CLADAG - VOC 2025

15th Scientific Meeting of the Classification and Data Analysis Group
1st International Scientific Joint Meeting of the Italian and
Dutch/Flemish Classification Societies

BOOK OF ABSTRACTS

Naples, 8–10 September 2025

Edited by

The Local Organizing Committee

9 788899 594244

Zaccaria Editore, Napoli

Scientific Committee

Michele La Rocca (Chair) (University of Salerno, Italy)
Mark de Rooij (Chair) (Leiden University, The Netherlands)

Andreas Alfons (Erasmus University Rotterdam, The Netherlands)
Zsuzsa Bakk (Leiden University, The Netherlands)
Ruggero Bellio (University of Udine, Italy)
Paula Brito (University of Porto, Portugal)
Carlo Cavicchia (Erasmus University Rotterdam, The Netherlands)
Petros Dellaportas (University College London, UK)
Katrijn van Deun (Tilburg University, The Netherlands)
Jeffrey Durieux (Erasmus University Rotterdam, The Netherlands)
Stefania Fensore (University of Chieti-Pescara "G. d'Annunzio", Italy)
Luca Frigau (University of Cagliari, Italy)
Sylvia Frühwirth-Schnatter (WU – Vienna University of Economics and Business, Austria)
Sabrina Giordano (University of Calabria, Italy)
Claire Gormley (University College Dublin, Ireland)
Leonardo Grilli (University of Florence, Italy)
Jos Hageman (Wageningen University, The Netherlands)
Krzysztof Jajuga (Wroclaw University of Economics and Business, Poland)
Agustin Mayo Iscar (University of Valladolid, Spain)
Mattieu Marbac (ENSAI/CREST Bruz, France)
Volodymyr Melnikov (University of Alabama, USA)
Marta Nai Ruscone (University of Genoa, Italy)
Roberto Rocci (Sapienza University of Rome, Italy)
Kim de Roover (KU Leuven, Belgium)
Laura Sangalli (Polytechnic University of Milan, Italy)
Michel van de Velden (Erasmus University Rotterdam, The Netherlands)
Tim Verdonk (KU Leuven, Belgium)
Adalbert Wilhelm (Constructor University Bremen, Germany)

Local Organizing Committee

Antonio D'Ambrosio (Chair) (University of Naples Federico II, Italy)

Enrico Cafaro (University of Naples Federico II, Italy)
Christian Capezza (University of Naples Federico II, Italy)
Rosanna Cataldo (University of Naples Federico II, Italy)
Luca Coraggio (University of Naples Federico II, Italy)
Anna Crisci (University of Naples Federico II, Italy)
Luca D'Aniello (University of Naples Federico II, Italy)
Rosa Fabbricatore (University of Naples Federico II, Italy)
Agostino Gnasso (University of Naples Federico II, Italy)
Alfonso Iodice D'Enza (University of Naples Federico II, Italy)
Carmela Iorio (University of Naples Federico II, Italy)
Lucio Palazzo (University of Naples Federico II, Italy)
Giuseppe Pandolfo (University of Naples Federico II, Italy)
Alfonso Piscitelli (University of Naples Federico II, Italy)

Massimiliano Politano (University of Naples Federico II, Italy)
Germana Scepi (University of Naples Federico II, Italy)
Rosaria Simone (University of Naples Federico II, Italy)
Maria Spano (University of Naples Federico II, Italy)
Giulia Vannucci (University of Naples Federico II, Italy)

Volunteers

Marco Cardillo (University of Naples Federico II, Italy)
Alessandra D'Alessio (University of Naples Federico II, Italy)
Alessia D'Ambrosio (University of Naples Federico II, Italy)
Giuseppe Gismondi (University of Naples Federico II, Italy)
Fabio Leone (University of Naples Federico II, Italy)
Antonella Meccariello (University of Naples Federico II, Italy)
Rebecca Rivieccio (University of Naples Federico II, Italy)
Daniele Rossi (University of Naples Federico II, Italy)
Luca Ruocco (University of Naples Federico II, Italy)
Marco Russo (University of Naples Federico II, Italy)
Carmine Santone (University of Naples Federico II, Italy)
Maria Tomas (University of Naples Federico II, Italy)
Roberta Accardo (University of Naples Federico II, Italy)

Table of Contents

Preface

This book collects the abstracts presented at CLADAG - VOC 2025, the 15th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS) joint with the Vereniging voor Ordinatie en Classificatie (VOC), the Dutch-Flemish classification society.
The conference held in Naples, Italy, from 8 to 10 September 2025, has been organized by a group of statisticians from the University of Naples Federico II under the auspices of the SIS.
The scientific program of CLADAG – VOC 2025 is noteworthy for the considerable richness of its content. It comprises 5 Keynote Lectures, 41 Invited Sessions promoted by the members of the Scientific Program Committee and 27 Solicited Sessions for a total of 250 talks. We would like to express our sincere gratitude to all the session organizers for inviting speakers of such distinguished repute, who have travelled from many different countries to attend. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts in this book. We would like to express our sincere gratitude all those who collaborated for CLADAG - VOC 2025. Finally, we express our deepest gratitude to all the authors and participants, whose invaluable contributions made this conference a reality.
A selection of the presented papers has already been regularly published in the conference proceedings entitled Supervised and Unsupervised Statistical Data Analysis published by Springer Cham (ISBN: 978-3-032-03042-9).

Naples, September 8, 2025
Antonio D'Ambrosio
on behalf of the CLADAG 2025 Local Organizing Committee

Plenary Talks

The Ewma Control Chart For Real-Time Detection Of Developing Mood Disorders: Key Principles, Optimization Of Performance, And Applications

Authors:

Eva Ceulemans[1*†], Evelien Schat[1], Marieke Schreuder[1]

[1] KU Leuven

[*] Corresponding author [†] Presenter

Contact: eva.ceulemans@kuleuven.be

Keywords:

Real-time, Monitoring, Statistical process control

Abstract:

Statistical methods that can accurately detect early signs of developing mood disorders in intensive longitudinal data (e.g., experience sampling method (ESM) data where people regularly report on their momentary feelings and thoughts using a smartphone app) in real-time are much needed, as such methods would allow to intervene preventively in order to prevent an episode from occurring or to mitigate its severity. Statistical process control (SPC) procedures, originally developed for monitoring production processes, seem promising and statistically sound methods to achieve this goal. SPC procedure capture the natural variation present in a set of in-control data (i.e., the baseline), used to establish control limits. Afterwards, incoming data are compared to the in-control distribution, to detect and test whether and when the incoming data go out-of-control (i.e., when the data go beyond the control limits). We start this talk by introducing SPC and argue why we selected the exponentially weighted moving average (EWMA) method as our SPC method of choice. Next we demonstrate that ESM data violate some crucial EWMA assumptions and discuss how we dealt with these violations by monitoring day averages rather than the individual measurement occasions. This approach of focusing on day statistics also offers a neat solution to the detection of variance changes, by applying EWMA to day statistics of variability. To illustrate the added value of the method, we examine whether the recurrence of depression can accurately be foreseen by applying the EWMA procedure. EWMA results of 41 formerly depressed patients are presented, who were now in remission and discontinuing antidepressant medication. Finally, we look into the baseline problem. One of the biggest challenges of applying the EWMA procedure to ESM data, is the amount of in-control data that is needed for optimal performance, which amounts to at least 50 days. Clearly, it is not trivial to obtain such a large amount of in-control data of a single person. We therefore investigate several potential solutions.

Independent component analysis by robust distance correlation

Authors:

Peter Rousseeuw[1*†], Sarah Leyder[2], Jakob Raymaekers[2]
Tom Van Deuren[2], Tim Verdonck[2]

[1] University of Leuven
[2] University of Antwerp

[*] Corresponding author [†] Presenter

Contact: peter@rousseeuw.net

Keywords:

Algorithms, Dependence measures, Independent variables, Multivariate statistics, Source separation

Abstract:

Independent component analysis (ICA) is a powerful tool for decomposing a multivariate signal or distribution into truly independent sources, not just uncorrelated ones. Unfortunately, most approaches to ICA are not robust against outliers. Here we propose a robust ICA method called RICA, which estimates the components by minimizing a robust measure of dependence between multivariate random variables. The dependence measure used is the distance correlation (dCor). In order to make it more robust we first apply a new transformation called the bowl ransform, which is bounded, one-to-one, continuous, and maps far outliers to points close to the origin. This preserves the crucial property that a zero dCor implies independence. RICA estimates the independent sources sequentially, by looking for the component that has the smallest dCor with the remainder. RICA is strongly consistent and has the usual parametric rate of convergence. Its robustness is investigated by a simulation study, in which it generally outperforms its competitors. The method is illustrated on three applications, including the well-known cocktail party problem.

Fusion Learning: fusing inferences from diverse data sources

Authors:

Regina Y. Liu[1*†]

[1] Department of Statistics, Rutgers University, USA

[*] Corresponding author [†] Presenter

Contact: rliu@stat.rutgers.edu

Keywords:

confidence distribution, data depth, fusion learning, heterogeneous studies

Abstract:

Advanced data acquisition technology has greatly increased the accessibility of complex inferences, based on summary statistics or sample data, from diverse data sources. Fusion learning refers to combining complex inferences from multiple sources to make a more effective overall inference for the parameters of interest. We focus on the tasks: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently? 3) How to combine inferences to enhance an individual study, thus named i-Fusion? We present a general framework for nonparametric and efficient fusion learning. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), developed by combining data depth, bootstrap and confidence distributions. We show that a depth-CD is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing inferential tool. The approach is efficient, general and robust, and readily applies to heterogeneous studies with a broad range of complex settings. The approach is demonstrated with an aviation safety analysis application in tracking aircraft landing performance and a zero-event studies in clinical trials with non-estimable parameters.

Model based clustering for non-continuous time series

Authors:

Dimitris Karlis[1*†]

[1] Athens University of Economics and Business.

[*] Corresponding author [†] Presenter

Contact: karlis@aueb.gr

Keywords:

Discrete valued time series, EM algorithm, Number of clusters

Abstract:

Model-based clustering has attracted significant attention in recent years. Here we focus on the case of time series data specifically on the problem of clustering multiple time series. While various approaches to time series clustering have been proposed in the literature, our emphasis is on model-based methods. We aim to present model-based clustering techniques tailored for time series data, with particular attention to cases with non-continuous observations. Specifically, we address clustering methods for count-valued and categorical time series. Such examples refer to timeline follow back (TLFB) data where students record the number of alcohol drinks per day for some large period and we wish to identify the different drinking patterns. Another example relates to data related to employment status of women, where for each year we have the employment status (unemployment, full time, part tie, students, etc) and we want to cluster the women based on their patterns. Theoretical developments will be illustrated together with real-world examples. We will also discuss computational challenges related to the implementation of these methods.

Directions on directions: a directional data analysis journey

Authors:

Giovanni C. Porzio[1*†]

[1] University of Cassino and Southern Lazio

[*] Corresponding author [†] Presenter

Contact: porzio@unicas.it

Keywords:

Directional data, Statistical manifolds, Hyperspherical analysis

Abstract:

Directional data lie on the circumference of the unit circle or, in higher dimensions, on the surface of hyperspheres. Consequently, their support is a nonlinear and bounded manifold, embedded in an Euclidean space. This, in turn, makes their analysis somewhat special, and the natural way of thinking about data structures changes. For instance, the space of position measures is also a bounded nonlinear manifold; the space of dispersion measures can also be bounded. This presentation aims to introduce this wonderful world, highlighting its specific characteristics and the potential applications of related statistical techniques. The speaker's contribution to the field will also be shared with the audience.

Invited Sessions

Session - Latent variable models for complex dependency structures

Organizers: Roberto Di Mari and Matteo Farnè

Relating Violations Of Measurement Invariance To Group Differences In Response Times

Authors:

Dylan Molenaar[1*†]

[1] University of Amsterdam

[*] Corresponding author [†] Presenter

Contact: d.molenaar@uva.nl

Keywords:

Measurement invariance, Mediation, Response times

Abstract:

Measurement invariance is an assumption underlying the regression of a latent variable on a background variable. It requires the measurement model parameters of the latent variable to be equal across the levels of the background variable. Item specific violations of this assumption are referred to as differential item functioning and are ideally substantively explainable to warrant theoretical valid and meaningful results. Past research has focused on developing statistical approaches to explain differential item functioning effects in terms of item or person specific covariates. In this study we propose a modeling approach that can be used to test if differences in item response times can be used to statistically explain differential item functioning. To this end, in this presentation, a model is studied in which item specific group differences are tested to be due to group differences on a latent response process factor. The properties of the model are investigated in a simulation study. In addition, the model is applied to reveal sex differences in response strategies underlying mathematical ability.

Endogenova: a Latent Variable Approach To Assess Endogenous Bank Performance From Ecb Supervisory Data

Authors:

Roberto Di Mari[1], Matteo Farnè[2*†], Angelos Vouldis[3]

[1] University of Catania
[2] University of Bologna
[3] European Central Bank

[*] Corresponding author [†] Presenter

Contact: matteo.farne2@unibo.it

Keywords:

endogenity, latent variables, model-based clustering, bank business model

Abstract:

In an era where banks and credit institutions are both primary actors of the worldwide financial system as well as essential players of a globalized real economy, one central question is whether we can correctly assess performances of banks, based on (possibly granular) balance sheet information. Extracting bank business models in a data-driven way by a proper clustering algorithm is certainly possible. A natural approach would leverage on balance sheet records to classify banks into distinct business models. Then, based on the classification outcome, one can profile business models over their different performance dimensions. The latter, which we call naive approach, overlooks endogeneity - in terms of reverse causality bias - of investment decisions behind balance sheet records, and performance outcomes. If reverse causality is not taken into account, any estimate of performance outcomes involving business model types will be biased. The classical instrumental variable (IV) approach is certainly a solution, which requires finding at least one good - i.e., relevant and exogenous - instrument. However, finding good instruments is sometimes hardly doable. In this work, we propose an alternative instrument-free strategy, i.e., to model directly the unobserved source of endogeneity by means of a straightforward probabilistic factorization. The empirical premise is as in the naive approach: separating the analysis of balance sheet indicators from the outcome variables. We show that this new approach, which we call endogeNOVA, allows recovering the business model latent class variable while simultaneously identifying the latent driver of endogeneity. The latter can be interpreted as the factor explaining the different performances of each business model, with almost no assumptions on the latent factor distribution. From our specification, we are able to provide unbiased performance measures related to each distinct business model. An extensive simulation study demonstrates that the key feature for separability of the two latent yet related dimensions is dispersion on the endogeneity latent factor. This holds irrespectively of sample size and separation on the latent class variable (i.e., business model type). In the empirical application, we consider nine typical balance sheet aggregates as choice variables, together with two among the most relevant performance indicators, i.e., ROE and ROA. By enforcing a penalization on the conditional cluster-class probabilities, we select the number of latent factors (namely, five) and the specification type (namely, nonlinear) based on BIC. In short, endogenova i) recovers a substantively meaningful four-cluster partition; ii) secures counter-causality effects; iii) demonstrates the association between the two latent dimentions and the observed variables. The meaningfulness of the derived partition is highlighted via a number of tests and models, also by comparison with

the partition derived by exogenova, which is like endogenova without extracting latent factors.

Copula-Based Hidden Semi-Markov Models For Cylindrical Time Series

Authors:

Francesco Lagona[1*†], Marco Mingione[1]

[1] University of Roma Tre

[*] Corresponding author [†] Presenter

Contact: francesco.lagona@uniroma3.it

Keywords:

circular data, copula, hidden semi-Markov model, environmental data

Abstract:

The statistical analysis of bivariate time series that include directional components is inherently different from traditional time series analysis, due to the wraparound nature of their domain and the difficulties in modeling dependence structures when directional measurements are involved. A hidden semi-Markov model is proposed for segmenting cylindrical time series according to a finite number of latent classes, associated with copula-based densities. Observations are modelled by a mixture of cylindrical densities, whose parameters are driven by a latent semi-Markov chain. The proposal integrates tools of directional statistics, used to specify copula-based cylindrical densities, and survival analysis tools, used to model the latent chain. It provides a parsimonious and computationally tractable approach to segment data in nonstandard domains, simultaneously predicting the time spent by the system within each class.

Two-Step Estimation Of Latent Trait Models

Authors:

Jouni Kuha[1*†], Zsuzsa Bakk[2]

[1] London School of Economics and Political Science
[2] Leiden University

[*] Corresponding author [†] Presenter

Contact: j.kuha@lse.ac.uk

Keywords:

Item response theory models, Latent variable models, Structural equation models, Pseudo-maximum likelihood estimation

Abstract:

General latent variable models combine two elements: measurement models for how the latent variables are related to observed measures of them, and structural models for relationships among the latent variables and observed explanatory and response variables for them. Often the structural model is the focus of substantive interest. Estimation of such models can be arranged in different ways. The one-step approach, such as standard full information maximum likelihood estimation, estimates all the parameters together. In alternative stepwise methods the estimation of different elements of the model is split into distinct steps. We describe two-step estimation, where just the measurement model is estimated in the first step and the measurement parameters are then fixed at their estimated values in the second step where the structural model is estimated. This idea can be used with any kinds of latent variable models; for example, it has been successfully developed for different kinds of latent class models. In this talk we show how the two-step approach can be implemented for latent trait models (item response theory models) where the latent variables are continuous and their measurement indicators are categorical variables. Examination of the properties of two-step estimators through simulation studies and applied examples shows that they perform well, and that they have attractive practical and conceptual properties compared to the alternative one-step and three-step methods. These results are in line with previous findings for other families of latent variable models. This provides strong evidence that two-step estimation is a flexible and useful general method of estimation for different types of latent variable models.

Session - Recent developments in finite mixture modeling

Organizer: Volodymir Melnykov

Fitting Gaussian Mixture Models With Uncertain Number Of Components

Authors:

Volodymyr Melnykov[1], Salvatore Ingrassia[2*†]

[1] University of Alabama (USA)
[2] University of Catania (Italy)

[*] Corresponding author [†] Presenter

Contact: s.ingrassia@unict.it

Keywords:

model-based clustering, finite mixture model, model selection criteria

Abstract:

Mixture models constitute one of the most important and fruitful approaches in statistics for data fitting and taking into account uncertainty in data clustering. In this framework, one of the most critical issues in data clustering via mixture modeling concerns the estimation of the number of groups. To this end, usually model selection criteria based on penalized log-likelihood, such as BIC or AIC, are taken into account and the model attaining the smallest value, among a set of candidates, is retained. Anyway, practice shows that sometimes models with different numbers of components may present quite close values of the adopted criterion and therefore crisp approaches can yield ineffective or misleading results. In this framework, we propose to include uncertainty in model selection criteria, in particular as far as the number of mixture components is concerned, in the framework of the maximum likelihood estimation. Throughout this paper this approach is referred to as fuzzy model selection. In other words, rather than selecting automatically a unique model, the idea is to provide a probability distribution on the number of components to be combined with other information coming from the reference domain the data come from, so that the final model results from a combination of statistical evidence and model interpretability. We remark that, with respect to Bayesian framework, here no preliminary information is required. The fuzzy approach will be illustrated on Gaussian mixtures, even though the extension to other kinds of mixtures of distributions is quite straightforward. Essentially, the idea is to consider the number of components of the mixture as a missing value whose distribution is to be estimated together with the parameters of the mixture model; in practice a large parameter space is considered that includes different possible numbers of components. Another issue discussed in the paper concerns the estimation of the number of components in the model in relation to the population the sample is drawn from. This issue will be referred to as “assessing the number of population components" throughout the paper. To this end, we analyze the distribution of the number of components, also referred to as the underlying population, by employing a non-parametric bootstrap sampling of the observed data sample. The proposal is also illustrated on the ground of a series of numerical studies based on both simulated and real data, in particular, we show how this fuzzy approach provides the user with different perspectives for the data clustering.

Mixture-Based Clustering For Ordinal Responses

Authors:

Marta Nai Ruscone[1*†]

[1] University of Genoa

[*] Corresponding author [†] Presenter

Contact: marta.nairuscone@unige.it

Keywords:

Mixture models, ordinal data, EM algorithm

Abstract:

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal data, using finite mixtures to cluster the rows (observations) of the matrix. These models can incorporate the main effects of individual rows and columns, as well as cluster effects, to model the matrix of responses. However, many real-world applications also include available covariates, which can provide insights into the main characteristics of the clusters. In our research, we have extended the mixture-based models to include covariates directly, to allow the clustering structures to be determined both by the individuals' similar patterns of responses and the effects of the covariates on the individuals' responses. We focus on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. We fit the models using the Expectation-Maximization algorithm and assess performance through a comprehensive simulation study. We also illustrate an application of the models.

