SDS
Statistica e
Data Science

Palermo, 11-12 April 2024

Proceedings of the Statistics and Data Science 2024 Conference

New perspectives on Statistics and Data Science

Edited by

Antonella Plaia – Leonardo Egidi
Antonino Abbruzzo

Proceedings of the SDS 2024 Conference
Palermo, 11-12 April 2024
Edited by: Antonella Plaia – Leonardo Egidi - Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN Ebook (Pdf)
978-88-5509-645-4

Contents

1 Keynote Sessions 11

1.1 Classification with imbalanced data and the (eternal?) struggle between statistics and data science. *Nicola Torelli* 11

1.2 Deep residual networks and differential equations. *Gérard Biau* 13

2 Invited - Complex data: new methodologies and applications . . 15

2.1 Link selection in binary regression models with the Model Confidence Set. *Michele La Rocca and Marcella Niglio and Marialuisa Restaino* 15

2.2 A cluster-weighted model for COVID-19 hospital admissions. *Daniele Spinelli, Paolo Berta, Salvatore Ingrassia and Giorgio Vittadini* 23

2.3 Multi-class text classification of news data. *Maurizio Romano and Maria Paola Priola* 28

3 Invited - Data science and dataspaces: challenges, results, and next steps 35

3.1 Data-Centric AI : A new Frontier emerging in Data Science. *Donato Malerba, Vincenzo Pasquadibisceglie, Vito Recchia and Annalisa Appice* 35

3.2 Data Spaces strategy to unleash agriculture data value: a concrete use case. *Nicola Masi, Delia Milazzo, Giulia Antonucci and Susanna Bonura*42

3.3 Addressing Agricultural Data Management Challenges with the Enhanced TRUE Connector. *Sergio Comella, Delia Milazzo, Mattia Giuseppe Marzano, Giulia Antonucci, Susanna Bonura and Angelo Marguglio* 48

4 Solicited - Data Science for Official Statistics 55

4.1 Data science at Istat for urban green. *Fabrizio De Fausti, Marco Di Zio, Giuseppe Lancioni, Stefano Mugnoli, Alberto Sabbi and Francesco Sisti* 55

4.2 Twitter (X) as a Data Source for Official Statistics: Monitoring Italian Debate on Immigration through Text Analysis. *Elena Catanese, Gerarda Grippo, Francesco Ortame and Maria Clelia Romano* 62

5 Solicited - Sustainable Artificial Intelligence in Finance 69

5.1 Feature Dependence and Prediction Explanations in P2P Lending. *Paolo Pagnottoni and Thanh Thuy Do* 69

6 Solicited - Young SIS 77

6.1 Merging data and historical information via optimal power prior selection in Bayesian models. *Roberto Macrì Demartino, Leonardo Egidi, Nicola Torelli and Ioannis Ntzoufras* 77

6.2 Hierarchical Mixtures of Latent Trait Analyzers with concomitant variables. *Dalila Failli, Bruno Arpino, and Maria Francesca Marino* 84

6.3 A Simultaneous Spectral Clustering for Three-Way Data. *Cinzia Di Nuzzo and Salvatore Ingrassia* 90

7 Solicited - From Data Analysis to Data Science 97

7.1 Optimal Scaling: New Insights Into an Old Problem. *Gilbert Saporta* 97

8 Solicited - Statistical methods for textual data 101

8.1 PROCSIMA: Probability Distribution Clustering Using Similarity Matrix Analysis. *Marco Ortu* 101

8.2 Exploring Anti-Migrant Rhetoric on Italian Social Media. *Lara Fontanella, Annalina Sarra, Emiliano del Gobbo, Alex Cucco and Sara Fontanella* 108

8.3 Causal inference from texts: a random-forest approach. *Chiara Di Maria, Alessandro Albano, Mariangela Sciandra and Antonella Plaia* 114

9 Solicited - Data analysis methods for data in non-Euclidean spaces 121

9.1 Riemannian Statistics for Any Type of Data. *Oldemar Rodrıguez Rojas* 121

9.2 PAM clustering algorithm for ATR-FTIR spectral data selection: an application to multiple sclerosis. *Francesca Condino, Maria Caterina Crocco and Rita Guzzi* 128

