with quadratic polynomials and nearest-neighbor bandwidths (17).

To quantify the performance of the methods in the simulation study, we call *positive* those points which should be outside the region and *negative* the others. Let FP, TP, FN and TN be, respectively, the number of false positive, true positive, false negative, and true negative points. Well-established measures of inefficiency are the False Negative (Positive) Rates (FNR and FPR), and the Total Error Rate (ER):

$$\text{FNR} = \frac{FN}{FN+TP}; \quad \text{FPR} = \frac{FP}{FP+TN}; \quad \text{ER} = \frac{FN+FP}{FN+FP+TN+TP}.$$

All methods are evaluated based on $\alpha = 0.05$, that is a coverage probability of 95%.

Results As depicted in Figure 1, the three methods lead to slightly different results, with the two copula-based approaches outperforming the direct KDE (Method1). Compared to the nonparametric Method3, the parametric copula-based approach (Method2) shows the lowest ER across all different scenarios and sample sizes, with an exception for the largest sample sizes, where the difference is negligible.

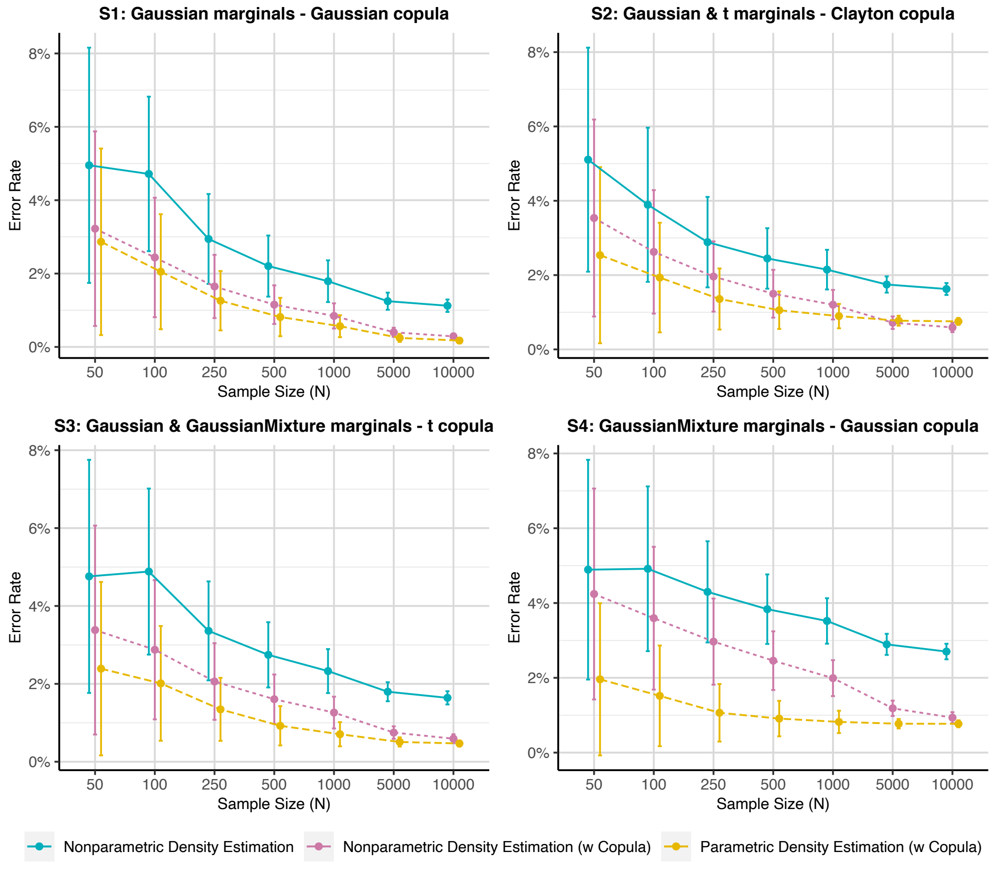

Figure 1: Total error rate (mean and error bars (mean ± SD), averaged across 10^4 MC samples) of the three compared methods across the different scenarios and for different sample sizes.

Looking at the FPR and FNR, results are very similar across different scenarios and we only discuss scenario S2. As displayed in Figure 2, Method1 has the best FPR performance, with a value close to 0. Rather than being a result of optimal performance, this is due to the fact almost or all data points were classified as highest density points, with no *positives* detected. This is also reflected in its FNR, highlighting a low ability to correctly place *positives* outside the HDR. The two copula-based approaches result in overall better performances, with a slight superiority of the parametric Method2. This was expected as simulation schemes consisted of only parametric copula families, and there is no misspecification.