Consensus, Constrained Parsimonious Gaussian Mixture Models: Labelling Pixels In Hyperspectral Images

Authors:

Ganesh Babu[1], Aoife Gowen[1], Michael Fop[1], Isobel Claire Gormley[1*†]

[1] University College Dublin

[*] Corresponding author [†] Presenter

Contact: Claire.Gormley@ucd.ie

Keywords:

latent variable modelling, high-dimensional data, constrained clustering, spatial dependence

Abstract:

Hyperspectral imaging (HI) is widely used in agriculture, medical diagnostics, and the food industry because of its ability to capture detailed spectral information, allowing identification and characterisation of materials. In the food industry, HI is commonly used to classify different types of food, evaluate quality, or detect contaminants using classification methods. The performance of the classification methods is highly dependent on accurately labelled training images. Typically, k-means clustering or threshold-based methods are commonly employed for pixel labelling, but these are subjective and are not generalisable across different imaging conditions or food types. Here, a suite of parsimonious Gaussian mixture models (PGMMs) are proposed to objectively label pixels in hyperspectral images. The latent variable model approach is employed to cluster the high-dimensional HI data, capturing the underlying structure using a small number of latent factors. Importantly, additionally available information is incorporated in the PGMMs: constraints on some pixels belonging to the same or different clusters are accounted for while clustering the rest of the pixels in an image, or pixel labels are modelled using a Markov random field to account for spatial interactions among neighbouring pixels. To ensure computational feasibility, a consensus clustering approach is also employed, where the data are divided into multiple randomly selected subsets of variables and clustering is applied to each data subset; the clustering results are then consolidated across all data subsets to provide a consensus clustering solution. The suite of proposed PGMMs is applied to simulated and benchmark datasets and to motivating near-infrared hyperspectral images of three types of puffed cereal: corn, rice, and wheat. Improved clustering performance and computational efficiency are demonstrated when compared to other current state-of-the-art approaches. Open-source R code is available to facilitate widespread implementation.

Dealing With Outliers In Model-Based Clustering

Authors:

Katharine Clark[1], Paul McNicholas[2*†]

[1] Trent University
[2] McMaster University

[*] Corresponding author [†] Presenter

Contact: paulmc@mcmaster.ca

Keywords:

Clustering, Mixture models, Outliers

Abstract:

The problem of outliers in model-based clustering is studied in several different scenarios, including data with skewed clusters, matrix-variate data, and functional data. Multiple approaches are considered, including approaches based on the subset likelihoods as well as approaches based on contamination. The considered approaches are illustrated using simulations and real data. The presentation concludes with comments on ongoing and future work.

Session - Complex latent class modeling

Organizers: Zsuzsa Bakk and Johan Lyrvall

A Two-Step Estimator For Growth Mixture Models With Covariates In The Presence Of Direct Effects

Authors:

Yuqi Liu[1*†], Zsuzsa Bakk[1], Ethan M. Mccormick[2], Mark de Rooij[1]

[1] Leiden University
[2] University of Delaware

[*] Corresponding author [†] Presenter

Contact: y.liu@fsw.leidenuniv.nl

Keywords:

Growth mixture model, Two-step estimator, Direct effects, Covariates

Abstract:

Growth mixture models (GMMs) are popular approaches for modeling unobserved population heterogeneity over time. GMMs can be extended with covariates, predicting latent class (LC) membership, the within-class growth trajectories, or both. However, current estimators are sensitive to misspecifications in complex models. We propose extending the two-step estimator for LC models to GMMs, which provides robust estimation against model misspecifications (namely, ignored and overfitted the direct effects) for simpler LC models. We conducted several simulation studies, comparing the performance of the proposed two-step estimator to the commonly used one-step and three-step estimators. Three different population models were considered, including covariates that predicted only the LC membership (I), adding direct effects to the latent intercept (II), or to both growth factors (III). Results show that when predicting LC membership alone, all three estimators are unbiased when the measurement model is strong, with weak measurement model results being more nuanced. Alternatively, when including covariate effects on the growth factors, the two-step, and three-step estimators show consistent robustness against misspecifications with unbiased estimates across simulation conditions while tending to underestimate the standard error estimates while the one-step estimator is most sensitive to misspecifications.

Guided Clustering Variational Autoencoder

Authors:

Christophe Biernacki[1*†], Violaine Courrier[2]

[1] Inria
[2] Withings and Inria

[*] Corresponding author [†] Presenter

Contact: Christophe.Biernacki@inria.fr

Keywords:

Generative model, Clustering, Variational inference

Abstract:

We present the Guided Clustering Variational Autoencoder (GCVAE), a model that addresses the limitations of traditional clustering methods by incorporating additional guiding variables. By combining Variational Autoencoders and Gaussian Mixture Models, GCVAE creates a latent space that captures both the structure of the input data and the context provided by the guiding variables. Unlike many previous approaches, GCVAE uses these additional variables solely in the generative part of the process, ensuring that inference relies only on the input data. The guiding variables can be adjusted to seamlessly reorient the clustering objective to suit various contextual needs, preserving the rich attributes of the original dataset and enabling context-driven partitioning. Experimental results on MNIST-SVHN data and original data from the Withings company demonstrate GCVAE's ability to improve both interpretability and segmentation quality in complex data settings.

Causal Structural Models For Stepwise Latent Class Analysis

Authors:

Felix Clouth[1*†]

[1] Tilburg University

[*] Corresponding author [†] Presenter

Contact: f.j.clouth@tilburguniversity.edu

Keywords:

latent variable models, stepwise latent class analysis, causal inference, g-computation

Abstract:

Two-step and bias-adjusted three-step latent class analysis (LCA) are popular methods for estimating the relationship between latent class membership and covariates or distal outcomes. Causal LCA provides a framework under which these relationships can be interpreted as causal effects. In this talk, I will discuss how causal effects involving latent classes can be formally defined under the potential outcomes framework and what assumptions are required for identifying these effects. Furthermore, I will present g-computation as a flexible method to estimate causal effects in a complex longitudinal setting. G-computation can be used to account for the problem of intermediate confounding. In a longitudinal setting where outcomes, exposures, and confounders can vary over time, intermediate confounding occurs when such time-varying confounders are affected by exposures or outcomes at previous timepoints. That is, time-varying confounders act as intermediate outcomes of past exposure but also as confounders affecting future exposure. Direct adjustment for these post-exposure confounders opens up the causal back-door path between exposure and outcome. G-computation with LCA solves this issue by estimating structural models for each outcome, exposure, and confounder. Potential outcomes, for instance, outcomes under hypothetical scenarios of 'always exposed' or 'never exposed', can then be obtained via micro-simulations that utilize these structural models. Stepwise approaches to LCA allow for estimating the structural models separately from the measurement model. This is not only beneficial for model estimation but also important for the definition of causal effects in the potential outcomes framework.

A Latent Variable Approach For Joint Modeling Of Item Responses, Response Times, And Item Position In Educational Testing

Authors:

Silvia Bacci[1], Rosa Fabbricatore[2*†], Maria Iannario[2]

[1] Università degli Studi di Firenze
[2] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: rosa.fabbricatore@unina.it

Keywords:

educational assessment, response time, item positioning, latent traits, B-GLIRT model

Abstract:

In the context of educational testing, incorporating collateral information beyond item responses can enhance measurement accuracy. In this vein, the use of computer-based testing platforms has facilitated the collection of response time data at the item level, offering new insights into test-taking behavior. Moreover, previous research has identified item position effects, whereby the placement of the same items in different test locations can influence performance due to factors such as fatigue or practice. Finally, since response accuracy may reflect varying degrees of correctness rather than a simple right-or-wrong dichotomy, greater attention should be devoted to its ordinal nature in current modeling approaches. This contribution presents a comprehensive modelling framework that integrates item responses, response times, and item position to better understand skill acquisition and latent speed changes. Our approach builds on the Bivariate Generalized Linear Item Response Theory model, capturing the dual impact of ability on response accuracy and the interplay between ability and speed on response times. The model further explores how item positioning affects test performance and provides diagnostic insights into individual differences. Several respondents' profiles are thus identified, and the effect of covariates on profile membership is explored. The empirical analysis is based on data from first-year Psychology students at the University of Naples Federico II, enrolled in an introductory Statistics course. The assessment included 30 multiple-choice items developed according to the three Dublin descriptors of Knowledge, Application, and Judgement. This study focuses on the 10 items designed to evaluate students' ability to apply knowledge in solving and interpreting problems (Application descriptor). The items were randomly presented via the Moodle platform, which also recorded item-level response times. Students' responses were scored on an ordinal scale: 0 credits for incorrect answers, 1 credit for partially correct answers, and 2 credits for fully correct answers. A set of socio-demographic and psychological factors (e.g., self-efficacy, test anxiety) is incorporated to explain differences in latent profiles. The findings support the effectiveness of our framework in providing a more nuanced and valid assessment of student competencies, offering valuable insights for both test design and formative feedback.

Session - New statistical approaches in life courses studies

Organizers: Dalit Contini and Giancarlo Ragozini

Exploring Student Mobility Trajectories In Higher Education During The Covid-19 Pandemic

Authors:

Vincenzo Giuseppe Genova[1*†], Maria Prosperina Vitale[2], Giuseppe Giordano[2], Giancarlo Ragozini[3]

[1] University of Palermo, Department of Economics, Business, and Statistics
[2] University of Salerno, Department of Political and Social Studies, Salerno, Italy
[3] University of Naples Federico II, Department of Political Sciences, Naples, Italy

[*] Corresponding author [†] Presenter

Contact: vincenzogiuseppe.genova@unipa.it

Keywords:

Educational inequalities, COVID-19 pandemic, Community detection algorithm, Territorial divide

Abstract:

The contribution compares student mobility trajectories in higher education before and during the COVID-19 pandemic, focusing on a regional perspective. Italy is used as a case study, which is characterised by persistent territorial inequalities, with traditional one-way mobility trajectories from Southern regions towards Northern and Central areas. Drawing on cohort data from the Italian Student National Archive, complex network data structures are extracted and analysed through community detection algorithms and regression models to reveal how the pandemic has shifted internal mobility flows, thereby altering long-established patterns. The main findings highlight a change in internal student mobility during the pandemic. Specifically, Northern universities maintained strong attractiveness nationwide, though slightly less than before the pandemic. The analysis also uncovers new mobility patterns towards the Central regions, reflecting the emergence of new educational routes. Furthermore, the persistent attractiveness of STEM programmes is confirmed.

Local Educational Supply And University Choice: Insights From Italy

Authors:

Cristian Usala[1*†], Mariano Porcu[1], Isabella Sulis[1]

[1] University of Cagliari, Department of Social and Political Sciences

[*] Corresponding author [†] Presenter

Contact: cristian.usala@unica.it

Keywords:

university mobility, territorial disparities, administrative data, inner areas

Abstract:

This study explores the relationship between the local supply of educational services and student mobility in the Italian higher education system, with a focus on territorial disparities and internal migration dynamics. Using administrative data from the MOBYSU.IT database on students enrolled in Italian universities, combined with a georeferenced database of over 53,000 schools, we examine how the spatial distribution of secondary education services influences students' decisions to pursue tertiary education outside their home municipalities. Particular attention is given to rural and inner areas, where service centers are sparse and distances from institutions are significant. We measure territorial accessibility to education in terms of the presence of upper secondary schools in students' areas of residence, the travel distance to the nearest institution, and the availability of school tracks and curricula. This information is used in a logistic regression framework to model the probability of students enrolling in a university that is not the closest available option and requires more than one hour of travel, conditional on local educational context and individual characteristics. Results show a strong negative association between the availability of local educational services and student mobility. Students from remote and underserved areas are more likely to bypass nearby institutions and opt for more distant universities, especially in Southern Italy. This suggests that when local access is limited, students prioritize perceived quality and future opportunities over proximity. However, in well-served municipalities, students show similar preferences across regions, with southern students exhibiting a lower probability of mobility compared to peers from other areas. These findings highlight the role of educational infrastructure in shaping mobility patterns and the risk of reinforcing demographic imbalances in marginalized territories. Future research will adopt a latent class logit framework to better classify students by individual and territorial characteristics.

Vaccination Timeliness As a Life Course Process: Patterns And Heterogeneity

Authors:

Chiara Chiavenna[1*†], Alessia Melegaro[2], Danilo Bolano[3]

[1] DONDENA research centre, Bocconi univeristy
[2] Department of Social and Political Science, Bocconi University
[3] University of Florence

[*] Corresponding author [†] Presenter

Contact: chiara.chiavenna@unibocconi.it

Keywords:

Vaccination trajectories, Sequence analysis, Discrepancy analysis

Abstract:

Timely vaccination during infancy is critical to protecting children from severe infectious diseases. However, conventional measures of vaccination uptake-such as coverage at fixed ages or binary indicators of delay-fail to capture the complexity of when and how vaccines are administered. In this work, we reconceptualize paediatric vaccination as a life course process, analysing individual vaccination trajectories through the lens of sequence analysis (SSA) and comparing methodologies for covariate inclusion. We use individual-level electronic health records (EHR) on hexavalent vaccination from over 800,000 children born in Lombardy (Italy) between 2006 and 2019. Each child's monthly vaccination history is encoded as a categorical sequence reflecting the number of doses received over time. This structure captures not only uptake, but also its timing, spacing, and potential interruptions-offering a more nuanced picture than dichotomous outcomes. We compute pairwise dissimilarities and apply hierarchical clustering to identify typical vaccination trajectories, assessing cluster quality using internal validity measures. Importantly, we move beyond descriptive clustering by examining the association between trajectories and sociodemographic factors. First, we use multinomial logistic regression to model how individual characteristics predict cluster membership, revealing structural determinants of adherence and delay. Second, we adapt a framework proposed by Studer et al. (2011), implementing a discrepancy-based approach to evaluate how covariates relate not only to average behaviour but also to within-group variability. Specifically, we compute dissimilarities between each individual and the gravity centre of their covariate group and apply generalised Levene-type tests to assess heterogeneity. This allows us to test whether certain social groups (e.g., children of foreign mothers) follow more diverse vaccination paths-potentially reflecting structural barriers to timely care. We complement these analyses with sliding-window visualisations to track how behavioural variability evolves across early childhood. By applying this methodological framework to the domain of vaccination timeliness, we demonstrate the value of combining sequence analysis, covariate-informed modelling, and discrepancy testing in life course research. This approach deepens our understanding of how behavioural standardisation or fragmentation unfolds across social contexts and offers tools to inform public health strategy.

Session - Safe Machine Learning

Organizer: Paolo Giudici

Safe Financial Time Series Agents

Authors:

Alessandro Piergallini[1*†], Paolo Giudici[1]

[1] University of Pavia

[*] Corresponding author [†] Presenter

Contact: alessandro.piergallini01@universitadipavia.it

Keywords:

Safe machine learning, Time series models, Agents

Abstract:

We address the problem of developing explainable Artificial Intelligence methods to interpret the results of AI models applied to time series data, taking temporal dependencies into account. To this end, we extend the normalized Shapley Lorenz methodology to time series models, including neural networks and recurrent neural networks. We illustrate the approach using Bitcoin prices as the dependent variable and a set of classical financial series as predictors. Our analysis shows that recurrent neural networks outperform classical neural networks in terms of both accuracy and robustness. Bitcoin prices are largely influenced by their own past values, with limited explanatory power from traditional financial assets. Nevertheless, recurrent models effectively capture the contribution of external variables. In addition, we design and implement a modular AI Crew Agent System to automate the econometric analysis of time series using ADL and ARIMAX models. The system consists of multiple agents: Data Loader, Stationarity Agent, Preprocessing Agent, Time Series Modeling Agent, SAFE AI Agent, and Documentation Agent. Once a dataset and a test or train split date are provided, the agents autonomously perform a stationarity analysis using Dickey Fuller tests, apply necessary transformations, select candidate models, and run residual diagnostic checks to eliminate those with autocorrelation issues. The final model is selected based on information criteria, such as AIC, BIC, and HQC, as well as RMSE. SAFE metrics are computed to assess robustness, explainability, and reliability, and a full report is automatically generated. This agent based pipeline enables efficient and transparent AI driven time series forecasting.

An Holistic Trustworthiness Assessment Of Ai Systems: The Safetyvalue

Authors:

Emanuela Raffinetti[1*†]

[1] University of Pavia

[*] Corresponding author [†] Presenter

Contact: emanuela.raffinetti@unipv.it

Keywords:

Machine Learning models, Deep Learning models, Artificial Intelligence, SAFE metrics, unified safety metric

Abstract:

In the last decade, we have witnessed a steady increase in the availability of open data which, together with the growing development of computational power, has contributed to the proliferation of highly complex Machine and Deep Learning models, enhancing the performance of Artificial Intelligence (AI) systems. Artificial Intelligence (AI) is a broad field within computer science focused on developing machines capable of performing tasks that typically require human intelligence. As a result, AI has gained significant relevance, particularly due to its potential to create new opportunities. However, alongside these benefits, AI methods also pose considerable risks. Their black-box nature can lead to automated decision-making processes that may result in incorrect actions due to the lack of transparency and interpretability. The urgent need to ensure the safety of AI systems has led to a growing body of literature. Several proposals, particularly within the statistical framework, have introduced novel approaches and metrics aimed at assessing the key principles of Sustainability (robustness), Accuracy, Fairness, and Explainability. A recent contribution, the SAFE approach, proposes a coherent set of metrics, each designed to independently measure one of the four safety principles. To the best of our knowledge, a unified metric that combines all four dimensions is not yet available in the literature. Within the SAFE framework, models under evaluation may be ranked differently depending on the principle under examination. For instance, a model may perform best in terms of accuracy but poorly in terms of fairness. This scenario leaves model selection at the discretion of regulators or policymakers, depending on the relative importance they assign to each safety principle. To address this gap, we introduce a composite safety index, the SAFEtyvalue, designed to quantify the overall trustworthiness of a Machine or Deep Learning model by summarizing performance across the four key principles. By design, the composite index assigns weights to each safety dimension, with their magnitude potentially reflecting the emphasis placed on each principle by standardisation bodies. The SAFEtyvalue appears as an effective tool to assess how robustness, explainability, and fairness contribute to preserving the original level of predictive accuracy. To further explore the utility of our proposal, we apply it to both white-box and black-box models using simulated data, with the aim of examining how the behavior of the SAFEtyvalue varies under different weight configurations.

The Gini Index As a Multivariate Coefficient Of Variation

Authors:

Gennaro Auricchio[1], Paolo Giudici[2*†], Giuseppe Toscani[1]

[1] University of Padua
[2] University of Pavia

[*] Corresponding author [†] Presenter

Contact: paolo.giudici@unipv.it

Keywords:

Gini Index, Multivariate Coefficient of Variation, Multivariate Analysis

Abstract:

The Gini index and the coefficient of variation are classic measures of sparsity inequality that are well-understood in the one dimensional case. In this talk we show that these two quantities are connected and use this relation to extend both quantities to the multivariate case.

Session - Trimming based robustness

Organizer: Marco Riani

Robust Principal Components By Casewise And Cellwise Weighting

Authors:

Mia Hubert[1*†], Fabio Centofanti[1], Peter Rousseeuw[1]

[1] KU Leuven

[*] Corresponding author [†] Presenter

Contact: mia.hubert@kuleuven.be

Keywords:

Casewise outliers, Cellwise outliers, Iteratively reweighted least squares, Missing values, Principal subspace

Abstract:

Principal component analysis (PCA) is a fundamental tool for analyzing multivariate data. Here the focus is on dimension reduction to the principal subspace, characterized by its projection matrix. The classical principal subspace can be strongly affected by the presence of outliers. Traditional robust approaches consider casewise outliers, that is, cases generated by an unspecified outlier distribution that differs from that of the clean cases. But there may also be cellwise outliers, which are suspicious entries that can occur anywhere in the data matrix. Another common issue is that some cells may be missing. We propose a new robust PCA method, called cellPCA, that can simultaneously deal with casewise outliers, cellwise outliers, and missing cells. Its single objective function combines two robust loss functions, that together mitigate the effect of casewise and cellwise outliers. The objective function is minimized by an iteratively reweighted least squares (IRLS) algorithm. Residual cellmaps and enhanced outlier maps are proposed for outlier detection. The casewise and cellwise influence functions of the principal subspace are derived, and its asymptotic distribution is obtained. Extensive simulations and two real data examples illustrate the performance of cellPCA.