9.3 Random Survival Forest for Censored Functional Data. *Giuseppe Loffredo, Elvira Romano and Fabrizio Maturo* 134

9.4 Advancing credit card fraud detection with innovative class partitioning and feature selection technique. *Mohammed Sabri, Antonio Balzanella and Rosanna Verde* 140

10 Solicited - Functional Data Analysis in Action 147

10.1 Functional Linear Discriminant Analysis for Misaligned Surfaces. *Tomas Masak* 147

10.2 Leveraging weighted functional data analysis to estimate earthquake-induced ground motion. *Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio* 155

10.3 Functional autoregressive processes on a spherical domain for global aircraft-based atmospheric measurements. *Almond Stöcker and Alessia Caponera* 161

11 Solicited - Bayesian Inference for Graphical Models 169

11.1 Log-likelihood approximation in Stochastic EM for Multilevel Latent Class Models. *Silvia Columbu, Nicola Piras and Jeroen K. Vermunt* 169

11.2 MCMC Sampling in Bayesian Gaussian Structure Learning. *Antonino Abbruzzo, Nicola Argentino, Reza Mohammadi, Maria De Iorio, Willem van den Boom and Alexandros Beskos* 176

12 Contributed - Promoting Equity: Statistical Insights into Tourism, Sustainability and Digital Divide 183

12.1 Lesson Learnt in the Data Science Worldview: New Dimension of Digital Divide. *Rita Lima* 183

12.2 An overview of Tourism Statistical Literacy. *Yasir Jehan, Giuseppina Lo Mascolo and Stefano De Cantis* 192

12.3 Scalable bootstrap inference via averaged Robbins-Monro approximations. *Giuseppe Alfonzetti and Ruggero Bellio* 198

12.4 The impact of sustainability on Initial Coin Offering: advantages in trading. *Alessandro Bitetto and Paola Cerchiello* 204

13 Contributed - High dimensional and functional data 211

13.1 Analysis of Brain-Body Physiological Rhythm Using Functional Graphical Models. *Rita Fici, Luigi Augugliaro and Ernst C. Wit* 211

13.2 A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects. *Ruggero Bellio and Cristiano Varin* 218

13.3 Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models. *Antonino Gagliano, Gianluca Sottile, Nicolina Sciaraffa, Claudia Coronnello and Luigi Augugliaro* 224

13.4 Investigating the association between high school outcomes and university enrollment choice: a machine learning approach. *Andrea Priulla, Alessandro Albano, Nicoletta D'Angelo and Massimo Attanasio* 230

14 Contributed - Statistical Analysis in economic and market dynamics 237

14.1 A comparison of multi-factor stochastic models for commodity prices C3. *Luca Vincenzo Ballestra, Christian Tezza and Paolo Foschi* 237

14.2 Nonparametric ranking estimation with application to the propensity for Circular Economy of Italian economic sectors. *Stefano Bonnini, Michela Borghesi and Massimiliano Giacalone* 246

14.3 Impact of the Russian invasion of Ukraine on coal markets: Evidence from an event-study approach. *Yana Kostiuk, Paola Cerchiello and Arianna Agosto* 252

14.4 Labour market and time series: a forecast approach for European countries from 1995 to 2022. *Paolo Mariani, Andrea Marletta and Piero Quatto* 258