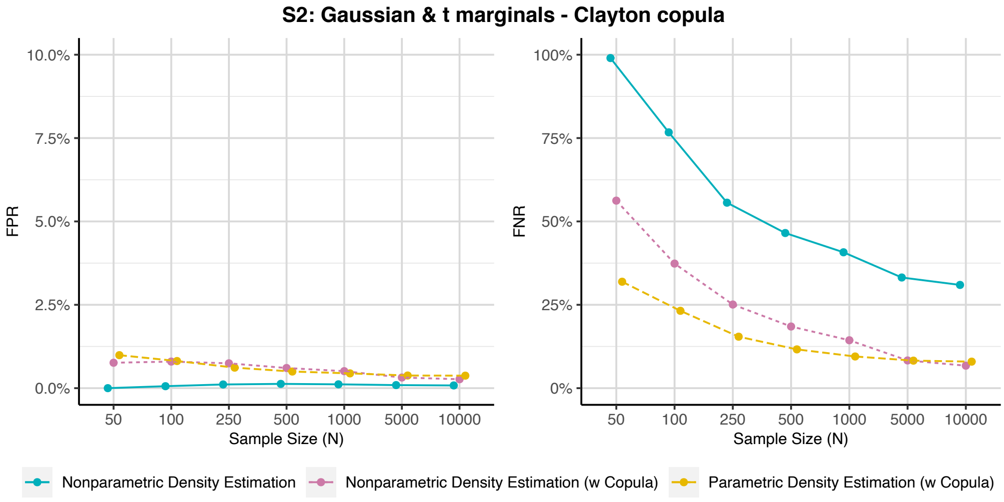

Figure 2: FPR and FNR (averaged across 10^4 MC samples) of the three compared methods for different sample sizes, relatively to scenario S2.

Application to MAGIC Data We apply the proposed methods for constructing a HDR for the joint distribution of two variables from the MAGIC dataset (`https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope`). The data simulate the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope. We focus on gamma observations (overall $n = 12,332$), and consider the two variables “fConc1” and “fM3Long”. In such a case (as deduced from Figure 3), the parametric approach is inappropriate for both the estimation of the marginal distribution and, more importantly, the copula model. Thus, we illustrate the derived HDR using the nonparametric Method1 and Method3 only. While in absence of the underlying truth it is not possible to reliably evaluate the two methods, it seems that the copula-based approach (Method3; right plot), more sensibly excludes the tail data points (which may be expected to have a lower density) from the HDR.

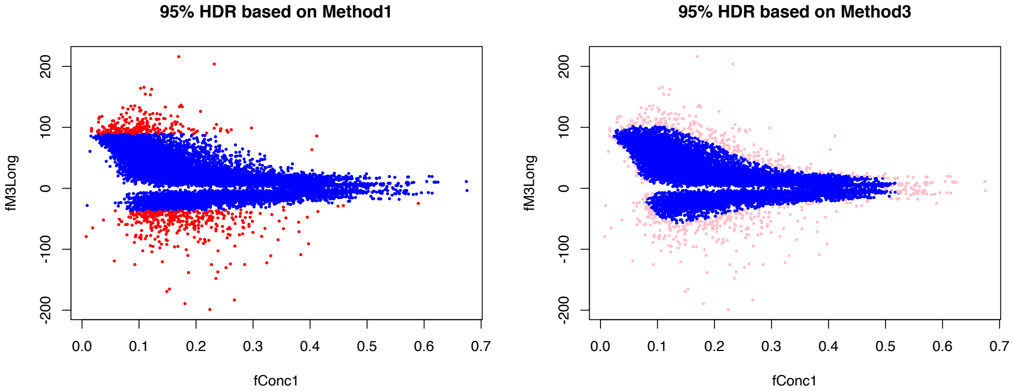

Figure 3: 95% HDR for two variables from the MAGIC dataset with Method1 and Method2.

5. Concluding Remarks

In this work, we proposed an alternative strategy for deriving HDRs in multivariate contexts using *copulae*, and evaluated both a parametric and a nonparametric approach. Compared to traditional kernel density estimation, the copula-based HDR resulted in lower missclasification errors in a number of simulation scenarios and possibly in real data. Although in this work we focused on the bivariate case ($d = 2$), we expect to see remarkable advantages over an increased number of variables $d > 2$. In fact, the extension of the common KDE to high dimensions has proven challenging in terms of both computational efficiency and statistical inference. We aim to pursue such a direction in future work, exploring, e.g., the use of vine copulae to construct flexible dependence models for an arbitrary number of variables using only bivariate building blocks.

References

[1] Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

[2] Jae H Kim, Iain Fraser, and Rob J Hyndman. Improved interval estimation of long run response from a dynamic linear model: A highest density region approach. *Computational Statistics & Data Analysis*, 55(8):2477–2489, 2011.

[3] Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[4] Paula Saavedra-Nieves. Nonparametric estimation of highest density regions for COVID-19. *Journal of Nonparametric Statistics*, 34(3):663–682, 2022.

[5] George EP Box and George C Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, 2011.

[6] Noyan Turkkan and T Pham-Gia. Computation of the highest posterior density interval in Bayesian analysis. *Journal of Statistical Computation and Simulation*, 44(3-4):243–250, 1993.

[7] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[8] Nils Lid Hjort and M Chris Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24(4):1619–1647, 1996.

[9] Han Liu, John Lafferty, and Larry Wasserman. Sparse nonparametric density estimation in high dimensions using the rodeo. In *Artificial Intelligence and Statistics*, pages 283–290. PMLR, 2007.

[10] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

[11] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[12] George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2021.

[13] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[14] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[15] José E. Chacón, Tarn Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21(2):807–840, 2011.

[16] Gery Geenens, Arthur Charpentier, and Davy Paindaveine. Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873, 2017.

[17] Thomas Nagler. kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities. *Journal of Statistical Software*, 84(7):1–22, 2018.