Robust Data Analysis And Clustering Under Heavy Tails

Authors:

Andrea Cerioli[1*†], Luis Angel García-Escudero[2], Agustín Mayo Iscar[2]
Domenico Perrotta[3], Francesca Torti[3]

[1] University of Parma
[2] University of Valladolid
[3] Joint Research Centre (JRC) - European Commission

[*] Corresponding author [†] Presenter

Contact: andrea.cerioli@unipr.it

Keywords:

Generalized radius process, Multivariate Student-t distribution, Robust distance, Outlier detection, Robust clustering

Abstract:

Elliptical heavy-tailed distributions, such as the Student-t distribution, have long been advocated as "robust" models for multivariate data in many fields. The underlying rationale is that robustness should be achieved by letting the classical maximum-likelihood estimators accommodate extreme observations naturally arising from the process under investigation. In that literature "robustness" is then interpreted as a generalization of the light-tailed normal model, which may be possibly unrealistic – and thus "weak" – for the generating process of the available data set. However, there is growing recognition that contamination might also occur under non-Gaussian scenarios, for example in the presence of clusters, heterogeneity or changes of regime that affect the non-Gaussian model assumed to be true for the majority of the data. We thus rely on a robust high-breakdown approach to multivariate data analysis under the Student-t distribution. In our framework the data generating process for the observable p-variate random vector X, with distribution function denoted as F_X, is defined by the contamination model

$$F_X = (1-\varepsilon)F_Y + \varepsilon F_Z, \tag{1}$$

where F_Y is the distribution function of random vector Y representing the postulated null model for the data, while F_Z is the distribution of a contaminant-model component which is usually left unspecified, except for the (often implicit) assumption that the distributions F_Y and F_Z do not overlap excessively, and $\varepsilon \in [0, 0.5)$ is the contamination rate. Recent developments suitable for an elliptical heavy-tail scenario assume that Y is distributed according to the p-variate Student-t law with location parameter vector μ, variance-covariance matrix Σ and $\nu > 2$ degrees of freedom. The main goal of the present work is to extend the heavy-tail version of model (1) to a clustering framework, where F_Y becomes itself a mixture of the postulated p-variate Student-t laws. Our robust clustering approach is based on trimming and generalizes the well-known TCLUST method developed for the case where F_Y in (1) is a Gaussian mixture. A crucial ingredient of our approach is the development of a suitable EM algorithm for the multipopulation and heavy-tail version of model (1). The proposed algorithm combines the usual constraints on the cluster-specific covariance matrices with trimming in the maximization step, computation of the appropriate consistency factor for the covariance estimates under the p-variate Student-t law and a further loop to estimate the usually unknown values of ν and ε. The possibility of improving the efficiency of the final cluster solution through reweighting is also investigated.

The Use Of Modern Robust Regression Analysis With Graphics: An Example From Marketing

Authors:

Anthony C. Atkinson[1], Marco Riani[2], Gianluca Morelli[2]
Aldo Corbellini[2*†]

[1] London School of Economics
[2] Università degli Studi di Parma

[*] Corresponding author [†] Presenter

Contact: aldo.corbellini@unipr.it

Keywords:

Box-Cox transformation, Generalized Additive Model (GAM), Forward Search, AVAS

Abstract:

Least squares (LS) regression, usually summarized through analysis of variance tables, may miss critical data characteristics. This paper demonstrates this by analyzing 1171 customer loyalty observations, examining the relationship between loyalty and factors like price and community outreach. We advocate for and demonstrate modern robust statistical tools, emphasizing graphical data interaction. Initial LS regression analysis shows significant relationships across all factors. However, diagnostic plots indicate that unexplained features still exist even after response transformation, suggesting data contamination. Consequently, a robust analysis employing a non-parametric model is proposed that significantly enhances the importance of transformations of explanatory variables: these transformations offer deeper insights into consumer behaviour, overcoming the limitations of non-robust methods. Building on a historical overview of regression, we detail our analysis process. First a conventional LS analysis is performed, followed by an advanced dynamic methods like the Forward Search, (Atkinson and Riani, 2020) which highlight the inadequacy of the simple LS model. We then apply a GLM, a method capable of emphasizing both robustness in data analysis and incorporating diverse graphical methods. A key extension involves non-parametric regression, including the AVAS (Tibshirani et al, 1988) a robust procedure for transforming both explanatory and response variables to achieve constant error variance, yielding an improved adjusted R^2 of 0.790. Despite the application of all these robust techniques, some outliers still persist. To address this, we introduce RAVAS, (Riani et al, 2023) a robust extension of AVAS, providing resilience against outliers. This robust analysis results in an even better-fitting model with clearly defined residuals and sharp conclusions regarding variable significance. The paper further provides a marketing-oriented interpretation of the transformed explanatory variables. Summarizing, we propose a structured approach to robust regression analysis, offering a checklist for comprehensive data analyses, moving beyond sole reliance on p-value tables. Links to the public software used are provided, encouraging the routine adoption of these powerful tools.

Session - Scalable estimation of large-scale models

Organizer: Ruggero Bellio

Hierarchical Item Response Theory

Authors:

Omiros Papaspiliopoulos[1*†], Max Goplerud[2]

[1] Bocconi
[2] Austin

[*] Corresponding author [†] Presenter

Contact: omiros@unibocconi.it

Keywords:

variational inference, mixed models, latent factors, large scale inference, saddlepoints

Abstract:

The talk relates to a book we are currently on variational inference and its applications to large scale inference for mixed models. In this talk I will discuss an important special case of the models we consider and involves random effects with multiplicative interactions. This model structure is pervasive throughout applied sciences: within political science is known as item response theory, within other parts of social sciences as factor models, within machine learning as matrix factorization. The type (and "density") of data differ across application areas and this has implications on current practices, existing theory and practical priorities. Nevertheless, there are important synergies. In this talk I will provide a synthesis of different results in this broad field, and I will present our work, motivated primarily by understanding political ideology using roll call data, on developing variational inference methods for such models, and some surprising challenges they arise. Among other things, our models learn an ANOVA-type decomposition of latent ideology. To put things in perspective and see why variational inference is a useful tool in this context, we train models of hundreds of thousands of parameters on millions of observations.

Fast M-Estimation For Exploratory Generalized Linear Latent Variable Models In High Dimensions

Authors:

Maria-Pia Victoria Feser[1*†], Giuseppe Alfonzetti[2], Stephane Guerrier[3]

[1] Department of Statistical Sciences, University of Bologna
[2] Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di UDINE
[3] Faculty of Sciences, University of Geneva

[*] Corresponding author [†] Presenter

Contact: maria.victoriafeser@unibo.it

Keywords:

factor analysis, maximum likelihood, joint likelihood inference, stochastic approximation algorithm

Abstract:

Dimension reduction for High Dimensional and Low Sample Size n (HDLSS) data settings is an important and challenging task, relevant to both machine learning and statistical applications. An example concerns the study of the gut microbiome, i.e. the collection of all the microorganisms (e.g. bacteria and archae) found in millions, that continuously communicate with the cells of the body through chemical signals. This type of data is typically discrete or in counts, as well as HDLSS. As a probabilistic alternative to matrix factorization when the data are of mixed types, whether discrete, continuous, we consider here Generalized Linear Latent Variable Models (GLLVMs). They achieve the reduction of dimensionality (p) by mapping the correlated multivariate data to so-called (q) latent variables, defined in a lower-dimensional space. The benefit of GLLVMs is twofold: the latent variables can be estimated and used as features to be embedded in another model, and the model parameters themselves are interpretable and provide meaningful indications on the very structure of the data. However, with a marginal likelihood based approach, GLLVM's estimation represents a tremendous challenge for even moderately large dimensions, essentially due to the multiple integrals involved in the likelihood function. Numerous methods based on approximations of this latter have been proposed: Laplace approximation, adaptive quadrature, or, recently, extended variational approximation. For GLLVMs, however, these methods do not scale well to high dimensions, and they may also introduce a large bias in the estimates. We propose instead an M-estimator, based on the profiled likelihood, which has a negligible efficiency loss compared to the (exact) marginal MLE. For large data sets in p, the proposed M-estimator, whose computational burden is linear in npq, remains applicable when the state-of-the-art likelihood based methods fail to converge, for example when p>n. It is different to alternative penalized estimators that can suffer from severe finite-sample biases. To compute the M-estimator, we propose to use a stochastic approximation algorithm, and it is used to analyse a dataset consisting of bacterial abundance clustered in 4707 operational taxonomic units extracted from faecal samples from 27 Hadza hunters in Tanzania and 16 Italian adults in Bologna. Our proposed method allows for summarising microbiota variation between the two groups in a low-dimensional and interpretable latent space.

Scalable Composite Likelihood Estimation Of Categorical Data Models With Crossed Random Effects

Authors:

Giuseppe Alfonzetti[1*†], Ruggero Bellio[2], Cristiano Varin[3]

[1] Università degli studi di Udine
[2] University of Udine
[3] Ca' Foscari University of Venice

[*] Corresponding author [†] Presenter

Contact: giuseppe.alfonzetti@uniud.it

Keywords:

categorical data, cross random effects, composite likelihood, stochastic approximations, scalable inference

Abstract:

Likelihood inference in mixed effects models for categorical responses with crossed random effects is notoriously challenging, as it requires integrating over the joint distribution of the random effects. For large-scale datasets, the dimension of the integrals may be so large to prevent the numerical computation of the maximum likelihood estimator. Several proposals exploit the concept of composite likelihood to address the numerical difficulties associated with likelihood inference for categorical data. A first proposal is the all-row-column method, which was recently proposed for binary or ordinal responses with a probit link. The all-row-column method exploits the closure propriety of the normal distribution to obtain scalable and provable consistent estimation in crossed-random effects models. Another composite likelihood method considers the pairwise likelihood built with all pairs of observations that share a random effect. The pairwise likelihood approach is rather general and provides consistent estimation as well, but it is not scalable since the number of pairs of correlated observations typically grows super-linearly with the sample size. In this contribution, we apply recent results on stochastic approximation for composite likelihoods to pairwise likelihood estimation for crossed random effects. The key idea is to implement a suitable sampling scheme to the reservoir of possible pairs of correlated observations, iterating towards the maximum pairwise likelihood estimate using a stochastic gradient update. The resulting estimator scales well and inherits the inferential properties of maximum pairwise likelihood estimation, although some care is required to evaluate the estimation variance in large settings. During the presentation, we will compare the all-row-column method, a Gaussian variational approximation and the proposed stochastic pairwise likelihood method in probit ordinal models for recommender systems. Furthermore, we will illustrate the stochastic pairwise likelihood method in logistic regression with crossed random effects, because all-row-column is currently not available with the logit link.

Boosting Strategies Of Stochastic Optimisation For High-Dimensional Latent Variable Models

Authors:

Motonori Oka[1*†], Yunxiao Chen[1], Irini Moustaki[1]

[1] London School of Economics and Political Science

[*] Corresponding author [†] Presenter

Contact: m.oka1@lse.ac.uk

Keywords:

Langevin diffusion, stochastic approximation, minibatch gradient, Markov chain Monte Carlo, marginal likelihood, empirical Bayes

Abstract:

Latent variable models are widely used in social and behavioural sciences, such as education, psychology, and political science. In recent years, high-dimensional latent variable models have become increasingly common for analysing large and complex data. Estimating high-dimensional latent variable models using marginal maximum likelihood is computationally demanding due to the complexity of integrals involved. To address this challenge, stochastic optimisation, which combines stochastic approximation and sampling techniques, has been shown to be effective. This method iterates between two steps – (1) sampling the latent variables from their posterior distribution based on the current parameter estimate, and (2) updating the model parameters using an approximate stochastic gradient constructed from the latent variable samples. In this paper, we investigate the strategies for boosting the performance of stochastic optimisation algorithms for high-dimensional latent variable models. The improvement of the stochastic optimisation algorithms is achieved mainly through the following three strategies: the Metropolis-adjusted Langevin (MALA) sampler that uses the gradient of the negative complete-data log-likelihood for the efficient posterior sampling of latent variables; the minibatch gradient method that uses a minibatch observations when sampling latent variables and constructing stochastic gradients; and the use of the second-order information of the objective function in the construction of the stochastic gradient. Our simulation investigations revealed that the algorithm with the MALA sampler and minibatch gradient method converges fastest in the beginning; however, in the end, the algorithm with the MALA sampler and fullbatch gradient method involving the second-order information converges best in terms of measure of accuracy. Furthermore, the aforementioned findings motivate us to propose the novel approach in constructing the stochastic optimisation algorithms, combining the MALA-sampler-based minibatch gradient algorithm and its fullbatch variant empowered with the second-order information. The utility of this novel, combined approach was confirmed via the additional simulation study. The proposed algorithm based on the combined approach is also shown to scale well to high-dimensional settings through simulation studies and a personality test application with 30,000 respondents, 300 items, and 30 latent dimensions.

Session - Biclustering: methodologies and applications

Organizer: Jos Hageman

A Genetic Algorithm Approach For Biclustering Diverse Structural Components In Complex Data

Authors:

Jos Hageman[1*†], Guus Nellissen[2]

[1] Wageningen University
[2] Statistics Netherlands (CBS)

[*] Corresponding author [†] Presenter

Contact: jos.hageman@wur.nl

Keywords:

Biclustering, Simplivariate Components, Genetic Algorithm, High-dimensional data analysis

Abstract:

As datasets continue to increase in size and complexity, identifying meaningful structure within high-dimensional data becomes more challenging. Simplivariate models aim to address this by detecting Simplivariate Components (SCs); submatrices within the data that involve only subsets of rows and columns and correspond to potentially interpretable structures, such as biological or clinical patterns. Previous work has typically focused on detecting either additive or multiplicative SCs, often requiring a priori assumptions about the model form. In this study, we extend the Simplivariate framework by introducing a more flexible and general approach. Using a modified Genetic Algorithm (GA) representation, our method allows for the simultaneous discovery of constant, additive, and multiplicative SCs, including both up- and downregulated patterns. It supports predefined model selection, uses F-tests to determine SC type, and assigns significance levels to facilitate interpretation in terms of explained variance. This classification step ensures that the selected SCs are not only internally coherent but also statistically meaningful. We benchmarked our approach through a simulation study using synthetic data containing embedded structure. Performance was compared with three other biclustering algorithms, under various noise and signal settings. Our method consistently recovered the target structure with high reliability, especially in scenarios where multiple types of signal coexisted. We further applied the method to several real-world datasets to assess practical utility. These analyses confirm that our approach can reveal relevant structure in complex data where conventional clustering or biclustering techniques fall short. The method is implemented in R and designed for extensibility. Future developments will focus on allowing user-defined weights and structural constraints to incorporate prior knowledge more directly into the modeling process. In doing so, we hope to improve both sensitivity and interpretability in applied contexts.

A Mixture Of Multivariate Poisson Lognormal Distributions And Its Extension To Biclustering

Authors:

Sanjeena Dang[1*†]

[1] Carleton University

[*] Corresponding author [†] Presenter

Contact: sanjeena.dang@carleton.ca

Keywords:

Multivariate Poisson-lognormal distribution, Multivariate count data, Biclustering, RNA-sequence data

Abstract:

Multivariate count data are commonly encountered through high-throughput sequencing technologies in bioinformatics. Although the Poisson and negative binomial distributions are routinely used to model these count data, their multivariate extensions are computationally expensive, thus restricting their use to small-dimensional datasets. Hence, independence between genes is assumed in most cases, and this fails to take into account the correlation between genes. Fitting such univariate models for multivariate analysis is not only biologically inappropriate, misspecifying the correlation (covariance) structure can result in poor fit to the data. Recently, we developed mixtures of multivariate Poisson lognormal (MPLN) models to analyze these multivariate count measurements. In the MPLN model, the observed counts, conditional on the latent variable, are modelled using a Poisson distribution, and the latent variable comes from a multivariate Gaussian distribution. Due to this hierarchical structure, the MPLN model can account for over-dispersion and allows for correlation between the variables. We will show that the univariate version of MPLN provides a similar fit to the widely used negative binomial distribution in terms of capturing the mean-variance trends of the RNA-seq data, and the multivariate version has some attractive characteristics. Moreover, we developed a computationally efficient framework for parameter estimation for MPLN models that utilizes variational Gaussian approximations. Building on this framework, a recent development of the MPLN mixture model-based biclustering approach that utilizes a block-diagonal covariance structure to allow for a more flexible structure of the covariance matrix to cluster modern biological data will be discussed.

An Innovative Approach To Co-Clustering Of Directional Data: a Methodological Framework With An Application On Interregional Mobility In The Italian National Health Care System

Authors:

Alessia D'Ambrosio[1], Giuseppe Gismondi[2], Marco Cardillo[3]
Giuseppe Pandolfo[2*†], Antonio D'Ambrosio[2]

[1] Department of Physics 'Ettore Pancini', University of Naples Federico II, Italy
[2] Department of Economics and Statistics, University of Naples Federico II, Italy
[3] Department of Agricultural Science, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: giuseppe.pandolfo@unina.it

Keywords:

Co-clustering, Directional Data, Data Depth, Healthcare Mobility

Abstract:

The aim of this work is to propose an innovative methodology for the co-clustering of transition tables expressed as directional data. The clustering method used is the Depth-Based Medoids Clustering Algorithm (DBMCA), and the evaluation of the quality of the cluster is carried out using the k-elbow method, duly adapted taking into account that the data represent values of angular depth. Transition tables are square asymmetric matrices of frequencies in which the rows and columns represent the same objects. The motivating example concerns particular frequency tables within the Italian national health care system, in which the objects in rows and columns are the Italian regions of residence and hospitalization, respectively. The distances between the clusters of the regions of hospitalization and residence highlight differences and similarities in healthcare mobility patterns, allowing for the identification of relationships between flows and regional behaviors. This approach provides a useful overview for understanding hospitalization dynamics and the connections between the analyzed territories.

Session - Advanced clustering methods for complex data I

Organizer: Marta Nai Ruscone

Community Detection In Financial Networks And The Importance Of Being Robust

Authors:

Pietro Coretto[1*†], Alfonso Peluso[2], Luca Coraggio[3]

[1] Università di Salerno
[2] Univeristy of Salerno
[3] Univerisity of Naples "Federico II

[*] Corresponding author [†] Presenter

Contact: pcoretto@unisa.it

Keywords:

Community Detection, Clustering, Network, Stock markets, Finance

Abstract:

Stock market networks represent the interconnected relationships between stocks or financial instruments. By studying these networks, researchers can gain insights into complex market dynamics and assess the impact of specific stocks on overall market behavior. The identification of stock networks has evolved significantly over the past few decades, primarily driven by physicists who have applied complex systems analysis to financial markets. This study advances the methodology for analyzing stock networks in two key directions. First, we introduce robust methods at each step of the analysis. Traditional approaches to constructing these networks rely on empirical correlation matrices and OLS-based methods to filter out variations driven by market trends. We replace these classical methods with robust alternatives capable of mitigating the well-documented influence of exceptional variations in stock returns that frequently occur during market shocks and sharp transitions between dynamic regimes. Second, we exploit the information gain provided by high-frequency data and demonstrate that high-frequency networks can reveal previously hidden market structures. We provide empirical evidence that networks reconstructed using robust methodologies and high-frequency data lead to the discovery of more compact clusters that exhibit greater stability and show improved alignment with economic sectors compared to traditional approaches. This enhanced methodology offers significant potential for advancing financial network analysis and improving our understanding of market structure dynamics.

Mixtures Of Dirichlet Distributions For Clustering Dynamic Compositional Data

Authors:

Igor Melnykov[1*†], Szymon Steczek[1]

[1] University of Minnesota Duluth

[*] Corresponding author [†] Presenter

Contact: imelnyko@d.umn.edu

Keywords:

clustering, finite mixture models, compositional data, Dirichlet distributions

Abstract:

We consider the modeling of population proportions with finite mixtures of Dirichlet distributions in a time-sensitive setting. The compositional data of interest to us are observed in blocks or sequences of points so that all observations in a block need to be classified together in the resulting clustering solution. The example motivating our model comes from the United Nations' demographic database that records the proportions of single, married, widowed, etc., participants at different ages for over 200 countries. For each country, the changing proportions basically form a curve or trajectory in the compositional data simplex. The clustering of these curves is complicated by the fact that their endpoints do not necessarily match and the proportions along the curve may be recorded at different ages for different countries. Previous research focused only on hierarchical and k-medoids clustering of such data. Our methodology uses a model-based approach that relies on the parameterization of the Dirichlet distribution exponents with respect to age. This parameterization was successfully carried out with the use of logistic functions and allowed to distinguish several patterns that exist among the countries when it comes to changes in marital status at different ages. Conducting separate analyses by gender, we observe that the most pronounced split appears between the countries with more traditional gender roles versus those where respondents of both genders tend to get married later in life. Another clear distinction emerges between the countries with lower versus higher proportions of married respondents. Using the Bayesian Information Criterion, we were able to distinguish four clusters in the data obtained from female respondents. The data obtained from males showed a smaller variety of patterns with only three clusters that were detected.