15 Contributed - Innovations in cluster and latent class models ... 263
15.1 Biclustering of discrete data by extended finite mixtures of latent trait models. *Dalila Failli, Maria Francesca Marino and Francesca Martella* 263
15.2 Seismic events classification through latent class regression models for point processes. *Giada Lo Galbo, Giada Adelfio and Marcello Chiodi* 270
15.3 Determining the optimal number of clusters through Symmetric Non-Negative Matrix Factorization. *Agostino Stavolo, Maria Gabriella Grassia, Marina Marino and Rocco Mazza* 276
16 Contributed - Modelling on spatial phenomena ... 283
16.1 Integrating computational and statistical algorithms in RT-GSCS for spatial survey administration. *Yuri Calleo, Simone Di Zio and Francesco Pilla* 283
16.2 Sensitivity mapping as a tool to support siting of offshore wind farms and increase citizens' acceptability. *Giovanna Cilluffo, Gianluca Sottile, Laura Ciriminna, Geraldina Signa, Agostino Tomasello and Salvatrice Vizzini* 290
16.3 Investigating hotel consumers' purchase intention on web analytics data through PLS-SEM. *Giuseppina Lo Mascolo, Chiara di Maria, Marcello Chiodi and Arabella Mocciaro Li Destri* 296
16.4 Spatio-temporal analysis of lightning point process data in severe storms. *Nicoletta D'Angelo, Milind Sharma, Marco Tarantino and Giada Adelfio*302
17 Contributed - Statistical machine learning for predictive modelling 309
17.1 Application of statistical techniques to predict the effective temperature of young stars. *Marco Tarantino, Loredana Prisinzano and Giada Adelfio* 309
17.2 Topological Attention for Denoising Astronomical Images. *Riccardo Ceccaroni and Pierpaolo Brutti* 316
17.3 LSTM-based Battery Life Prediction in IoT Systems: a proof of concept. *Vanessa Verrina, Andrea Vennera and Annarita Renda* 322
17.4 Predictive modeling of drivers' brake reaction time through machine learning methods. *Alessandro Albano, Giuseppe Salvo and Salvatore Russotto* 328
18 Contributed - Ordinal and preference data analysis . 335
18.1 OSILA (Order Statistics In Large Arrays): an original algorithm for an efficient attainment of the order statistics. *Andrea Cerasa* 335
18.2 The Mallows model with respondents' covariates for the analysis of preference rankings. *Marta Crispino, Lucia Modugno and Cristina Mollica* 343
18.3 Value-Based Predictors of Voting Intentions: An Empirical and Explainable approach. *Luca Pennella and Amin Gino Fabbrucci Barbagli* 349
18.4 A dynamic version of the Massey's rating system with an application in basketball. *Paolo Vidoni and Enrico Bozzo* 355
19 Contributed - Advances in text mining ... 361
19.1 Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?. *Andrea Sciandra and Arjuna Tuzzi* 361

19.2 A Fuzzy Topic Modeling approach to legal corpora. *Antonio Calcagnì and Arjuna Tuzzi* 368
19.3 EmurStat: a digital tool for statistical analysis of emur flow. *Simone Paesano, Maria Gabriella Grassia, Marina Marino, Dario Sacco and Rocco Mazza* 374
19.4 Graph Neural Networks for clustering medical documents. *Vittorio Torri and Francesca Ieva* 380

Preface

The development of large-scale data analysis and statistical learning methods for data science is gaining more and more interest, not only among statisticians, but also among computer scientists, mathematicians, computational physicists, economists, and, in general, all experts in different fields of knowledge who are interested in extracting insight from data. Cross-fertilization between the different scientific communities is becoming crucial for progressing and developing new methods and tools in data science. In this respect, the Statistics & Data Science group of the Italian Statistical Society has organized its 3rd international conference held in Palermo on the 11st and 12nd of April 2024, attended by over 100 researchers from different scientific fields. A collection of the presented papers is available in the present Proceedings showing a huge variety of approaches, methods, and data-driven problems, always tackled according to a rigorous and robust scientific paradigm.

The Statistics & Data Science group

Palermo, April 11st and 12th, 2023

Antonella Plaia - Leonardo Egidi - Antonino Abbruzzo
Editors

13. Contributed - High dimensional and functional data

13.1 - Analysis of Brain-Body Physiological Rhythm Using Functional Graphical Models
13.2 - A comparison of scalable estimation methods for large-scale logistic regression models with crossed random effects
13.3 - Single-cell Sequencing Data: Critical Analysis and Definition of Statistical Models
13.4 - Investigating the association between high school outcomes and university enrolment choice: a machine learning approach

Analysis of Brain–Body Physiological Rhythm Using Functional Graphical Models

Fici R., Augugliaro L. and Wit E.C.