Model-Based Clustering And Variable Selection For Multivariate Count Data

Authors:

Thomas Brendan Murphy[1*†], Julien Jacques[2]

[1] University College Dublin
[2] Université Lumière Lyon 2

[*] Corresponding author [†] Presenter

Contact: brendan.murphy@ucd.ie

Keywords:

Clustering, variable selection, Count data

Abstract:

Model-based clustering provides a principled way of developing clustering methods. We develop a new model-based clustering methods for count data. The method combines clustering and variable selection for improved clustering. The method is based on conditionally independent Poisson mixture models and Poisson generalized linear models. The method is demonstrated on simulated data and data from an ultra running race, where the method yields excellent clustering and variable selection performance.

Mixed-Type Fuzzy Spectral Clustering With Kernel-Based Similarity

Authors:

John R.J. Thompson[1*†], Jesse S. Ghashti[1], Daniel Krasnov[2]
Nicolas Bosteels[2]

[1] The University of British Columbia
[2] McGill University

[*] Corresponding author [†] Presenter

Contact: john.thompson@ubc.ca

Keywords:

fuzzy clustering, spectral clustering, mixed-type data, kernels, bandwidth selection, distance metric learning, reinforcement learning, bootstrap

Abstract:

Fuzzy clustering algorithms, such as the popular fuzzy C-means algorithm, extend hard clustering to allow for a degree of uncertainty in the cluster assignments through a fuzzifying parameter in the objective function. However, challenges remain in selecting the fuzzy parameter, incorporating mixed continuous and categorical data types (called mixed-type data), handling nonlinear cluster shapes and boundaries, and balancing variables based on their importance to fuzzy clustering. To address these challenges, we extend spectral clustering to the fuzzy setting through a fuzzy C-means. We propose using a mixed-type kernel similarity function, where bandwidth selection controls variable importance and reduces the influence of selecting a fuzzy parameter. We further propose extending the fuzzy C-means algorithm to incorporate mixed-type kernel distance metrics into the objective function. We compare to fuzzifying spectral clustering through a bootstrap k-means procedure that requires no specification of a fuzzifying parameter. Throughout this talk, we will explore the influence of distance metric learning on fuzzy clustering algorithms, with a particular focus on the relationship between the kernel bandwidths and the cluster fuzzifying parameter. For bandwidth selection methods, we consider several methods, including an iterative locally-adaptive eigengap maximization method and a reinforcement learning-based genetic algorithm. We apply these methods in fuzzy clustering to simulated and real continuous-only and mixed-type benchmark datasets to evaluate clustering performance in comparison to current fuzzy clustering algorithms. We find that the bandwidth controls variable importance and mitigates the influence of selecting a fuzzy parameter. We find that fuzzy spectral clustering with kernel-based similarity provides intuitive fuzzy boundaries between clusters without the a priori specification of cluster shapes.

Session - Data science applications for environmental quality control and sustainability

Organizer: Monica Palma

Characteristic-Based Fuzzy Clustering Of Mcs-Garch Volatility Components In Traffic Flow Data

Authors:

Rodolfo Metulini[1*†], Maurizio Carpita[2], Manlio Migliorati[2]

[1] University of Bergamo
[2] University of Brescia

[*] Corresponding author [†] Presenter

Contact: rodolfo.metulini@unibg.it

Keywords:

time series analysis, sustainable mobility, urban dynamics, dimensionality reduction

Abstract:

Understanding and forecasting human mobility in city areas is critical for urban planning, risk management, environmental quality control, and sustainability. Managing mobility can directly influence air pollution, energy use, and the effectiveness of interventions aimed at improving urban livability. In previous work, we applied a trivariate Vector AutoRegressive model with eXplanatories and Dynamic Harmonic Regression (VARX + DHR) to one year of hourly mobile phone traffic data related to specific zones of Brescia, which displays strong evidence of multiple seasonality. While forecasting accuracy was reasonable, the model's residuals exhibited heavy tails and time-varying heteroskedasticity, pointing to unmodeled volatility. In this study, we address this issue by modelling residual volatility using Multiplicative Component Standard Generalized AutoRegressive Conditional Heteroscedasticity (MCS - GARCH) approach, decomposing it into three components: daily, hourly, and intraday volatility. All components are then used in a fuzzy clustering strategy, enabling the grouping of areas with similar traffic behaviors even when clear-cut cluster boundaries are absent, thus capturing the overlapping and non-discrete structure of urban mobility. To determine the list of features, we extract time series characteristics based on "Wang, Smith and Hyndman" framework, and apply a principal component analysis to further reduce dimensionality before clustering. This yields interpretable groups of areas with similar volatility behavior. Another research direction regards incorporating the estimated volatility components as covariates in the VARX+DHR model, to improve its ability to capture latent mobility dynamics and forecast accurately. By linking volatility to human movements, our approach provides deeper insights into mobility patterns impacting air quality and congestion, supporting more accurate forecasting and contributing a novel, data-driven perspective to sustainable urban and environmental management.

An Enhanced New Multivariate Gwr Approach For Spatio-Temporal Pm10 Levels Prediction

Authors:

Antonella Congedi[1*†], Sandra De Iaco[2], Monica Palma[3]
Lucia Simmini[1]

[1] University of Salento
[2] University of Salento; National Centre for HPC, Big Data and Quantum Computing; National Biodiversity Future Center
[3] University of Salento; National Centre for HPC, Big Data and Quantum Computing (Bologna)

[*] Corresponding author [†] Presenter

Contact: antonella.congedi@unisalento.it

Keywords:

PM10 predictions, Multivariate GWR, Spatio-temporal analysis

Abstract:

Nowadays, the global awareness of climate change has led to an increasing attention in scientific literature on environmental pollution, which is defined as physical, chemical and biological alterations to the natural environment caused mainly by human activities. Among the several issues that impact on the well-being of humans and the Earth, the air quality is one of the most critical dimensions. Consequently, monitoring air pollutants is essential for supporting public policies for the safeguarding of human health and environmental sustainability. In this paper, air pollution due to particulate matter with a diameter of 10 micrometers or less (PM10), has been modelled using multivariate spatio-temporal techniques. Differently from the previous studies, this work simultaneously captures the influence of demographic, economic, and environmental variables over time, at a highly disaggregated (municipal) spatial scale. To this end, a time-indexed geographically weighted regression (GWR) multivariate model has been applied to monthly observations of PM10 across Italian municipalities and to the covariates which most affect the level of particulate matter. This approach allows the estimation of regression coefficients over the spatial domain for each recorded time point. Then, by computing the sample spatio-temporal variogram of the obtained coefficients, the behaviour of the dependent variable with respect to each predictor has been evaluated. Furthermore, the kriging technique has been carried out to predict the regression coefficients, and, consequently, to forecast the dependent variable PM10 in future time points through the multivariate GWR. In order to assess the reliability of the proposed procedure, the jackknife technique has been applied to a properly selected test set. This new multivariate spatio-temporal approach, based on the integration of the GWR with geostatistical techniques, improves the understanding of the complex dynamics that drive air pollution on a fine spatial scale, providing valuable insights that can inform targeted and timely environmental policies. Fundings Funded by European Union-NextGenerationEU with the Cascade Open Calls published by ALMA MATER STUDIORUM- University of Bologna, inside the Project GRINS funded by PNRR - Mission 4, Component 2, Investment 1.3 "Partnership extended to Universities, Research Centers, Firms and research projects funding", D.D. 341 of 15/03/2022. "ECoST-DATA, Exploring Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques", CUP: J33C22002910001 - CODICE: PE00000018 - Project Manager Prof. Sandra De Iaco: sandra.deiaco@unisalento.it

Advances Of Spatio-Temporal Clustering For Evaluating The Interaction Between Tourism And Environment

Authors:

Veronica Distefano[1], Sandra De Iaco[2*†]

[1] Università del Salento, European Centre for Living Technology Cá Foscari University of Venice; University Pegaso
[2] University of Salento; National Centre for HPC, Big Data and Quantum Computing; National Biodiversity Future Center

[*] Corresponding author [†] Presenter

Contact: sandra.deiaco@unisalento.it

Keywords:

spatial-temporal clustering model, sustainability, space-time metric

Abstract:

Nowadays, there is a growing interest to assess sustainability, which is assuming a key role in various socio-economic areas. In this context, the tourism sector, which is an important factor in terms of economic and social development, has a significant impact on the environment and contributes to climate change. For this reason, there is more and more attention on any forms of sustainable and green tourism and consequently on studies that combine the two underlying domains, that is the tourism growth and the environment safeguard. The aim of this work is to develop a methodological framework for a spatio-temporal clustering model to investigate the interaction between tourism and environmental factors in Italy, as well as the pattern recognition in order to reduce the intrinsic heterogeneity of Italian territory. In particular, an innovative boostrap spatio-temporal clustering is proposed, where the Ward hierarchical clustering algorithm that includes spatial and temporal constraints is included. The space-time component is built as a combination of a spatial distance matrix and a temporal distance matrix based on the concept of a space-time metric. This new unsupervised algorithm can be performed for time series that may differ in length. The results obtained by using the proposed procedure are illustrated through a real data set characterized by several tourism and environmental indicators, taken during a five-year period for the administrative regions (NUTS2 level) of Italy. The proposed methodology aims to guide and support the development of regionally tailored environmental policies. Keywords: spatial-temporal clustering model, sustainability, space-time metric. Funding Funded by European Union NextGenerationEU with the Cascade Open Calls pub- lished by ALMA MATER STUDIORUM â University of Bologna, inside the Project GRINS funded by PNRR â Mission 4, Component 2, Investment 1.3 Partnership extended to Uni- versities, Research Centers, Firms and research projects funding, D.D. 341 of 15/03/2022. ECoST-DATA, Exploring Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques, CUP: J33C22002910001 â CODICE: PE00000018 â Project Manager Prof. Sandra De Iaco: sandra.deiaco@unisalento.it

Session - Advances in clustering algorithms: contrastive hierarchical approaches and dynamic time warping

Organizers: Anna Denkowska and Stanisław Wanat

Large-Scale Benchmarking Of Glms And Machine Learning Models In Auto Insurance Ratemaking

Authors:

Sebastian Baran[1*†]

[1] Institute of Quantitative Methods in Social Sciences, Cracow University of Economics, Poland

[*] Corresponding author [†] Presenter

Contact: sebastian.baran@uek.krakow.pl

Keywords:

generalized linear models, machine learning, neural networks, insurance ratemaking

Abstract:

In recent years, the insurance industry has increasingly embraced advanced analytics and machine learning techniques to enhance pricing accuracy and risk segmentation. Despite this trend, Generalized Linear Models (GLMs) remain the standard approach for insurance ratemaking due to their interpretability, regulatory acceptance, and statistical robustness. This presentation offers a comprehensive empirical comparison between classical GLMs and selected machine learning models in the context of auto insurance ratemaking. The analysis is conducted on an extensive dataset comprising over 30 million policy records, representing one of the largest auto insurance datasets examined in this type of study. We evaluate and compare model performance using insurance-specific metrics. The modeling framework includes separate frequency and severity models, as well as combined pure premium models, reflecting common structures used in ratemaking practice. The machine learning models considered include gradient boosting machines (GBM), random forests, and neural networks, with particular attention given to model calibration, computational efficiency, and scalability. In addition to predictive performance, we assess the interpretability and transparency of the models using tools such as partial dependence plots and SHAP values, reflecting the practical needs of actuarial professionals and regulators. The study highlights the trade-offs between predictive power and interpretability, and discusses scenarios where machine learning models can complement or enhance traditional GLMs without compromising compliance or explainability. This session aims to provide a realistic perspective on the strengths and limitations of each approach, helping actuaries and data scientists understand when and how machine learning can complement traditional models. The goal is not to advocate for one method over another, but to explore their roles in the evolving landscape of insurance analytics.

Contrastive Hierarchical Clustering

Authors:

Przemysław Rola[1*†]

[1] Cracow University of Economics

[*] Corresponding author [†] Presenter

Contact: przemyslaw.rola@uek.krakow.pl

Keywords:

contrastive learning, hierarchical clustering, soft decision tree

Abstract:

Contrastive Hierarchical Clustering Deep clustering has been dominated by flat models, which split a dataset into a predefined number of groups. Although recent methods achieve an extremely high similarity with the ground truth on popular benchmarks, the information contained in the flat partition is limited. We introduce CoHiClust, a Contrastive Hierarchical Clustering model based on deep neural networks, which can be applied to typical image data. We use a soft binary decision tree to create a hierarchical structure, where leaves play the role of clusters. In contrast to hard decision trees, every internal node defines the probability of taking a left/right branch. To train CoHiClust, we introduce the hierarchical contrastive loss function designed for trees. Our idea is based on maximizing the likelihood that similar data points will follow the same path. The more similar data points, the longer they should be routed through the same nodes. Since we work in an unsupervised setting, we use a self-supervised approach and generate similar images using data augmentations. As a result, CoHiClust distills the base network into a binary tree. The hierarchical clustering structure can be used to analyze the relationship between clusters, as well as to measure the similarity between data points. Experiments demonstrate that CoHiClust generates a reasonable structure of clusters, which is consistent with our intuition and image semantics. Moreover, it obtains superior clustering accuracy on most of the image datasets compared to the state-of-the-art flat clustering models. The hierarchical structure constructed by CoHiClust provides significantly more information about the data than typical flat clustering models. In particular, we can inspect the similarity between selected groups by measuring their distance in the hierarchy tree and, in consequence, find super-clusters. The presentation is based on joint work "Contrastive Hierarchical Clustering".

Dtw-Based Time Series Clustering With Application To The Identification And Measurement Of Systemic Threats In The Insurance Sector

Authors:

Anna Denkowska[1*†], Maciej Denkowski[2], João Paulo Vieito[3]
Stanisław Wanat[1]

[1] Cracow University of Economics
[2] Jagiellonian University
[3] Polytechnic Institute of Viana do Castelo

[*] Corresponding author [†] Presenter

Contact: anna.denkowska@uek.krakow.pl

Keywords:

Dynamic Time Warping, Minimum Spanning Trees, Systemic Risk

Abstract:

DTW-Based Time Series Clustering with Application to the Identification and Measurement of systemic threats in the Insurance Sector Anna Denkowska1, Maciej Denkowski2, João Paulo da Torre Vieito 3 and Stanisław Wanat 4 1,4 Department of Mathematics, Cracow University of Economics, e-mail: anna.denkowska@uek.krakow.pl, wanats@uek.krakow.pl 2 Faculty of Mathematics and Computer Science, Jagiellonian University, e-mail: maciej.denkowski@uj.edu.pl 3 School of Business Studies, Polytechnic Institute of Viana do Castelo, e-mail: joaovieito@esce.ipvc.pt Clustering time series data presents unique challenges compared to clustering traditional data due to the temporal ordering and potential misalignment of time steps. Dynamic Time Warping (DTW) is a key technique that has significantly advanced time series clustering, offering more flexible similarity measurements than traditional metrics like the Euclidean distance. DTW measures similarity between two time series that may vary. It works by non-linearly aligning sequences to minimize distance, making it ideal for clustering real-world time series data where patterns may be similar but not aligned. In the context of increasing financial exposure to risks, especially in the insurance sector, there is a growing need for advanced analytical tools that can detect emerging patterns of systemic instability. This study explores the use of DTW for clustering time series of topological indicators obtained for Minimum Spanning Trees (MST) constructed based on return rates of insurance companies. The MST framework captures the evolving structure of interdependencies within the insurance market, providing a compact representation of systemic connectivity, and thus bringing information about possible path of contagion. From these dynamic networks, we extract time series of topological indicators such as node degree, betweenness centrality, clustering coefficient, and average path length. These indicators reflect changes in market structure over time and can signal stress propagation or structural shifts. We apply DTW-based clustering to group similar temporal behaviours of these indicators, allowing us to identify periods of synchronized stress or anomaly. Unlike traditional clustering methods, DTW accommodates time lags and non-linear shifts, making it particularly suitable for irregular or delayed responses to external shocks - such as climate-related events or regulatory changes. The goal is to integrate this methodology into early-warning systems for systemic and climate risk detection. This work contributes to the growing field of network-based financial risk analysis and provides practical insights for regulators, risk managers, and actuaries seeking to incorporate climate-related risk signals into their systemic risk monitor-

ing frameworks. Future work includes refining the mapping between topological patterns and specific climate risk scenarios, and extending the approach to cross-sectoral financial networks.

Analyzing The Impact Of Tail Dependencies On Value At Risk (Var) Using Distorted Mix Copula

Authors:

Krystian Szczęsny[1*†], Stanisław Wanat[1], Valentina Lorenzoni[2]

[1] Krakow University of Economics
[2] Scuola Superiore Sant'Anna in Pisa

[*] Corresponding author [†] Presenter

Contact: szczesnk@uek.krakow.pl

Keywords:

Tail dependencies, Value at Risk (VaR), Distorted Mix Copula (DMC), Solvency II

Abstract:

While dependencies in the central part of a distribution can be estimated with relative precision, tail dependencies are significantly more challenging to capture yet they have a profound impact on final Value at Risk (VaR) estimates. Traditional methods based on a single copula often fail to accurately represent both central and extreme dependencies simultaneously, which leads to the underestimation of risk and greater model uncertainty. This difficulty stems from the scarcity of data in the tails and the inherently less stable nature of extreme dependencies. Moreover, many classical copulas such as the Gaussian or Frank copula exhibit weak or no tail dependence, making them inadequate for analyzing extreme risk. In addition, conventional approaches lack the flexibility to adapt the model to varying dependency structures across different regions of the distribution. This results in oversimplifications that are insufficient for stress testing. In practice, while the model may capture dependencies well in the central region, where data is plentiful, it may entirely miss the crucial interactions that occur under extreme conditions, such as during market crises. To address this, we employ a Distorted Mix Copula (DMC) approach, which enables separate modeling of dependence structures in the central region and in the tails. In the first stage, central dependencies are estimated using classical methods, while the tail dependencies are modeled using various copulas specifically tailored to extreme co-movement behaviors. To identify the combination of tail copulas that yields the highest VaR, we use an optimization algorithm based on simulated annealing. The combination of the DMC framework and Monte Carlo optimization allows for a more realistic and comprehensive estimation of maximum VaR, assuming the central dependency structure is known. This approach is particularly relevant in the context of regulatory frameworks such as Basel III and Solvency II, which require the inclusion of worst-case scenarios in risk modeling, ensuring more conservative and robust financial risk assessments. Ultimately, this method helps mitigate the risk of regulatory capital underestimation and enhances the resilience of investment portfolios to abrupt, systemic shocks. It enables financial institutions to better prepare for stress events, where previously neglected tail dependencies could lead to rapid and difficult to manage losses even when the model performs well under normal market conditions.

How European Countries Cluster With Respect To The Ability To Satisfy Health Needs And Ensure The Health Of Citizens By Their Expenditure On Health?

Authors:

Valentina Lorenzoni[1*†]

[1] Scuola Superiore Sant'Anna

[*] Corresponding author [†] Presenter

Contact: valentina.lorenzoni@santannapisa.it

Keywords:

model-based clustering, health needs, health expenditure

Abstract:

Background: The ability of countries to meet the health needs of citizens is essential to preserve health which is a leverage for the development and the economic growth of the country. Methods: Using data from the Eurostat database about health expenditure (expressed as proportion of the Gross Domestic Product), self-reported unmet needs, for medical examinations as well as long standing illness for European countries, we used a model-based clustering relying on multivariate beta distribution order to take into account the skewed and constrained (in the [0,1] interval) distribution of variables that were all expressed as proportion. The number of G mixtures (i.e., clusters) better fitting data could then be chosen on the basis of the solution maximizing the Bayesian Information Criterion (BIC). Results: On the basis of the BIC value, the solution using three mixture components (each corresponding to a different cluster, CL) better fit the data. The first component (CL1) grouped countries with "high health needs and high unmet needs"; the second component (CL2) represented countries with "high unmet health needs, low health needs as well as low healthcare expenditure", while the third component grouped countries with "high unmet needs, high health needs and high expenditure". The mixing proportion estimated suggested that CL3 is the most represented (40%), CL1 and CL2 comprised each 30% of the countries analysed. Interestingly, each cluster contains countries from different geographical areas. Conclusions: Understanding how countries are able to satisfy health needs is essential, and further studies are necessary to better depict how countries behave.