Abstract This paper presents an analysis of physiological data derived from a recent investigation on network physiology, adopting the conceptual framework that views the human organism as a complex network of interacting organs. The study explores coordinated interactions among organs using functional conditional Gaussian Graphical Models (fcGGM). Organ functions are modelled as networks with individual regulatory mechanisms, forming a broader system through continuous interactions. The focus of the analysis is on the interactions within and between two subnetworks: brain activity and a composite network comprising the RR interval of the electrocardiographic waveform, respiration amplitude and blood volume pulse.

Key words: EEG waves, network physiology, Functional Data Analysis, conditional functional Graphical Models

1 Introduction

The paper considers the intricate dynamics of physiological interactions, offering insights into the temporal synchronization of organ functions at the whole organism level, using functional data analysis (FDA). This analysis contributes to exploring new statistical tools to advance our comprehension of physiology interaction, shedding light on the intricate mechanisms that underlie the orchestration of physiological states within the human organism. Adopting a functional framework, the

Rita Fici
University of Palermo, Viale delle Scienze - Building 13, Palermo, e-mail: rita.fici@unipa.it

Luigi Augugliaro
University of Palermo, Viale delle Scienze - Building 13, Palermo, e-mail: luigi.augugliaro@unipa.it

Ernst-Jan Camiel Wit
Università della Svizzera italiana, Via Buffi 13, Lugano, e-mail: wite@usi.ch

vector of records for each variable and each statistical unit represents a discretized observation of a functional variable. Both the response and the explanatory sets of variables constitute two functional processes which take place in an interval of the real line $\mathscr{T} \in \mathrm{R}$ and denoted by $\boldsymbol{Y}(t)$ and $\boldsymbol{X}(t)$, where $t \in \mathscr{T}$. This analysis aims to recover the conditional independent structure of brain waves and the effect of some external covariates on the expected function of each brain wave. These relations can be represented by a graph $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$, where $\mathscr{V}$ is the set of variables given by the response and the explanatory functions, and $\mathscr{E}$ represents underacted edges between functions. We apply the conditional functional Gaussian Graphical model proposed by [1], where the cross-covariance operator is considered as in [4] and the functional processes are assumed to have a partially separable covariance operator defined in [7].

2 Analysis

In this section, we first provide details on the analyzed data. Then, in Section 2.2, we offer a brief overview of the applied model and the estimation procedure used. Lastly, in Section 2.3, the estimation results are presented.

2.1 Brain-Body data

The data utilized in this study originate from a recent investigation on network physiology [3]. Within their conceptual framework, the human organism is viewed as a network where coordinated interactions among organs are essential for generating distinct physiological states and maintaining health. Each organ is considered as a network with its own regulatory mechanisms, continuously interacting with other body organs to contribute to the broader system. Physiological interactions occur at multiple levels across the temporal scale to synchronize organ functions at the whole organism level. Our analysis focuses on the interaction within and between two subnetworks; $\boldsymbol{Y}(t)$ is the network consisting of the α-waves of brain activity, and $\boldsymbol{X}(t)$ describes the activity of three organs, namely: the time interval between two successive R-waves on the electrocardiographic waveform (RR), respiration amplitude, and blood volume pulse (BVP). Brain activity is recorded using the EPOC PLUS wireless headset, consisting of 14 electrodes that record signals from different brain regions. The ECG (sampled at 250 Hz) and respiration signals (sampled at 25 Hz) are acquired through a sensorized t-shirt, while the E4 wristband captures the BVP signal (sampled at 64 Hz) with a photoplethysmographic (PPG) sensor. Each electrode records the amplitude (α) of EEG waves. As described by [8], the data were collected by monitoring eighteen young, healthy volunteers during a measurement protocol consisting of three experimental conditions corresponding to different levels of mental stress. In this study, we analyze the data corresponding to a resting

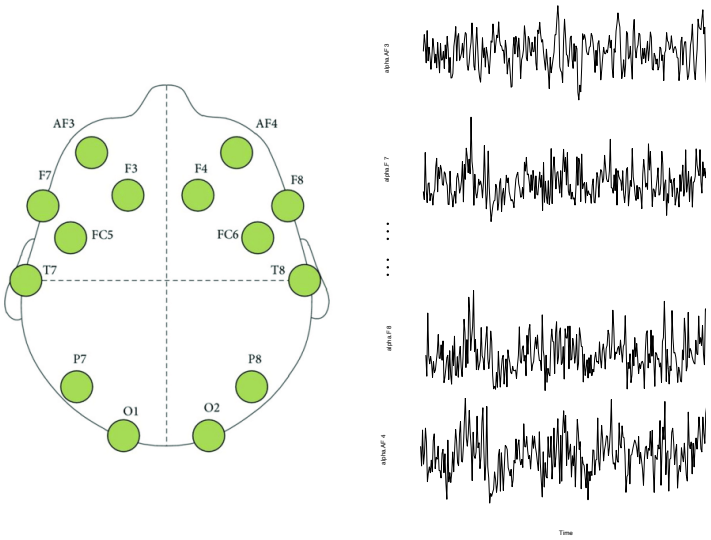

Fig. 1 Left: Graphical representation of the positioning of the 14 EEG electrodes over the scalp.

condition, lasting 12 min and consisting of watching a video showing landscapes with relaxing background music. The data consists of 300 time instants. Figure 1 shows the placement of the EEG electrodes over the scalp and some of the lines designed by the observed 300-dimensional discretized version of the brain waves.

2.2 Functional analysis methods

To recover the edge set representing the relation within and between the two processes, we analyze the relations among the coefficients resulting from the truncated joint vectorized Karhunen-Loéve expansion (see [1]):

$$\boldsymbol{Z}(t) = \sum_{m=1}^{M} \boldsymbol{\zeta}_m \varphi_m(t)$$

where $\boldsymbol{Z}(t) = \left[\boldsymbol{Y}(t)^\top \boldsymbol{X}(t)^\top\right]^\top$, and each element in $\boldsymbol{\zeta}_m$ is equal to $\zeta_{m,i} = \int_{\mathscr{T}} Z_i(t)\varphi_m(t)dt$ where $\{\varphi_m\}_{m=1}^{M}$ is a system of orthonormal basis. To estimate the basis, we perform the eigen-decomposition on the observed matrices of data (see [2]). We analyze $\{\boldsymbol{\zeta}_m\}_{m=1}^{18}$, which constitute 90% of the total variability. Figure 2 shows the kernel estimations of the densities of the scores for $M = 1$ for some of the variables under examination. The assumption of Gaussianity underlying the fcMMs appears quite appropriate for the scores of the explanatory variables. However, an observed asymmetry is noted in the scores of the response variables. To recover the edge-set, we

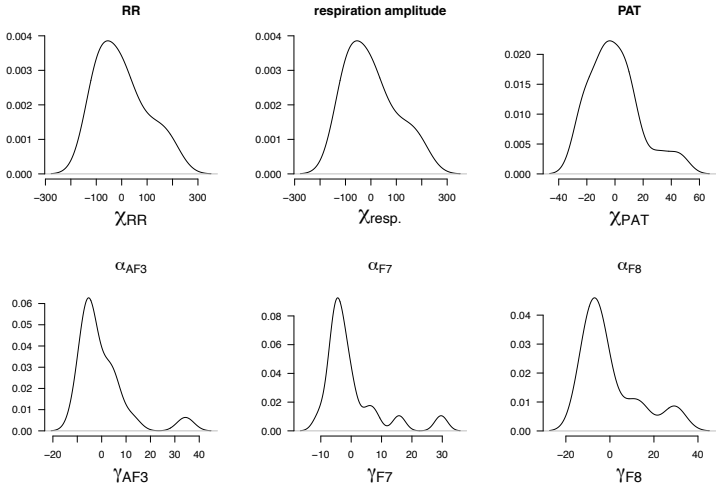

Fig. 2 Estimated densities for the scores of the explanatory variables and some of the response variables.

use the double-penalized joint group lasso estimator proposed by [5]. The optimization problem is performed iteratively, fixing one of the two tuning parameters and varying the other one with 20 values. The tuning parameter value for the next iteration is selected using the joint version of Kullback-Leibler cross-validation (KLCV, [6]) proposed in [1]. The total number of iterations is 3.