Session - Advanced methods for cellwise outlier detection

Organizer: Giorgia Zaccaria

Cellwise And Casewise Robust Covariance In High Dimensions

Authors:

Fabio Centofanti[1*†], Mia Hubert[1], Peter Rousseeuw[1]

[1] KU Leuven

[*] Corresponding author [†] Presenter

Contact: fabio.centofanti@kuleuven.be

Keywords:

Casewise outliers, Covariance estimation, Cellwise outliers, High-dimensional data, Regularization

Abstract:

The sample covariance matrix is a cornerstone of multivariate statistics, but it is highly sensitive to outliers. These can be casewise outliers, such as cases belonging to a different population, or cellwise outliers, which are deviating cells (entries) of the data matrix. Recently some robust covariance estimators have been developed that can handle both types of outliers, but their computation is only feasible up to at most 20 dimensions. To remedy this we propose the cellRCov method, a robust covariance estimator that simultaneously handles casewise outliers, cellwise outliers, and missing data. It relies on a decomposition of the covariance on principal and orthogonal subspaces, leveraging recent work on robust PCA. It also employs a ridge-type regularization to stabilize the estimated covariance matrix. We establish some theoretical properties of cellRCov, including its casewise and cellwise influence functions as well as consistency and asymptotic normality. A simulation study demonstrates the superior performance of cellRCov in contaminated and missing data scenarios. Furthermore, its practical utility is illustrated in a real-world application to anomaly detection. We also construct and illustrate the cellRCCA method for robust and regularized canonical correlation analysis.

Casewise And Cellwise Robust Tensor-On-Tensor Regression

Authors:

Mehdi Hirari[1*†], Fabio Centofanti[1], Mia Hubert[1]
Stefan Van Aelst[1]

[1] KU Leuven

[*] Corresponding author [†] Presenter

Contact: mehdi.hirari@kuleuven.be

Keywords:

Anomaly detection, Casewise outliers, Cellwise outliers, Robust statistics, Tensor data, Tensor regression.

Abstract:

Tensor-on-tensor (TOT) regression is an important tool for analyzing tensor data. Its aim is to estimate a tensor response from a set of predictor tensors. However, standard TOT is sensitive to outliers, which may be present in both the response and the predictors. It is particularly affected by observations that deviate from the bulk of the data, known as casewise outliers, as well as by individual outlying cells within the tensors, referred to as cellwise outliers. The latter are especially likely to occur in tensor data, as tensors typically contain a large number of cells. This paper introduces a novel robust TOT method that can handle both types of outliers simultaneously, and can cope with missing values as well. This method uses a single loss function to reduce the influence of both casewise and cellwise outliers in the response. The outliers in the predictor are handled using Robust Multilinear Principal Component Analysis, a newly proposed tensor decomposition method that is robust to both casewise and cellwise outliers. Graphical diagnostic tools are also proposed to identify the different types of outliers detected by the new robust TOT method. The performance of the method and associated graphical displays is assessed through simulations and illustrated on a real dataset.

Challenges Of Cellwise Outliers

Authors:

Jakob Raymaekers[1*†], Peter Rousseeuw[2]

[1] University of Antwerp
[2] University of Leuven

[*] Corresponding author [†] Presenter

Contact: Jakob.Raymaekers@uantwerpen.be

Keywords:

robust, cellwise outliers, covariance, regression

Abstract:

It is well-known that real data often contain outliers. The term outlier typically refers to a case, typically denoted by a row of the nxd data matrix. In recent times a different type has come into focus, the cellwise outliers. These are suspicious cells (entries) that can occur anywhere in the data matrix. Even a relatively small proportion of outlying cells can contaminate over half the cases, which is a problem for robust methods. In this talk, we discusses the challenges posed by cellwise outliers, and some methods developed so far to deal with them. New results are obtained on cellwise breakdown values for location, covariance and regression. We conclude by looking to the future and by formulating some points for debate.

Session - Symbolic data analysis

Organizer: Paula Brito

Spatial Clusterwise Functional Regression For Predicting Distributional Data: An Application To Co_2 Emissions

Authors:

Rosanna Verde[1*†], Gianmarco Borrata[1], Antonio Balzanella[1]

[1] University of Campania "Luigi Vanvitelli

[*] Corresponding author [†] Presenter

Contact: Rosanna.verde2@gmail.com

Keywords:

Distributional Data, Clusterwise Regression Model, Spatial dependence

Abstract:

This work deals with a Functional Clusterwise Regression (CWFR) model based on a new regression method for distributional data. The first main contribution is the definition of a new regression model that maps probability density functions into a Hilbert space through the Logarithmic transformation of Derivative Quantile functions (LDQ). This distributional processing of the data has the advantage of allowing an analysis of the new functions and being able to return from the achieved results to the original quantile functions, through an inverse transformation. Each distributional variable is processed through this LDQ transformation and then treated as functional data by applying B-spline smoothing over the quantile domain. This facilitates a flexible and interpretable modeling of the distributional variables. The proposed CWFR model predicts a response distribution by partitioning the observations into K clusters, each associated with its local regression model, thereby capturing potential heterogeneity in the relationships between covariates and the response. As extension, we introduce the Spatial Functional Clusterwise Regression (SCWFR) model to analyze and predict Carbon Dioxide (CO2) emissions across different geographic areas. This model incorporates spatial covariates into the CWFR model, enabling the identification of the most relevant spatial variables for estimating CO_2 emissions at the local level. The SCWFR approach accounts for spatial variability in the distribution of covariates across regions and improves prediction accuracy by exploiting the spatial characteristics of the data. The novelty of the SCWFR lies in its interpretability: it allows for the evaluation of the influence of explanatory variables across different quantiles of their distributions, highlighting which value ranges contribute most to the relationship with the response distribution. Furthermore, it shows how spatial covariate variability affects these relationships differently across regions. Preliminary results, based on the Italian Municipalities Dataset, confirm the effectiveness of the proposed approach, showing that local estimates obtained through SCWFR outperform global models in terms of prediction accuracy and interpretability.

Entropy-Based Discriminant Analysis For The Classification Of Density-Valued Symbolic Data

Authors:

Francesca Condino[1*†], Paula Brito[2]

[1] Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria
[2] Faculdade de Economia, Universidade do Porto and LIAAD - INESC TEC

[*] Corresponding author [†] Presenter

Contact: francesca.condino@unical.it

Keywords:

Parametric distribution, Mixture, Kullback-Leibler divergence, Entropy

Abstract:

In recent decades, traditional discriminant analysis techniques have been extended to address more complex data structures. These include data represented as intervals, histograms, or distributions, typically organized within symbolic data tables. Symbolic Data Analysis provides a powerful framework for managing such richly structured datasets, which go beyond the simplicity of conventional unit-variable formats. In response to these challenges, this study proposes a novel discriminant analysis approach for density-valued data, leveraging the Jensen-Shannon divergence and its properties in terms of dissimilarity measure among distributions. Indeed, this entropy-based divergence allows us to quantify the discrepancy between objects and offers some important advantages, making it especially suitable in this context. The main idea of the proposed method is to assign each statistical unit, represented by a parametric density function, to one of the predefined groups in a way that minimizes the Jensen-Shannon divergence within groups. Indeed, the proposed methodology allows for a decomposition of total divergence into within-group and between-group components, in accordance with Huygens' theorem, thereby enabling the derivation of an effective classification rule. This is achieved by computing the barycentre of each group as a mixture of the constituent densities, so that the barycentre of each group belongs to the space of representation. An application on real data is proposed. The dataset concerns air time and departure delays recorded in 2013 for airline companies operating at New York airports, categorized into Main and Regional carriers. Each unit in the symbolic data table represents a specific airline in a given month, described by parametric density functions modeling departure delays and air time. The Generalized Extreme Value distribution and a flexible multimodal model are employed to capture the characteristics of the data. Classification results demonstrate that the proposed method performs with high accuracy for each descriptor in distinguishing between Main and Regional carriers. Finally, the dependence among symbolic variables is addressed by adopting a copula-based approach to link the marginal distributions, and an extension of the methodology to the bivariate case is proposed.

Principal Component Analysis Of Distributional Data: Method And Applications.

Authors:

Sónia Dias[1*†], Paula Brito[2]

[1] Escola Superior de Tecnologia e Gestão - Instituto Politécnico de Viana do Castelo and LIAAD-INESC TEC
[2] Faculty of Economics, University of Porto and LIAAD-INESC TEC

[*] Corresponding author [†] Presenter

Contact: sdias@estg.ipvc.pt

Keywords:

Histogram-valued variables, Principal Component Analysis, DSD regression model

Abstract:

The development of models and methods for representing, analysing, interpreting, and organising distributional data has been increasing. Linear models serve as the foundation for several statistical techniques, including linear regression, linear discriminant analysis, and principal component analysis. The Distribution and Symmetric Distribution (DSD) linear regression model allows predicting the distribution of the target variable from other histogram-valued variables, and is obtained optimizing a criterion based on the Mallows distance between the observed and the predicted distributions. In this work, a Principal Component Analysis (PCA) that uses the definition of linear combination considered in the DSD Model is proposed. Each principal component is obtained by a linear combination of the p original correlated histogram-valued variables, represented by the respective quantile functions. Since the space of quantile functions is a semi-vectorial space, the linear combination definition uses the quantile functions of the observed histograms and of the corresponding symmetric histograms, consequently non-negativity constraints are imposed on the parameters. The definitions of covariance and variance between histogram-valued variables are adapted to these type of data and are based on the Mallows distance. Similarly to the classical case, the (normed) vector of parameters defining the first principal component (PC1) is estimated by maximizing its variance, subject to the condition of non-negativity of the parameters. Note that in this case PC1 is a quantile function. Since, as usual, the variables used in PCA are measured in different scales, the original histogram-valued variables should be standardized. The proposed approach for the determination of the first principal component may be particularized to interval-valued variables, which constitute a special case of histogram-valued variables. To analyse and interpret the behaviour of the results obtained for the First Principal Component, two applications were studied. For a data set with 33 car models described by four strongly correlated interval-valued variables - price, engine capacity, top speed, and acceleration - the PC1 accounts for 91.35% of the total variance of the original variables. In the second study, scientific journals were aggregated in eight scientific areas described by five histogram-valued variables: number of published papers, impact factor, immediacy index, total citations, cited half-life. The conclusions were that the PC1 accounts for 54.2% of the total variance of the original variables. Moreover, the PC1 is strongly and positively correlated with impact factor, immediacy index, and total citations. On the other hand, PC1 is weakly (negatively) correlated with the number of published papers and cited half-life. The results obtained in these applications show the importance and relevance of continuing the generalization of the method to p principal components. Acknowledgments This work is funded by

national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2023 (https://doi.org/10.54499/UID/50014/2023).

Session - Advances in clustering and classification for mixed Data

Organizer: Carlo Cavicchia

New Robust Distance-Based Clustering Algorithms For Large Mixed-Type Data

Authors:

Aurea Grané[1*†], Fabio Scielzo-Ortiz[1]

[1] Universidad Carlos III de Madrid

[*] Corresponding author [†] Presenter

Contact: aurea.grane@uc3m.es

Keywords:

clustering, fast k-medoids, generalized Gower distance, multivariate heterogeneous data, outliers, robust Mahalanobis distance

Abstract:

In this work, new robust efficient clustering algorithms able to deal with big data are developed, namely Fast k-medoids and q-Fold Fast k-medoids, both implemented in a new Python package, called FastKmedoids. Their performance is analyzed in rather complex mixed-type datasets, whose size go from 35k to 1M, with outlier contamination and different patterns of underlying correlation structure. The simulation study comprises four computational experiments, where the stability, accuracy and efficiency of the new proposals is tested and compared to existing clustering alternatives. MDS is used to visualize clustering results.

A Regularized Wishart Mixture Model For Clustering Covariance Objects

Authors:

Andrea Cappozzo[1*†], Alessandro Casa[2]

[1] Università Cattolica del Sacro Cuore
[2] Free University of Bozen-Bolzano

[*] Corresponding author [†] Presenter

Contact: andrea.cappozzo@unicatt.it

Keywords:

Covariance matrices, Model-based clustering, Penalized likelihood, EM algorithm, Sparse estimation, Covariance graph models

Abstract:

Covariance objects are fundamental in many scientific disciplines, as they encode linear relationships among variables and thereby facilitate understanding of complex, multidimensional phenomena. In finance, they form the basis of portfolio theory by informing both risk management and investment decisions. In genomics, they capture gene dependencies essential for identifying trait associations and understanding genetic variation. In neuroscience, covariance matrices derived from brain imaging data are instrumental in analyzing functional connectivity between brain regions, offering critical insights into neural dynamics. This talk focuses on the problem of clustering when the data consist of sample covariance matrices. To this extent, a natural model-based clustering approach involves the use of mixture models, with each component modeled by a Wishart distribution. However, in high-dimensional settings, this approach encounters significant scalability challenges due to the quadratic increase in the number of parameters with the number of variables. To address this limitation, we propose a sparse Wishart mixture model that introduces cluster-specific sparsity in the component scale matrices. Estimation is carried out by maximizing a penalized log-likelihood, incorporating a covariance graphical lasso penalty within a tailored EM algorithm. This regularization promotes both interpretability and computational efficiency by shrinking weak or spurious associations toward zero, thereby highlighting the most salient variable interactions within each cluster. We demonstrate the effectiveness of our method through an application to functional magnetic resonance imaging (fMRI) data. The model successfully clusters individuals based on their functional brain connectivity, revealing meaningful patterns of neural interaction and offering insights into potential neurological differences. The imposed sparsity also enhances visualization and interpretation of connectivity structures within and across clusters.

Cluster Analysis From An Information-Theoretic Viewpoint

Authors:

Efthymios Costa[1*†], Ioanna Papatsouma[1], Angelos Markos[2]

[1] Imperial College London
[2] Democritus University of Thrace

[*] Corresponding author [†] Presenter

Contact: efthymios.costa17@imperial.ac.uk

Keywords:

clustering, information theory, information bottleneck, rate distortion theory, mixed-type data

Abstract:

Cluster analysis is the task of assigning objects into groups. This allocation into clusters is typically conducted based on pairwise dissimilarities or distances among objects, or by assuming clusters to have been generated by some probabilistic mechanism. Besides these two approaches, namely dissimilarity-based and model-based clustering, information-based clustering has recently emerged as a competitive alternative. Information-based clustering presents several advantages over traditional clustering approaches, such as avoiding making non-trivial assumptions about the data structure, or not requiring the selection of a dissimilarity metric. The rationale behind information-based clustering is that cluster analysis seeks to obtain a compressed representation of the data, so that each observation is identified by the cluster it belongs to, while ensuring that similar objects are assigned into the same cluster. There is a straightforward analogy of this task to the formulation of rate-distortion theory problems, where interest lies in the optimal compression of an input signal by a communication channel so that it can be reconstructed by a receiver with minimal distortion. We present the foundations of information-based clustering and refer to the rate-distortion problem and how it is formulated using information-theoretic quantities. We then discuss the information bottleneck algorithm and its variants, which provide us with a powerful and versatile framework that can be used for clustering purposes, while controlling the contribution of variables to the final solution. We implement information-based clustering using the deterministic variant of the information bottleneck algorithm on artificially generated and publicly available mixed-type data sets, that is data sets consisting of both continuous and categorical variables. The proposed framework is compared against state-of-the-art dissimilarity-based clustering methods for mixed-type data, making it clear that it offers a competitive approach in a variety of settings, while minimising the amount of required user input.

Session - Data-Driven decision making

Organizer: Tim Verdonck

Ensemble Cost-Sensitive Logistic Regression Models With Multi-Type Lasso Penalty

Authors:

Bing Yang[1*†], Stefan Van Aelst[1], Tim Verdonck[2]

[1] KU Leuven
[2] University of Antwerp

[*] Corresponding author [†] Presenter

Contact: bing.yang@kuleuven.be

Keywords:

Decision analysis, Cost-sensitive classification, Diverse ensemble, Interpretability

Abstract:

In many real-world applications, such as diagnosing a medical condition or detecting fraudulent transactions, a false negative (missing a case) is worse or more costly than a false positive. Therefore, it makes sense to take (mis)classification costs into account in the decision process to minimize the risks for stakeholders. To achieve this, cost-sensitive methods have been developed, such as cost-sensitive logistic regression. For high-dimensional data, it is well-known that ensemble models often perform much better than a single (sparse) model. However, ensembles of a large number of models are difficult to interpret, while explainability of decisions is often desirable or even legally required. To combine the interpretability of a single model with the performance of ensemble models, the split-learning framework has been developed recently. We developed a cost-sensitive split learning method that induces sparsity via the lasso penalty to obtain an interpretable model. However, real-world datasets usually contain a combination of different types of predictor variables for which the lasso is not the most appropriate choice. In particular, we focus on handling categorical predictors which frequently occur in real data. In this talk, we use the split-learning framework to introduce a diverse ensemble of cost-sensitive logistic regression models with multi-type lasso penalty to handle categorical predictors. To solve the non-convex optimization problem, a novel algorithm based on a partial conservative convex separable quadratic approximation is developed. The proposed method demonstrates substantial cost savings in extensive simulations and real-world applications. Moreover, the ensemble model obtained by the proposed method is fully interpretable as a logistic regression model.

Applied Robust Statistics Through The Monitoring Approach

Authors:

Marco Riani[1*†], Anthony C. Atkinson[2], Domenico Perrotta[3]
Aldo Corbellini[1], Valentin Todorov[4]

[1] University of Parma
[2] LSE
[3] JRC
[4] UNIDO

[*] Corresponding author [†] Presenter

Contact: mriani@unipr.it

Keywords:

Robust statistics, Monitoring Approach, Rregression, Clustering, Transformation

Abstract:

In this talk I will present the keypoints of the forthcoming book "Applied Robust Statistics through the Monitoring Approach: Applications in Regression" Heidelberg: Springer Nature. by Atkinson, Riani, Corbellini, Perrotta, and Todorov (2025). This open access book presents robust statistical methods and procedures through the monitoring approach, with an emphasis on applications to linear regression. Illustrating the theory, it explores both large and small-sample properties. The book describes the results of many years' work of the authors in the development of powerful methods of robust regression analysis. Robust methods are designed to analyse contaminated data. The well-established static robust methods estimate model features, such as parameter estimates, assuming the amount of contamination in the data is known. These methods are described in detail in Chapter2 for estimation in a simple sample. The extension to regression is in Chapter 3, with an emphasis on S-estimation and related procedures as well as on LTS. The monitoring methods of Chapter 4, including the forward search, find the appropriate level of robustness for each data set and so avoid biased estimation from the inclusion of outliers and inefficiency due to the deletion of uncontaminated observations. This analysis is followed by examples which illustrate the use of the interactive graphical analyses associated with our FSDA toolbox. Numerical comparisons of the size and power of outlier tests are in Chapter 5. Later chapters illustrate applications to response transformation in regression and to non-parametric regression. Extensions of the robust multiple regression model include Bayesian, heteroskedastic, time series and compositional regression, together with the clustering of regression models. Finally, several approaches to model selection are investigated and robust analyses of regression data are presented that illustrate the use of the techniques introduced earlier. The GitHub repo of the book is https://github.com/UniprJRC/MonitoringBook. In this repo it is possible to reproduce all the figures and tables. All routines which implement the methods in the book can also be used to explore your own data. They are available in the FSDA MATLAB toolbox which can be freely installed and used through MATLAB on line.

Robust Forecasting With Lstm

Authors:

Christophe Croux[1*†], Klaus Nordhausen[2], Mika Sipila[3]
Sara Taskinen[3]

[1] KU Leuven
[2] University of Helsinki
[3] University of Jyväskylä

[*] Corresponding author [†] Presenter

Contact: christophe.croux@kuleuven.be

Keywords:

Forecasting, Robustness, Machine Learning

Abstract:

Long Short-Term Memory (LSTM) models are special cases of recurrent neural networks. They have become a standard tool in the deep learning community for time series prediction. Standard LSTM models turn out not to be robust to outliers, despite the belief that deep neural networks can cope with highly non-linear and noisy patterns. In this paper we introduce a robust version of LSTM. Using simulation experiments, we show that robust LSTM can cope with different types of outliers, including patches of outliers and level shifts. Finally, we investigate how robust LSTM models may be used for time series outlier detection, and how accurate such a detection method is.

Frequentist Estimation Of Microclustering Models With Applications To Record Linkage

Authors:

Edoardo Redivo[1*†], Cinzia Viroli[1]

[1] Università di Bologna

[*] Corresponding author [†] Presenter

Contact: edoardo.redivo@unibo.it

Keywords:

clustering, record linkage, entity resolution, microclustering

Abstract:

Common approaches to model-based clustering, such as finite mixture models, implicitly assume that cluster sizes grow with the sample size, resulting in a few large clusters even as the number of observations increases. However, this situation does not properly reflect some clustering applications. One such case is record linkage, also known as entity resolution, where the aim is to identify observations recorded on the same statistical unit across one or more datasets. Record linkage can be used to improve data quality by removing duplicates or to integrate multiple datasets. This task can be framed as a clustering problem, where each cluster corresponds to a unique individual or entity. However, each individual is typically observed only a few times, and moreover, the number of duplicates is not expected to grow indefinitely with the sample size. To better describe scenarios such as record linkage, where many small clusters are expected, and in particular where cluster sizes should grow sublinearly with the sample size, the microclustering property has been recently formalized. This property, along with models adhering to it, have been introduced within the Bayesian framework of random partition models. In this work, we aim to characterize models exhibiting the microclustering property from a frequentist perspective. In particular, we establish sufficient conditions under which microclustering emerges in a frequentist setting, focusing on models where the number of clusters is a random variable. For practical modelling, we propose a novel mixture model formulation that distinguishes between unmatched observations, those forming singletons, and matched observations. For this model, we show that the model log-likelihood is not monotonically increasing in the number of clusters and prove the existence of a finite optimal number of clusters. Through simulation studies, we show that this model effectively recovers microclustering structures. Overall, this work aims to provide a principled and interpretable frequentist framework for microclustering, complementing Bayesian approaches and broadening their practical applicability.

Session - Advances in mixture models

Organizer: Paul McNicholas

Parsimonious Ultrametric Manly Mixture Models

Authors:

Alexa Sochaniwsky[1*†], Paul McNicholas[1]

[1] McMaster University

[*] Corresponding author [†] Presenter

Contact: sochaa1@mcmaster.ca

Keywords:

Model-based clustering, Ultrametricity, Manly transformation, Skewness

Abstract:

A family of parsimonious ultrametric mixture models with the Manly transformation is developed for clustering high-dimensional and asymmetric data. In model-based clustering literature, we have seen consistent advances in Gaussian mixture modeling, particularly in addressing high-dimensional data; however, such methods often fail to account for the presence of skewness. These advances frequently include decomposing and constraining covariance matrices. While these methods reduce the number of free parameters and improve clustering performance, they often provide limited insight into the structure and interpretation of the clusters. This research addresses this shortcoming by implementing the extended ultrametric covariance structure and the Manly transformation. This covariance structure reduces the number of free parameters while also identifying latent hierarchical relationships between and within groups of variables. This phenomenon allows the visualization of hierarchical relationships within individual clusters, improving cluster interpretability. Incorporating the Manly distribution with the extended ultrametric covariance structure and constraining said covariance structure results in the parsimonious ultrametric Manly mixture model family. As with many classes of mixture models, model selection remains a fundamental and unresolved challenge. It can be demonstrated that, for the proposed family, common model selection criteria fail to consistently identify the correct model; i.e., the correct number of clusters, number of groups of variables, and covariance structure. We propose a two-step model selection procedure. In the first step, the number of clusters and the number of groups of variables are selected; in the second step, the covariance structure is selected. With simulation studies and real data analyses, we demonstrate improved model selection via the proposed two-step method, and the effective clustering performance of the proposed family.

Model-Based Clustering Of Mixed-Type Compositional-Continuous Data

Authors:

Antonello Maruotti[1*†], Alfonso Russo[2], Fabio Divino[3]

[1] Libera Università Maria Ss. Assunta
[2] Università degli Studi di Roma Tor Vergata
[3] Università del Molise

[*] Corresponding author [†] Presenter

Contact: a.maruotti@lumsa.it

Keywords:

finite mixture, latent variables, EM algorithm

Abstract:

In this work, we develop a model-based clustering framework for multivariate data consisting of mixed-type variables, focusing on compositional and continuous variables. Our approach is motivated by the increasing need to analyse heterogeneous data structures where multiple response types coexist and may exhibit complex patterns of dependence and unobserved heterogeneity. Rather than modelling each variable independently or assuming overly simplistic dependence structures, we propose a unified strategy based on finite mixture models that incorporates flexible associations between outcomes through multidimensional discrete latent variables. At the core of our methodology lies a multidimensional finite mixture model in which each response variable is conditionally modelled given its own outcome-specific latent discrete variable. The latent variables underlying the finite mixture are not assumed to be independent; instead, they are jointly modelled, capturing the dependence among the multivariate responses. This formulation generalizes standard finite mixture models by relaxing the unidimensionality assumption, which typically forces all variables to share a single latent structure. In our framework, each margin (i.e., each outcome variable) can have a different number of mixture components, allowing for more tailored and interpretable modelling of variable-specific heterogeneity. As a result, the clustering structure is not driven by a single common membership but by a composition of outcome-specific memberships, leading to richer and more flexible clustering representations. Our model is particularly relevant for data scenarios where some variables are compositional-that is, they represent parts of a whole and are constrained to lie in the simplex-adding further challenges to the modelling. We address the unique challenges of compositional data by employing appropriate transformations and modelling strategies that preserve their relative and constrained nature, while integrating them seamlessly into the mixture framework. Estimation of model parameters is carried out using a likelihood-based approach via the Expectation-Maximization (EM) algorithm, which we adapt to handle the mixed-type and multidimensional latent structure of the data. Through simulations and applications to real-world data, we demonstrate the effectiveness of our proposed model in uncovering meaningful clusters, capturing complex inter-variable associations, and outperforming standard mixture approaches.

Change Point Detection In Categorical Sequences

Authors:

Volodymyr Melnykov[1*†], Lingge Wang[1]

[1] University of Alabama

[*] Corresponding author [†] Presenter

Contact: vmelnykov@ua.edu

Keywords:

categorical sequence, change point, finite mixture model

Abstract:

The majority of cluster analysis techniques have been developed for quantitative data. However, there is an abundance of applications with the goal of grouping qualitative data, particularly categorical sequences. Such sequences commonly arise in the analysis of Web navigation patterns, DNA, and medical records. In recent years, several papers on the topic addressed the problem of clustering categorical sequences. However, sometimes the objective of a study is to detect a change point in these sequences. For example, such a change point may indicate a moment when medical records of a healthy individual begin to exhibit new features that can be associated with the development of a certain health condition. We propose an approach for the effective detection of change points in heterogeneous categorical sequences.

Dimension-Wise Kurtosis Control And Parsimony In Model-Based Clustering

Authors:

Salvatore Daniele Tomarchio[1], Luca Bagnato[2], Antonio Punzo[1*†]

[1] University of Catania
[2] Università Cattolica del Sacro Cuore

[*] Corresponding author [†] Presenter

Contact: antonio.punzo@unict.it

Keywords:

Mixture models, EM algorithm, Heavy-tailed distributions

Abstract:

Dimension-wise scaled normal mixtures (DSNMs) form a recently proposed family of d-variate continuous distributions that extend the multivariate normal (MN) distribution by enabling: (1) a more general form of central symmetry, and (2) dimension-specific excess kurtosis. Like the MN distribution, DSNMs share the notable property that zero correlation implies independence. These features arise within a scale mixture of MNs, where the mixing is governed by a d-variate random variable composed of independent, similarly distributed components-each affecting one dimension individually. In this paper, we develop parsimonious finite mixtures of DSNMs for model-based clustering, targeting scenarios with symmetric clusters that exhibit varying degrees of excess kurtosis across dimensions. As a concrete example, we consider mixtures of dimension-wise scaled shifted exponential normal (DSSEN) distributions-a subclass of DSNMs where the mixing variables follow a shifted exponential distribution. This specific choice ensures a closed-form expression for the joint density of the DSSEN distribution. To induce parsimony in DSNM mixtures, we impose structured constraints on the conditional correlation, scale, and tailedness parameters. This leads to a family of 60 interpretable models. For the DSSEN case, we detail the estimation procedure-based on the expectation-conditional maximization algorithm-for different model configurations to compute maximum likelihood estimates. Finally, we demonstrate the practical utility of our models through applications to both simulated and real datasets, comparing their performance with established mixtures of symmetric heavy-tailed distributions from the literature.

Session - Preference Data

Organizer: José Luis Garcia Lapresta

Challenges In Preference-Approval Of Opportunity Sets

Authors:

José Carlos R. Alcantud[1*†]

[1] University of Salamanca

[*] Corresponding author [†] Presenter

Contact: jcr@usal.es

Keywords:

Opportunity set, Preference-approval, Ranking

Abstract:

Let us consider a finite collection named X of potential options. An opportunity set is defined as a non-empty collection of available alternatives from X. The field of opportunity freedom aims to establish procedures for the comparison of opportunity sets. Broadly speaking, there are two fundamental approaches to this topic. The first one views the size of an opportunity set as a gauge of freedom. On the contrary, the indirect utility approach assesses this set based on the value of its most beneficial elements. The later approach typically resorts to a complete and transitive preference relationship on X to establish comparisons between elements. Dispensing with welfaristic arguments, the cardinality rule pertains to the position that the number of options present in the opportunity set determines its importance. A significant restriction is that we cannot use a cardinality index in infinite sets. In addition, its lack of concern about the quality of the options was criticized by many authors. The indirect-utility ranking is the fundamental example of valuation by the most beneficial component. Both views are reconciled in several works. Other authors have studied rankings combining both types of arguments (e.g., leximax, cardinality-first lexicographic, preference-first lexicographic, and dominance). This work is the first attempt to introduce preference-approval analysis in this field. Conceptually, the improvement of preference-approval structures with respect to a preference-based approach is the inclusion of a set of "approved" objects that is appropriately related to a preference on the set of objects. Preference-approval structures appeared in the context of (hybrid) voting systems in 2009, and later on they were applied to consensus measurement and to clustering, especially with the help of distances defined for these concepts. In this contribution we expand their scope to the literature on opportunity freedom. The specific goals that we pursue are as follows. First, natural procedures are defined that use the knowledge contained in a preference-approval structure on the options to rank the opportunity sets that they define. Furthermore, preference-approval structures on opportunity sets are produced that supplement these and other known rankings with suitable sets of "approved" opportunity sets. Finally, we point out that it is possible to define preference-approval structures on opportunity sets from rankings of the options.

The Majority Principle Is Adequate Only For Purely Ordinal Individual Preferences

Authors:

Remzi Sanver[1*†]

[1] Université Paris Dauphine, CNRS

[*] Corresponding author [†] Presenter

Contact: remzi.sanver@dauphine.fr

Keywords:

majority principle, preference-approval, cardinal preference

Abstract:

Consider two possible social outcomes x and y. We define the majority principle as follows: If the number of individuals who prefer x to y exceeds those who prefer y to x, then the social ranking doesn't put y above x. We show that equal treatment of individuals and equal treatment of alternatives combined with a very weak monotonicity condition implies the majority principle. The three conditions are plausible when individual preferences are purely ordinal. We define ways of introducing intensities of different strength to individual preferences, the weakest being the preference-approval framework. For each of these ways, we present instances where the majority principle becomes indefensible. We conclude that the majority principle is very natural with purely ordinal individual preferences and only in this case.

Aggregating Ternary Preferences In a Scoring Context

Authors:

José Luis García-Lapresta[1], Miguel Martínez-Panero[1*†]

[1] Universidad de Valladolid

[*] Corresponding author [†] Presenter

Contact: panero@eco.uva.es

Keywords:

voting systems, Borda rule, approval voting, preference-approvals

Abstract:

Voting systems aggregate individual opinions on a set of alternatives to generate an outcome, usually a single winning alternative, a subset of winning or acceptable alternatives or a weak order on the set of alternatives. In this contribution, we extend preference-approval structures (each voter ranks the alternatives by means of a weak order and, additionally, assesses each alternative as either acceptable or unacceptable) to a more general situation, where voters sort the alternatives in three disjoint classes instead of two (for instance, acceptable, neutral and unacceptable). Voters not only sort the alternatives in these classes, but they may rank those acceptable or unacceptable. In this framework, we propose a parameterized family of voting systems related to the Borda rule, where positive (negative) individual scores are assigned to acceptable (unacceptable) alternatives in a decreasing way from best to worst, while neutral alternatives obtain null scores. As in the Borda rule and the approval voting system, the alternatives are collectively ranked by the sum of their individual scores. The parameters used in the voting systems allow greater or lesser importance to be given to acceptable, neutral and unacceptable alternatives: (1) if an acceptable alternative A is preferred to another acceptable alternative B, and A becomes unacceptable, then the score of B should increase; (2) it is less disgraceful for an unacceptable alternative to be defeated by an acceptable alternative than by a neutral one; (3) it is more meritorious to beat an acceptable alternative than to a neutral one; and (4) it is more meritorious to tie with an acceptable alternative than to beat a neutral one. We analyze the role of parameters and provide some properties that satisfy the proposed voting systems: anonymity, neutrality, reciprocity, weak preference unanimity, strong preference unanimity, indifference unanimity, strict Pareto, and cancellation.

Marginal Contribution To Consensus In Ternary Preferences

Authors:

Alessandro Albano[1*†], José Luis García-Lapresta[2], Antonella Plaia[1]
Mariangela Sciandra[3]

[1] University of Palermo
[2] Universidad de Valladolid
[3] Dipartimento di Scienze Economiche Aziendali e Statistiche

[*] Corresponding author [†] Presenter

Contact: alessandro.albano@unipa.it

Keywords:

Preferences, approval voting, preference-approvals, consensus, Banzhaf value

Abstract:

This study measures the overall consensus of a voter group and evaluates each voter's influence on shaping that consensus in ternary preference settings. In this framework, each voter ranks alternatives using a weak order and simultaneously assigns each alternative to one of three categories: acceptable, neutral, or unacceptable. This approach extends preference-approvals, which combines a weak order over alternatives with a binary classification into acceptable and unacceptable options, by introducing an intermediate category. We propose a novel distance-based measure that quantifies the overall consensus in a group of voters under ternary preferences. The proposed measure satisfies desirable properties and allows us to assess the degree of alignment in the expressed preferences. Building on this, we introduce the concept of marginal contribution to consensus, which evaluates how much each individual voter contributes (positively or negatively) to the overall agreement within the group. This measure is conceptually linked to the Banzhaf value in cooperative game theory. To address computational challenges in large-scale settings, we develop a sampling-based estimation procedure that efficiently approximates marginal contributions. Simulation studies confirm the validity and stability of the method as the number of voters increases. We apply our methodology to two real datasets-one from the Italian National Institute of Statistics (ISTAT) and one from the Balkan Barometer. The analysis reveals regional and national patterns in consensus and voter influence, demonstrating the interpretability and relevance of the proposed tools. This framework contributes to the study of collective decision-making by offering new metrics to evaluate both agreement and individual impact in complex preference settings.

Session - New approaches to dimensionality reduction: applications in the social sciences

Organizers: Alfonso Piscitelli and Leonardo Salvatore Alaimo

Modelling Social Inequalities With Identity Spline And Lasso Regression

Authors:

Ida Camminatiello[1*†], Rosaria Lombardo[1], Mario Musella[2]

[1] Università della Campania "Luigi Vanvitelli
[2] Università di Napoli Federico II

[*] Corresponding author [†] Presenter

Contact: ida.camminatiello@unicampania.it

Keywords:

Identity Spline, LASSO Regression, Socio-economic Inequalities

Abstract:

Addressing socio-inequalities in the digital era requires the development of integrated policies informed by advanced statistical modelling. These policies often involve coordinated investment in digital infrastructure, education, inclusive technological design, and the regulation of dominant technology platforms, alongside efforts to promote equitable digital labour practices. From a statistical perspective, modelling the impact of such multifaceted interventions across geographically and economically diverse regions poses significant challenges, especially due to structural heterogeneity, correlated covariates, and the need for interpretable scenario analysis. Cross-sectional data provide a valuable setting for examining the relationships between socio-economic indicators and investment patterns in education, infrastructure, and technology. However, assessing the implications of different policy strategies under uncertainty necessitates flexible modelling tools that can accommodate complex dependencies and enable exploratory perturbations of the response. To this end, we propose a modelling framework that combines identity splines with LASSO regression to support scenario-based analysis of social inequalities. The identity spline is a specialised spline function that enables controlled perturbations of a variable by adjusting its nodal coefficients. When used to represent the response variable-here an indicator of social inequalities-this construction allows for structured and interpretable modifications, facilitating the exploration of alternative policy scenarios. The LASSO regression model (Least Absolute Shrinkage and Selection Operator) is a widely used penalised regression technique that performs simultaneous variable selection and regularisation, making it well-suited for high-dimensional or multicollinear settings commonly encountered in socio-economic data. Building on this, we introduce a functional-scalar LASSO regression model, in which the response variable is transformed into a functional object via the identity spline. This formulation enables the response to be perturbed in a structured manner, allowing statisticians to investigate the sensitivity of outcomes to hypothetical interventions. The proposed model thus offers a principled statistical framework for scenario-based policy evaluation in the presence of complex predictor structures and uncertain outcomes.

Density-Based Clustering For The Detection Of High-Intensity Regions

Authors:

Samuela L'Abbate[1*†], Paola Perchinunno[2], Leonardo Salvatore Alaimo[3]

[1] Department of Humanities Research and Innovation
[2] Department of Economics, Management and Business Law
[3] Department of Statistical Sciences, University of Rome La Sapienza

[*] Corresponding author [†] Presenter

Contact: samuela.labbate@uniba.it

Keywords:

Density-based clustering, DBSCAN, HDBSCAN, High-intensity regions, Pattern recognition

Abstract:

This paper proposes an improved density-based clustering approach for identifying dense regions within data patterns, with a specific focus on radiographic analysis. The methodology leverages DBSCAN and HDBSCAN, two widely used clustering algorithms known for their ability to detect clusters of arbitrary shape and effectively manage noise. These methods are particularly suitable for applications where the number of clusters is unknown in advance and the data distribution may be irregular or complex. Our approach is designed to enhance the recognition of high-density areas that may reflect meaningful variations in the observed phenomenon. By carefully tuning algorithmic parameters and exploring different distance metrics, we demonstrate how density-based clustering can provide deeper insight into data structures that might be overlooked by conventional techniques. This is especially relevant in scenarios where traditional thresholding methods may fail due to noise, shape irregularity, or overlapping intensity levels. The method was applied to radiographic data to assess its ability to highlight regions with increased signal intensity. These areas, characterized by compactness and density, are often associated with features of potential clinical interest. The analysis showed that DBSCAN and HDBSCAN, when properly configured, perform robustly in detecting such regions. Their flexibility makes them particularly valuable for exploratory data analysis and as supportive tools in diagnostic workflows. Beyond its direct application, this study contributes to a broader understanding of how unsupervised learning techniques can be adapted and refined to enhance pattern detection in complex data. The results suggest that density-based clustering algorithms, even in their standard forms, are powerful tools for extracting meaningful insights from challenging datasets.

Assessing Foods Environmental Footprints Using Clustering Hierarchical Disjoint Principal Component Analysis

Authors:

Mariaelena Bottazzi Schenone[1], Maurizio Vichi[1*†]

[1] Sapienza University of Rome

[*] Corresponding author [†] Presenter

Contact: Maurizio.Vichi@uniroma1.it

Keywords:

Environmental impact, food sustainability, composite indicators, clustering

Abstract:

This study evaluates the environmental impacts of food production by proposing a novel methodology called Clustering Hierarchical Disjoint Principal Component Analysis, which simultaneously performs clustering and dimensionality reduction to provide a comprehensive understanding of complex multivariate data while simplifying their interpretation. The approach constructs a hierarchical set of Specific Composite Indicators (SCIs) that aggregate into a General Composite Indicator (gCI). Using data from 43 food items, the methodology synthesizes environmental dimensions such as land use, water consumption, and greenhouse gas emissions into SCIs, which are then combined into a gCI representing the overall environmental footprint of the foods. By identifying meaningful clusters of food items based on the gCI, the method uncovers variations in environmental burdens, offering valuable insights for promoting sustainable dietary choices. Ultimately, this approach supports policymakers and consumers in making informed decisions to reduce the environmental footprints associated with food production.