2.3 Results

Figure 3 shows the selected links representing the response conditional independence structure. The right side of the scalp exhibits a higher number of links compared to the left side. The most connected alpha wave is $P8$ with 12 links. The waves $T8$ and $F8$ have 11 connection each. Table 1 provides details on the joint KLCV (jKLCV) regressors selection, indicating which explanatory variables have impacts on each of the response variables. The RR result affects all the α waves. The respi-

	AF3	F7	F3	FC5	T7	P7	O1	O2	P8	T8	FC6	F4	F8	AF4
RR	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Resp	1	1	1	1	1	1	1	1	0	1	1	1	1	1
PAT	0	0	0	1	1	1	0	0	0	0	0	0	0	0

Table 1 jKLCV: regressors selection as results of jKLCV selection method.

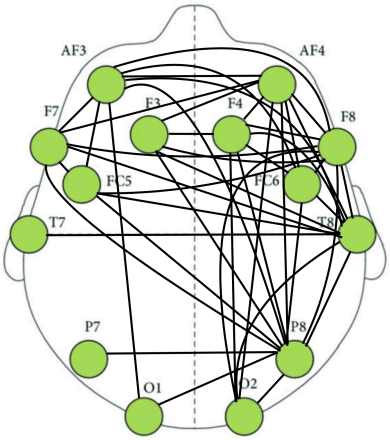

Fig. 3 jKLCV: conditional independent response network.

ration amplitude affects all the response waves except $P8$. Lastly, the blood volume pulse affects the α waves of $FC5, T7$ and $P7$ sensors.

3 Conclusion

The present work aims to contribute to the refinement and development of methods that allow addressing questions related to physiology by adopting a functional point of view.

This study delves into the intricate dynamics of physiological interactions, employing functional data analysis (FDA) to explore the temporal synchronization of organ functions at the whole organism level. By applying the conditional functional Gaussian Graphical model, we aimed to uncover the conditional independent structure of brain waves and the influence of external covariates. The adopted methodology includes the truncated joint vectorized Karhunen-Loève expansion and the double-penalized joint group lasso estimator. Those tools allow us to recover the conditional independence structure and identify influential factors on brain wave activities. The results highlighted the significance of the RR interval and respiration amplitude in shaping the expected value of almost all the α brain waves. The blood volume pulse has a significant effect on some of the sensors.

References

1. Fici, R., Augugliaro L., Wit E.C.: Functional Conditional Gaussian Graphical Models. (2024) doi: 10.48550/arXiv.2401.10196
2. Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators.John Wiley & Sons. (2015) doi: 10.1002/9781118762547
3. Pernice R, Antonacci Y, Zanetti M, Busacca A, Marinazzo D, Faes L, Nollo G.: Multivariate Correlation Measures Reveal Structure and Strength of Brain-Body Physiological Networks at Rest and During Mental Stress. Front Neurosci. Frontiers in Neuroscience. **14** (2021) doi: 10.3389/fnins.2020.602584
4. Qiao, X., Guo, S., James, G.M.: Functional Graphical Models. Journal of the American Statistical Association. **114** (525) pp. 211-222 (2019) doi: 10.1080/01621459.2017.1390466
5. Sottile G., Augugliaro L., Vinciotti V, Arancio W., Coronnello C.: Sparse inference of the human hematopoietic system from heterogeneous and partially observed genomic data. (2022) doi: 10.48550/arXiv.2206.09863
6. Vujačić, I., Abbruzzo, A., Wit E.C.: A computationally fast alternative to cross-validation in penalized Gaussian graphical models. Journal of Statistical Computation and Simulation. **85** (18) pp. 3628-3640 (2015) doi: 10.1080/00949655.2014.992020
7. Zapata J., Oh S.Y., Petersen A.:Partial separability and functional graphical models for multivariate Gaussian processes. Oxford University Press (OUP). **109**(3) pp. 665–681 (2021) doi: 10.1093/biomet/asab046
8. Zanetti, M., Mizumoto, T., Faes L., Fornaser, A., Cecco, M., Maule, L. Valente, M., Nollo, G.: Multilevel assessment of mental stress via network physiology paradigm using consumer wearable devices. Journal of Ambient Intelligence and Humanized Computing. **12** pp. 4409–4418 (2021) doi: 10.3389/fnins.2020.602584