Session - Advances in text mining for data analysis

Organizer: Luca Frigau

Text-Based Propensity Scores For Analyzing Comorbidities In Ehrs

Authors:

Chiara Di Maria[1*†], Alessandro Albano[1], Mariangela Sciandra[1]
Antonella Plaia[1]

[1] Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli studi di Palermo

[*] Corresponding author [†] Presenter

Contact: chiara.dimaria@unipa.it

Keywords:

propensity score, EHR, embedding, xgboost, NLP

Abstract:

Understanding comorbidity patterns is fundamental to improving patient outcomes and optimizing healthcare delivery strategies. Traditional approaches to studying disease associations have primarily relied on structured Electronic Health Record (EHR) fields, such as diagnosis codes and laboratory values, often overlooking the rich, multidimensional information embedded within clinical narratives. This study presents a novel methodological approach that leverages natural language processing (NLP) and machine learning techniques to investigate comorbidities from unstructured clinical discharge notes, specifically focusing on the association between diabetes and chronic kidney disease (CKD). Our methodology consists of three main components: first, we reduce the dimensionality of clinical discharge notes through appropriate text representations, i.e. Term Frequency-Inverse Document Frequency (TF-IDF) weighted document-term matrices and dense text embeddings generated via the pre-trained all-MiniLM-L6-v2 model. Second, we estimate propensity scores that represent the probability of receiving a specific diagnosis conditional on the extracted textual information, employing various machine learning models including XGBoost, and multilayer perceptrons (MLPs) and LASSO regression. Finally, we incorporate these text-derived propensity scores as covariates in logistic regression models to analyze disease associations while adjusting for potential confounding factors present in clinical narratives. We applied this approach to the MIMIC-III database, analyzing 50,560 discharge summaries to investigate the relationship between diabetes and CKD. Our results demonstrate that text-based propensity score adjustment substantially improves model performance, with the embedding-based XGBoost model achieving the highest performance. Importantly, after adjusting for textual confounders, the estimated odds ratios were significantly attenuated compared to naive models, suggesting that failing to account for information in clinical texts may lead to inflated association estimates. The embedding-based XGBoost model yielded a more conservative and accurate estimate of the true association between diabetes and CKD compared to the naive model's odds ratio. These findings highlight the critical importance of incorporating unstructured clinical text data in comorbidity research and demonstrate the potential of NLP-based propensity score methods in advancing observational healthcare research and clinical decision-making.

hypertargetabs130

Unsupervised Topic Relationship Discovery Using Generalized Structured Component Analysis: A Novel Approach To Document Clustering

Authors:

Marco Ortu[1*†]

[1] University of Cagliari

[*] Corresponding author [†] Presenter

Contact: marco.ortu@unica.it

Keywords:

Structural Topic Modeling, General Component Analysis, Document Clustering

Abstract:

The exponential growth of digital text collections has intensified the need for sophisticated document organization and analysis methods. Topic modeling techniques, particularly Latent Dirichlet Allocation (LDA) and its variants, have become fundamental tools in understanding document collections by discovering latent themes. However, the common practice of clustering documents based on their dominant topics often oversimplifies the rich, probabilistic nature of topic distributions and fails to capture complex inter-topic relationships. Traditional document clustering approaches typically follow one of two paths: either assigning documents to clusters based on their dominant topic, or using similarity measures computed from topic distribution vectors. While these methods have proven useful for basic organization tasks, they suffer from several limitations. First, the reduction to dominant topics discards valuable information about secondary themes and topic proportions. Second, simple similarity-based clustering fails to capture structural relationships and indirect connections between topics. Third, these approaches often struggle to incorporate additional document-level covariates that might influence topic relationships. This paper introduces a novel methodological framework that addresses these limitations by leveraging Generalized Structured Component Analysis (GSCA) in an unsupervised manner. GSCA, originally developed for analyzing relationships between observed and latent variables, provides an ideal foundation for our approach due to its flexibility in handling complex relationships and its component-based nature that aligns well with topic distributions. Our framework makes several key contributions to the field. First, it preserves and utilizes the full probabilistic nature of topic distributions, allowing for more nuanced analysis of document relationships. Second, it employs a path significance assessment procedure that combines multiple statistical approaches, including bootstrap-based validation, regularization for stability, and out-of-bag prediction error evaluation. Third, it provides a natural mechanism for incorporating document-level covariates, enabling richer analysis of factors influencing topic relationships. The proposed method begins with a fully connected path matrix representing all possible relationships between topics. Through a systematic procedure combining statistical significance testing, effect size assessment, and stability analysis, it identifies and retains only the most meaningful relationships. This approach not only reveals the underlying structure of topic relationships but also provides a more robust foundation for document clustering based on these structural patterns. The theoretical foundations of our approach draw from several key areas: topic modeling, structural equation modeling, and component analysis. By integrating these perspectives, we

develop a comprehensive framework that bridges the gap between probabilistic topic modeling and structural analysis of relationships. The method's unsupervised nature makes it particularly valuable for exploratory analysis of large document collections where prior knowledge of topic relationships may be limited or unavailable.

Bridging Textual And Network Data Analysis: Exploring Trends In Ethereum Developer Discussions

Authors:

Silvia Bartolucci[1*†]

[1] University College London

[*] Corresponding author [†] Presenter

Contact: s.bartolucci@ucl.ac.uk

Keywords:

Topic modelling, Complex networks, Open source development

Abstract:

Understanding how complex socio-technical systems evolve over time requires integrated approaches that can capture both the content of communication and the structure of interaction. This talk presents a novel framework that combines natural language processing, topic modeling, and complex network analysis to study sustainability engagement within open-source software communities. As a case study, we focus on Ethereum, one of the world's most prominent blockchain ecosystems, where decentralized governance, energy-intensive infrastructure, and rapid technological change converge. Using over 65,000 GitHub issues and comments spanning nearly a decade, we apply BERT-based topic modeling-augmented with large language models-to extract and classify sustainability-related discussions. These are organized using the Sustainability Awareness Framework (SusAF), which structures sustainability across five key dimensions: environmental, economic, social, individual, and technical. We map how developers' attention to these dimensions has evolved over time, revealing a landscape heavily skewed toward technical optimization, but with a growing awareness of environmental and economic issues, particularly in response to major events such as Ethereum's shift from Proof-of-Work to Proof-of-Stake. To complement the textual analysis, we construct bipartite and projected developer-topic networks, allowing us to assess influence, engagement patterns, and the formation of thematic sub-communities. Our network results show that a small group of core developers consistently drive sustainability-related discourse, and that contributors tend to specialize in specific domains, leading to distinct knowledge clusters within the project. These dynamics suggest that while sustainability is gaining visibility, its integration remains uneven and often reactive to external pressures. This work offers new insights into how values, incentives, and responsibilities are negotiated in open-source environments. Ultimately, we highlight the importance of developers' communities not only as technical actors, but as central agents in shaping the sustainability trajectory of digital technologies. By integrating textual and network analysis, our approach offers a transferable methodology for studying users' engagement in other communication platforms and social media outlets.

Modeling The Impact Of Review Content On Tourist Satisfaction: The Case Of The Sardinian Hotel Reviews

Authors:

Giulia Contu[1*†], Marco Ortu[1]

[1] Università di Cagliari, Dipartimento in scienze economiche e aziendali

[*] Corresponding author [†] Presenter

Contact: giulia.contu@unica.it

Keywords:

tourism review, Partial least square, topic analysis, sentiment analysis, customer satisfaction

Abstract:

Understanding the drivers of tourist satisfaction is essential for enhancing service quality and destination competitiveness. In recent years, online reviews have emerged as a rich source of information for analyzing customer experiences. However, the causal relationship between the textual content of reviews-particularly expressed topics and emotions-and overall satisfaction ratings remains underexplored. This study addresses this gap by analysing reviews of hotels in Sardinia, with a focus on potential differences between coastal and inland locations. We employ the TOpic modeling Based Index Assessment through Sentiment (TOBIAS), an approach that integrates natural language processing with causal inference techniques to assess how specific topics, moods, and emotional tones in written reviews influence customer ratings. Our findings contribute to the growing body of literature at the intersection of text analytics and causal modeling and offer insights into the factors tourist satisfaction across different regional contexts.

Session - Classification for biomedical data

Organizer: Cinzia Viroli

A Flexible Latent Dirichlet Model For Modeling Taxa Communities

Authors:

Alice Giampino[1*†], Roberto Ascari[1], Sonia Migliorati[1]

[1] University of Milano-Bicocca

[*] Corresponding author [†] Presenter

Contact: alice.giampino@unimib.it

Keywords:

latent variable models, flexible Dirichlet distribution, collapsed Gibbs sampling, gut microbiome

Abstract:

The study of microbial communities residing in the human digestive tract, collectively referred to as the gut microbiota, along with their interactions, provides critical insights into human physiological and biochemical processes, disease mechanisms, and potential therapeutic strategies. Key challenges in this context include sparsity, high dimensionality, and substantial interindividual variability. A powerful method for analyzing microbiome data is Latent Dirichlet Allocation (LDA), a mixed-membership model originally developed in natural language processing to identify latent patterns in textual data. LDA can be adapted to identify bacterial communities that represent latent microbial abundance profiles, i.e., unique taxa compositions that are functionally dependent on each other and must be learned from data. LDA is a three-level hierarchical Bayesian model that imposes a Dirichlet prior distribution on the vectors of community compositions. Despite its widespread use, LDA's effectiveness is limited by the Dirichlet prior, which enforces near-independence-i.e., all forms of independence developed for compositional data analysis-among the components of community proportions, thereby precluding the modeling of relationships between taxa within each community. Independence is closely linked to the rigidity and limited richness of the Dirichlet parameterization, which, among other things, restricts correlations to be negative. To overcome this limitation, we propose a novel mixed-membership model that incorporates an extended flexible Dirichlet distribution as the prior for community composition. This distribution is an identifiable mixture with Dirichlet components, which allows for positive correlations, thus enabling the modeling of co-occurrence relationships between taxa. Moreover, its conjugacy to the multinomial model allows the implementation of an efficient collapsed Gibbs sampler for inferential purposes. The proposed model enhances flexibility in community detection and the interpretation of taxa associations, thereby contributing significantly to a deeper understanding of gut microbiome complexity.

A New Prior For Bayesian Graphical Modeling: The S-Bartlett

Authors:

Pierfrancesco Alaimo Di Loro[1*†], Gianluca Mastrantonio[2], Marco Mingione[3]

[1] LUMSA University, Department GEPLI
[2] Politecnico di Torino, Mathematical Science Department
[3] Roma Tre University, Political Science Department

[*] Corresponding author [†] Presenter

Contact: p.alaimodiloro@lumsa.it

Keywords:

Graphical Modeling, Sparse Positive Definite Matrix, Bayesian Hierarchical Modeling, Cholesky

Abstract:

We introduce a general strategy for defining distributions over the space of sparse symmetric positive definite matrices. Our method utilizes the Cholesky factorization of the precision matrix, imposing sparsity through constraints on its elements while preserving their independence and avoiding the numerical evaluation of normalization constants. In particular, we develop the S-Bartlett as a modified Bartlett decomposition, recovering the standard Wishart as a particular case. By incorporating a Spike-and-Slab prior to model graph sparsity, our approach facilitates efficient Bayesian estimation through a tailored MCMC routine based on a transformational Dual Averaging Hamiltonian Monte Carlo update. This framework extends naturally to the Generalized Linear Model setting, enabling applications to non-Gaussian outcomes via latent Gaussian variables. The method is tested through an extensive simulation study, highlighting how the S-Bartlett prior offers a computationally efficient and flexible alternative for estimating sparse precision matrices. It is then applied to gene expression data, revisiting previous results obtained via standard approaches.

Clustering Microbiome Data Via Diversity-Based Mixture Models

Authors:

Silvia Dallari[1*†], Laura Anderlucci[1], Angela Montanari[1]

[1] Department of Statistical Sciences, University of Bologna

[*] Corresponding author [†] Presenter

Contact: silvia.dallari2@unibo.it

Keywords:

Model-based clustering, Diversity measures, Microbiome data

Abstract:

Research on the gut microbiome is becoming essential for gaining insights into human health and into how the microbiota may influence biological systems and disorders. Specifically, finding microbiome communities that can be grouped to uncover community-types associated with specific health or environmental conditions is of interest. A key characteristic of biological communities is the biodiversity. In the literature, diversity measures are generally divided into two main classes: the class of *alpha* and the one of *beta* diversities. The former type of diversity evaluates the within-sample richness of taxa and can be quantified via, e.g., the Shannon-Wiener diversity index. The second category, i.e. the class of *beta*-diversities, measures differences between two or more samples, thereby enabling the description of how many taxa are common between communities or individuals. In ecological studies a widely used *beta*-diversity measure is the Bray-Curtis dissimilarity. In this work, the latter is employed in the definition of a distance-based density function and extended to the model-based clustering framework as a Bray-Curtis dissimilarity-based mixture model. The proposed method can be further extended to incorporate the *alpha*-diversity information as a covariate in the mixture weights. Indeed, healthy individuals are expected to present a high degree of gut microbiome heterogeneity (i.e., high *alpha*-diversity). Therefore, integrating the *alpha*-diversity in the model may help, for example, in differentiating individuals exhibiting dysbiotic gut microbiome configurations from those with healthy microbiome compositions. A thorough simulation study has been performed to test the ability of the model in recovering the true clustering structure and the potential of the proposal is demonstrated through a real data application.

Session - Regularization and latent variables

Organizer: Andreas Alfons

Beyond Regularization: Inherently Sparse Principal Component Analysis

Authors:

Jan O. Bauer[1*†]

[1] VU Amsterdam

[*] Corresponding author [†] Presenter

Contact: j.bauer@vu.nl

Keywords:

HDLSS, Principal Component Analysis, Singular Value Decomposition, Sparse Principal Component Analysis

Abstract:

Sparse principal component analysis (sparse PCA) is a widely used technique for dimensionality reduction in multivariate analysis, addressing two key limitations of standard PCA. First, sparse PCA can be implemented in high-dimensional low sample size settings, such as genetic microarrays. Second, it improves interpretability as components are regularized to zero. However, over-regularization of sparse singular vectors can cause them to deviate greatly from the population singular vectors, potentially misrepresenting the data structure. Additionally, sparse singular vectors are often not orthogonal, resulting in shared information between components, which complicates the calculation of variance explained. To address these challenges, we propose a methodology for sparse PCA that reflects the inherent structure of the data matrix. Specifically, we identify uncorrelated submatrices of the data matrix, meaning that the covariance matrix exhibits a sparse block diagonal structure. Such sparse matrices commonly occur in high-dimensional settings. The singular vectors of such a data matrix are inherently sparse, which improves interpretability while capturing the underlying data structure. Furthermore, these singular vectors are orthogonal by construction, ensuring that they do not share information. We demonstrate the effectiveness of our method through simulations and provide real data applications.

Sparse Clusterpath Gaussian Graphical Modeling And Covariance Estimation

Authors:

Daniël J.W. Touw[1], Andreas Alfons[1*†], Patrick J.F. Groenen[1]
Ines Wilms[2]

[1] Erasmus University Rotterdam
[2] Maastricht University

[*] Corresponding author [†] Presenter

Contact: alfons@ese.eur.nl

Keywords:

hierarchical clustering, precision matrix, covariance matrix, block structure, unsupervised learning

Abstract:

Graphical models serve as effective tools for visualizing conditional dependencies between variables. However, as the number of variables grows, interpretation becomes increasingly difficult, and estimation uncertainty increases due to the large number of parameters relative to the number of observations. To address these challenges, we introduce the Clusterpath estimator of the Gaussian Graphical Model (CGGM) that encourages variable clustering in the graphical model in a data-driven way. Through the use of an aggregation penalty, we group variables together, which in turn results in a block-structured precision matrix. A unique feature of our estimator is that the found block structure is preserved in the covariance matrix. The CGGM estimator is formulated as the solution to a convex optimization problem, making it easy to incorporate other popular penalization schemes. We illustrate this flexibility through the combination of an aggregation and a sparsity penalty. To this end, we achieve simplified interpretation of the estimated graphical model via both node aggregation and edge sparsity. Importantly, we present a computationally efficient implementation of the CGGM estimator by using a cyclic block coordinate descent algorithm. In simulations, we find that CGGM not only matches but oftentimes outperforms other state-of-the-art methods for variable clustering and sparsity in graphical models. Furthermore, we present some challenges when an approximate block structure is present in the covariance matrix (rather than the precision matrix) as the primary structure of interest, and we discuss how the CGGM algorithm can be used for finding a block-structure in covariance matrices. In this context, the found block structure implies that the variables of a given block have equal loadings in latent factor models. We illustrate the practical versatility of CGGM on data from a survey in the behavioral sciences.

Block-Regularized Exploratory Approximate Factor Analysis For Multidomain Data

Authors:

Tra Le[1*†], Jeroen Vermunt[1], Katrijn Van Deun[1]

[1] Tilburg University

[*] Corresponding author [†] Presenter

Contact: t.t.le_1@tilburguniversity.edu

Keywords:

high-dimensional data, latent variable modeling, multidomain data, regularization, approximate factor model

Abstract:

The current trend of intensive data collection brings both opportunities and challenges for behavioral science research. On the one hand, behavior and cognition are no longer studied from the psychological perspective only, but also from other disciplinary perspectives such as environmental, social, clinical, and biomolecular. This often leads to so-called high-dimensional multidomain data. In analyzing this type of data, it is of great importance to disentangle unique mechanisms underlying each data domain from common mechanisms shared by all (or multiple) data domains. Current latent variable methods are not appropriate to address this challenge. To this end, we propose a Block-regularized Exploratory Approximate Factor Analysis (BREAFA). The method adopts the approximate factor model framework with hard cardinality constraints to impose sparsity across and within data domains. That is, the model is estimated under the constraint that exactly C blocks of loadings are equal to zero to identify shared and unique mechanisms. In addition, within each data domain, exactly K loadings are imposed to be zero to encourage variable selection to ease interpretation. Model selection is done using the Index of Sparseness. Both factor scores and loadings are estimated in an alternating optimization scheme. The performance of the proposed method is evaluated in an extensive simulation study in comparison with other methods such as JIVE, RegularizedSCA, and SCD-CovR. We also demonstrate the use of the method using two empirical datasets.

A Generalized Additive Partial-Mastery Diagnostic Classification Model

Authors:

Camilo Cardenas-Hurtado[1*†], Irini Moustaki[1], Yunxiao Chen[1]

[1] London School of Economics and Political Science

[*] Corresponding author [†] Presenter

Contact: c.a.cardenas-hurtado@lse.ac.uk

Keywords:

Diagnostic classification models, Latent variable models, Restricted latent class models, Semi-parametric models, Item response theory

Abstract:

Diagnostic classification models (DCMs), also known as cognitive diagnosis models (CDMs), are restricted latent class models widely used to measure attributes of interest in diagnostic assessments in education, psychology, biomedical sciences, and related fields. Partial-mastery CDMs (PM-CDMs) are a recently proposed extension of CDMs that allow for partial-mastery levels for each latent attribute. They substantially relax a restrictive assumption in most CDMs that individuals can only have two statuses for each attribute, namely mastery and non-mastery. As a result, PM-CDMs tend to yield a better fit for real-world data and refined measurement of latent attributes of interest. However, despite its great improvement over the traditional CDMs, a PM-CDM is required to specify i) a Q-matrix that describes the relationship between the latent attributes and the items, and ii) parametric item response functions that characterize how the distribution of each item response depends on the relevant latent attributes. Both tasks require domain knowledge about the measurement problem, and mistakes in these specifications may be detrimental to the measurement. This paper proposes a generalized additive partial-mastery CDM (GaPM-CDM) that overcomes both limitations by replacing the parametric item response functions in PM-CDMs with semiparametric generalized additive forms. The new model does not require parametric item response functions nor a Q-matrix specification, while still allowing for partial-mastery levels for each latent attribute. We present real-world applications on self-reported social health and educational testing, and evaluate the model's performance in terms of item response function recovery and measurement accuracy of the latent attributes through extensive simulation studies.

Session - Statistical perspectives on fairness in classification algorithms

Organizer: Anna Gottard

Measuring Discrimination In Decision-Making Algorithms: An Approach Based On Causal Inference

Authors:

Francesco Pauli[1], Roberta Pappadà[1*†]

[1] University of Trieste

[*] Corresponding author [†] Presenter

Contact: rpappada@units.it

Keywords:

Discrimination, Machine learning, Coarsened Exact Matching, Sensitive attribute

Abstract:

Machine learning algorithms are routinely used for business decisions that may directly affect individuals in various contexts, such as credit scoring, employment, and criminal justice. When such algorithms are used in the decision process, their behavior concerning discrimination depends on the information it is given, and discrimination may occur unconsciously or explicitly based on sensitive attributes. Statistical tools and methods are then required to handle such potential biases. We propose to exploit the Coarsened Exact Matching (CEM) algorithm to measure discrimination against a protected group to be used in data pre-processing for discrimination detection and removal. Experiments are conducted to test the proposed methodology on real data and a comparison with related work is also discussed.

Society-Centered Ai: An Integrative Perspective On Algorithmic Fairness

Authors:

Isabel Valera[1*†]

[1] Saarland University

[*] Corresponding author [†] Presenter

Contact: ivalera@cs.uni-saarland.de

Keywords:

fair ML, algorithmic decision making, society-centered ML

Abstract:

In this talk, I will share my never-ending learning journey on algorithmic fairness. I will give an overview of fairness in algorithmic decision making, reviewing the progress and wrong assumptions made along the way, which have led to new and fascinating research questions. Most of these questions remain open to this day, and become even more challenging in the era of generative AI. Thus, I will provide only few answers but many open challenges to motivate the need for a paradigm shift from owner-centered to society-centered AI. With society-centered AI, I aim to bring the values, goals, and needs of all relevant stakeholders into AI development as first-class citizens to ensure that these new technologies are at the service of society. To that end, I will show how to optimize stochastic polices to jointly maximize both the decision-maker utility and the welfare of decision subjects in a multi-objective approach, thus leading to better fairness-utility trade-offs. The resulting decision-making mechanisms improve fairness by leveraging the uncertainty of decision outcomes without compromising (the decision-maker's) utility.

Removing The Influence Of Sensitve Attributes Via Variational Approximations

Authors:

Emanuele Aliverti[1*†]

[1] Dipartimento di Scienze Statistiche, Università degli Studi di Padova

[*] Corresponding author [†] Presenter

Contact: emanuele.aliverti@unipd.it

Keywords:

Bayesian inference, binary regression, factor model, fairness, variational inference

Abstract:

In many application areas, predictive models are used to support or automate critical decision-making processes. However, there is growing awareness that such models may encode unwanted associations with sensitive attributes, such as gender or ethnicity. In this talk, I will illustrate statistical adjustments leveraging Variational Inference (VI). In particular, VI can be used to target a purposely misspecified posterior distribution that imposes independence among parameters of interest and sensitive attributes, leading to algorithms whose predictions are independent from such information. The proposed methods are illustrated in the context of Gaussian factor models and semiparametric binary regression, providing efficient algorithms that scale well with the number of observations.

Multi-Class Classification Under System Constraints: a Unified Approach Via Post-Processing

Authors:

Evgenii Chzhen[1*†], Mohamed Hebiri[2], Gayane Taturyan[2]

[1] CNRS
[2] Université Gustave Eiffel

[*] Corresponding author [†] Presenter

Contact: evgenii.chzhen@cnrs.fr

Keywords:

fairness, constraints, optimization

Abstract:

We study the problem of multi-class classification under system-level constraints expressible as linear functionals over randomized classifiers. We propose a post-processing approach that adjusts a given base classifier to satisfy general constraints without retraining. Our method formulates the problem as a linearly constrained stochastic program over randomized classifiers, and leverages entropic regularization and dual optimization techniques to construct a feasible solution. We provide finite-sample guarantees for the risk and constraint satisfaction under minimal assumptions. The framework accommodates a broad class of constraints, including fairness, abstention, and churn requirements.

Session - Analysis of complex data

Organizers: Domenico Perrotta and Andrea Cerioli

Robust Estimation Of Mixed Models

Authors:

Fabrizio Laurini[1*†], Agustín Mayo Iscar[2], Luis Angel García-Escudero[2]

[1] University of Parma
[2] University of Valladolid, Valladolid, Spain

[*] Corresponding author [†] Presenter

Contact: fabrizio.laurini@unipr.it

Keywords:

Mixed models, Robustness, Trimming

Abstract:

Parameter estimation for mixed models can be severely biased by outliers. The outliers can be in the conditional distribution of the response variable or in the inter-individual variability. Our contribution is a maximum likelihood estimator modified in order to get resistance to the effect of outliers by the introduction of trimming. It trims fixed proportions of observations for addressing both of the sources of outlyingness separately, the mentioned ones related to intra and inter individual variability. The corresponding levels of trimming are input hyper-parameters and they can be tuned and monitored with proper diagnostics. We show some results with artificial data and compare with available competitors. We also apply the proposed robust method to real data, based on trade data of the European Union.

Weighted Likelihood Estimation Of Multivariate Location And Scatter With Simultaneous Outliers And Missing Values.

Authors:

Luca Greco[1*†], Claudio Agostinelli[2]

[1] University Giustino Fortunato - Benevento
[2] Department of mathematics, University of Trento

[*] Corresponding author [†] Presenter

Contact: l.greco@unifortunato.eu

Keywords:

Robustness, Missing values, Weighted likelihood

Abstract:

The presence of outliers and missing values constitutes a fundamental challenge in statistical analysis, potentially compromising the validity of inferential conclusions. These data inadequacies can occur simultaneously, requiring integrated methodological approaches for their proper treatment. Outliers may be classified as either case-wise, affecting entire observational units, or cell-wise, involving isolated anomalous values within otherwise valid observations. While case-wise contamination renders complete observations unreliable, cell-wise outliers preserve the informational content of unaffected cells within the same observation. Importantly, in high-dimensional settings, even modest rates of cell- wise contamination can induce significant case-wise outlier propagation through the accumulation of anomalous values across dimensions. Missing values present a distinct but related challenge, necessitating appropriate imputation strategies when the missingness mechanism is ignorable. Our methodological development operates under the Missing Completely at Random (MCAR) framework, where the probability of missingness is independent of both observed and unobserved data values. This work addresses the simultaneous occurrence of outliers and missing values in the estimation of multivariate location and scatter parameters within elliptical distributions. Building upon the well- established Expectation-Maximization (EM) algorithm framework, we incorporate robust estimation through weighted likelihood methods. Specifically, weighted likelihood estimation is a soft-trimming technique that down-weights observations based on their agreement with the assumed model, thereby reducing the influence of anomalous data points. For case-wise contamination, we develop an enhanced EM algorithm incorporating weights updates, thereby providing protection against anomalous data points while properly accounting for missing values. Alternative weighting schemes are considered. In the cell-wise contamination scenario, we implement a two-stage approach: first, potential outlying cells are identified using optimized filtering criteria; second, these cells undergo snipping, a process where identified outliers are treated as missing values, prior to final estimation using robust case-wise techniques. Our contribution advances the field of robust statistical analysis by providing a unified framework for handling these dual data quality challenges, with particular relevance to modern high-dimensional datasets in data science applications.

Robust Clustering Based On Trimming With Increasing Dimensionality

Authors:

Luis Angel García-Escudero[1*†], Agustín Mayo Iscar[1], Lucia Trapote-Reglero[1]

[1] Universidad de Valladolid

[*] Corresponding author [†] Presenter

Contact: lagarcia@uva.es

Keywords:

robustness, model-based clustering, trimming, high-dimensional data, anomaly detection

Abstract:

Outliers are known to significantly distort the results of many widely applied Cluster Analysis methods. While increasing the number of clusters to be detected might seem like a straightforward way to accommodate outliers, this strategy is often ineffective and frequently impractical. To address this issue, robust clustering techniques have been developed, which not only mitigate the influence of outliers on clustering results but also help detect meaningful anomalies in data, especially when subpopulations are naturally present. This presentation focuses on robust clustering methods based on trimming, with particular emphasis on TCLUST, an extension of the Minimum Covariance Determinant (MCD) method designed for Cluster Analysis. TCLUST is highly effective for low-dimensional datasets, but its performance deteriorates in high-dimensional settings due to the complexity involved in estimating the scatter matrices of multiple components. Although constraints on the eigenvalues of the scatter matrices can help, such an approach forces clusters to be nearly spherical with equal dispersion. An alternative approach is the Robust Linear Grouping (RLG) algorithm, which assumes that clusters lie around lower-dimensional affine subspaces, combining clustering with dimensionality reduction. However, the RLG approach assumes errors orthogonal to the approximating subspaces and can face challenges with intersecting subspaces. A new approach will also be presented that offers a compromise between TCLUST and RLG. This compromise results from robustifying the High Dimensional Data Clustering (HDDC) approach, already available in the literature, through the implementation of trimming and the enforcement of additional constraints on eigenvalues. The HDDC approach assumes parsimonious conditions on the scatter matrices, making it tractable in higher dimensions, but can benefit from convenient robustification. An algorithm implementing this novel robust approach will be introduced, along with illustrative examples and diagnostics for detecting outlyingness of interest. In this algorithm, it is important to provide adequate random initialization and to robustly estimate the intrinsic dimensions of the approximating affine subspaces. Furthermore, the presentation will also address how replacing traditional casewise trimming with cellwise trimming can help retain more information by discarding only individual cells instead of entire observations, which is particularly important as dimensionality increases.

Session - Statistical approaches for measuring and analysing educational imbalance

Organizer: Rosaria Romano

Enhancing Student Resilience Through Socio-Emotional Skills: Evidence From Pisa 2022

Authors:

Tommaso Agasisti[1*†], Mara Soncin[1], Chiara Masci[2]
Sergio Longobardi[3]

[1] Politecnico di Milano School of Management
[2] University of Milan
[3] University Naples Parthenope

[*] Corresponding author [†] Presenter

Contact: tommaso.agasisti@polimi.it

Keywords:

Socio-Emotional Skills, Student Resilience, Multilevel Analysis, Machine Learning Techniques, Disadvantaged Students, OECD PISA

Abstract:

Student resilience - the ability of students from disadvantaged backgrounds to achieve strong academic outcomes despite socio-economic challenges - remains a critical concern in educational research and policy. While cognitive skills and academic preparedness are well-documented predictors of student success, growing evidence suggests that socio-emotional competencies also play a fundamental role in fostering resilience. This study explores the relationship between socio-emotional skills and academic resilience by analyzing data from the 2022 OECD PISA assessment. By employing both traditional regression models and advanced machine learning techniques, our analysis provides a comprehensive examination of the main factors influencing student resilience. The findings reveal that socio-emotional skills (such as perseverance, self-efficacy, and emotional regulation) significantly contribute to academic success, particularly for students from disadvantaged backgrounds. Moreover, integrating machine learning methods allows for a more nuanced understanding of the complex interactions between socio-economic status, school characteristics, and student outcomes. Our results confirm student-level control variables' expected direction and magnitude, aligning with existing literature and reinforcing their robustness. At the school level, variables related to extracurricular activities and school climate consistently emerge as strong predictors of resilience. By highlighting the potential of targeted interventions aimed at enhancing these skills, our study provides valuable guidance for policymakers and educators: implementing structured programs that integrate socio-emotional learning into school curricula could serve as a powerful tool to bridge educational disparities and promote resilience among disadvantaged students.

Discovering Profiles Of Resilient Students In Pisa: An Information-Theoretic Approach To Clustering Mixed-Type Educational Data

Authors:

Angelos Markos[1*†], Efthymios Costa[2], Ioanna Papatsouma[2]

[1] Democritus University of Thrace
[2] Imperial College London

[*] Corresponding author [†] Presenter

Contact: amarkos@eled.duth.gr

Keywords:

Academic resilience, Information-theoretic clustering, Mixed-type data methods, PISA data analysis, Educational equity

Abstract:

Academic resilience-where students from socioeconomically disadvantaged backgrounds perform well on international assessments-is a crucial focus in educational equity research. Yet most empirical analyses of resilience rely on threshold-based definitions and linear modeling, approaches that obscure the heterogeneity of disadvantaged student experiences and often fail to respect the diverse measurement types present in educational data. PISA datasets exemplify this challenge: they combine continuous variables (achievement scores, motivation, belonging) and categorical attributes (gender, immigrant background). We focus on uncovering profiles of disadvantaged students across three PISA cycles (2015, 2018, 2022) in Greece, capturing the country's transition from economic recovery through pandemic disruption, and examining how resilience patterns evolve over time within a single educational system-addressing a critical gap in the literature where virtually no studies examine temporal dynamics of resilience profiles. Our approach employs information-theoretic clustering methods designed to handle mixed-type educational data while preserving the natural structure of different measurement types.

Assessing Policies For Schools With Low Socioeconomic Opportunities: Insights From Invalsi Tests

Authors:

Pasquale Sannino[1*†], Cristina Davino[2], Rosaria Romano[2]

[1] Università di Macerata
[2] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: p.sannino2@unimc.it

Keywords:

Regression Discontinuity, INVALSI data, Discrete Running Variable

Abstract:

Assessing students' competencies is one of the fundamental tasks for policymakers. In fact, the issue of inequalities in the education sector represents one of the most pressing and challenging problems to address. Educational inequalities, often linked to socioeconomic, geographic, or cultural factors, can undermine not only students' opportunities but also a country's social cohesion and economic development. This study analyzes the results of Italian students in the INVALSI Mathematics tests, with a particular focus on the students' socioeconomic and cultural background and how this influences test outcomes. More specifically, the study aims to analyze whether the ministry's intervention of suggesting smaller class sizes for schools attended by students with more disadvantaged socioeconomic and cultural conditions helps to reduce the performance gap among students. To do this, we use the Regression Discontinuity (RD) design, which has recently emerged as one of the most credible non-experimental strategies for the analysis of the causal effects. The RD approach allows for the estimation of the causal effect of a covariate, known as the running variable or score, on an outcome by comparing results just above and just below a known and exogenous threshold. In this specific application, we focus on the discrete nature of the running variable and the complications this introduces in RD designs, given that the socioeconomic and cultural background is inherently discrete. We propose possible solutions that enable effective estimation and management of the causal effect despite the discrete nature of the running variable, and we strengthen our analysis with a simulation study.

Advancing Educational Research With Multilevel Quantile Regression: Evidence From Large-Scale Data

Authors:

Clelia Cascella[1*†], Marco Geraci[2], Domenico Vistocco[3]

[1] Italian National Institute for Evaluation of Educational System (INVALSI), Rome, Italy
[2] Sapienza University of Rome, Rome, Italy
[3] University of Naples Federico II, Naples, Italy

[*] Corresponding author [†] Presenter

Contact: Clelia.cascella@invalsi.it

Keywords:

Multilevel quantile analysis, Multilevel modelling, Quantile regression, Large-scale assessment data

Abstract:

Hierarchical data are commonly encountered in research domains such as education, sociology, healthcare, and psychology. The complexities inherent in such data structures require advanced analytical methodologies. This paper presents an application of multilevel quantile regression to large-scale assessment (LSA) data and examines its added value in comparison to traditional multilevel modelling and standard quantile regression. Multilevel modelling - also known as hierarchical modelling - is widely used to analyse nested data structures, such as students within schools or patients within hospitals. It allows researchers to incorporate both individual-level and group-level variables, accounting for intra-group correlations. Quantile regression, on the other hand, estimates conditional quantiles of the response variable, providing a more nuanced understanding of the distribution than models focusing solely on the mean. In this study, we employ multilevel quantile regression, which integrates the strengths of both approaches, making it particularly suitable for hierarchical data structures. This method is especially valuable when analysing LSA data. While traditional multilevel models estimate effects at the average level, multilevel quantile regression reveals how predictors influence different points of the distribution - such as the lower, median, and upper quantiles - allowing for richer insights. For example, it enables researchers to explore how personal and contextual student factors affect various levels of achievement, rather than just the average performance. This granularity is crucial in educational research. For instance, mathematics education studies have shown that gender differences in performance on specific items are more pronounced at the upper end of the ability distribution than at the lower or average levels. Such distributional effects are often obscured in traditional models. Standard approaches may fail to detect these subtle but important variations, potentially leading to misleading interpretations and policy implications. In contrast, multilevel quantile regression provides a powerful tool for managing the complexity of LSA data, offering a more detailed and accurate understanding of educational phenomena. This paper offers empirical evidence supporting the use of multilevel quantile regression and critically discusses its contributions relative to both traditional multilevel and quantile regression approaches.

Session - Data-Driven classification and statistical modeling for tackling environmental challenges

Organizer: Luca Merlo

Flexible And Robust Modeling Of Tidal Meteorological Residuals In The Venice Lagoon Using Hidden Semi-Markov Models

Authors:

Antonello Maruotti[1], Lorena Ricciotti[2*†], Alfonso Russo[3]
Sondre S. Hølleland[4]

[1] Libera Università Maria Ss. Assunta
[2] Università degli Studi di Bari
[3] Università di Roma Tor Vergata
[4] NHH Norwegian School of Economics

[*] Corresponding author [†] Presenter

Contact: lorena.ricciotti@uniba.it

Keywords:

Hidden Semi-Markov Models, Sea Level Regimes, Elastic-net Regularization, Clustering, Sea Conditions

Abstract:

This study introduces a robust and flexible statistical modeling framework to analyze sea level dynamics in the Venice Lagoon, with a specific focus on the tide component influenced by meteorological conditions. We propose a novel class of robust Hidden Semi-Markov Regression Models (HSMRMs) capable of capturing key features of marine data such as regime-switching behavior, time-varying heteroskedasticity, heavy tails, skewness, and outliers. The framework extends traditional Hidden Markov Regression Models by relaxing the assumption of geometrically distributed sojourn times and by incorporating variable selection through elastic-net regularization. To enhance robustness against outliers and skewed data, we ran the model using the conventional Gaussian distribution and the heavy-tailed Student-t and Johnson's SU distributions. These distributions allow for more accurate modeling of both the central tendency and variability of sea levels under different environmental regimes. Empirical analysis is conducted using hourly data from a tide gauge at the Lido di Porto inlet, covering various meteorological variables such as wind speed and direction, air and water temperature, pressure, and humidity. The proposed model identifies four distinct environmental regimes influencing meteorological residuals, each associated with specific weather conditions and temporal dynamics. For example, one regime captures extreme residual events associated with the Bora wind, while others reflect typical conditions of the Venice Lagoon. The model also distinguishes between true and apparent contagion in sea level data by incorporating autoregressive components, thereby addressing both latent regime shifts and explicit temporal dependencies. Simulation studies confirm the ability of the Johnson's SU-based HSMRM in parameter estimation accuracy and classification performance under heavy-tailed, skewed data generation processes. The regularization approach effectively selects relevant covariates and lags, enhancing model interpretability and forecasting ability.

Robust Clustering Using Maximized Mutual Information

Authors:

Mackenzie Neal[1*†], Paul McNicholas[1], Arthur White[2]

[1] McMaster University
[2] Trinity College Dublin

[*] Corresponding author [†] Presenter

Contact: nealm6@mcmaster.ca

Keywords:

Robust clustering, Mutual information, Mixture models, Flow cytometry

Abstract:

Often, we can describe clustering algorithms as using either a generative approach or a discriminative approach. A generative approach provides information on geometric properties of clusters, whereas a discriminative approach aims to determine boundaries between clusters. We incorporate ideas from both approaches to present a fully unsupervised probabilistic, discriminative clustering method capable of capturing irregular sub-populations common in biological datasets such as those arising from flow cytometry experiments. Flow cytometry experiments use lasers to produce cellular light signals that are then converted into electronic signals and analyzed. As these experiments increase in complexity, so does the data, making these datasets increasingly more difficult to cluster via traditional methods, such as sequential manual gating. This is because sequential manual gating is incredibly subjective, irreproducible, and highly influenced by the order of marker selection. We overcome these problems by using a regularized mutual information objective function and model each component density as a Gaussian uniform mixture. In doing so, we introduce a fully unsupervised clustering algorithm that is robust to outliers and high-density regions. We demonstrate this robustness on various simulated datasets. Since the proposed method pulls from both generative and discriminative clustering, we can capture well-separated, highly intuitive sub-populations, while also obtaining cluster characteristics. Although flow cytometry is the primary motivation for the work proposed herein, this method can be applied to any dataset wherein the goal is to obtain intuitive clustering solutions. We demonstrate this by comparing the proposed work to popular clustering methods on various simulated and real datasets.

Climate-Risk Salience And Public Support For Mitigation: Causal Evidence From The 2021 German Floods

Authors:

Giulio Grossi[1*†], Lea Anna Cozzucoli[2]

[1] University of Florence
[2] University of Trieste

[*] Corresponding author [†] Presenter

Contact: giulio.grossi@unifi.it

Keywords:

Bayesian causal inference, extreme weather, floods, spatial latent factors, climate-policy attitudes

Abstract:

Understanding how perceived climate risk shapes citizens' willingness to engage in mitigation is a defining challenge of the twenty-first century. While the frequency of extreme weather events is rising, many communities newly exposed to such hazards have limited experience from which to form risk perceptions. We study whether direct exposure to extreme climatic events-specifically floods and severe storms-alters public beliefs about climate change and attitudes toward more ambitious, including self-protective, mitigation actions. Prior research has documented correlations between disasters and pro-environmental attitudes, yet few studies provide clear causal identification. Leveraging the spatial nature of both treatment (disaster exposure) and outcomes (attitudinal measures), we develop a Bayesian latent-factor framework that imputes the counterfactual attitude matrix by exploiting shared spatial and cross-item structure. This approach allows us to isolate treatment effects while fully accounting for latent heterogeneity and spatial spillovers. Additionally, we provide insights about the link between our proposal, balancing weights and outcome regression methods for estimating causal effects. Our empirical application examines the devastating effects of July 2021 floods in Germany. Combining high-resolution flood-impact data with the ARIADNE climate-attitude survey (Hertie School), we compare directly and peripherally affected areas with unaffected counties to estimate the causal impact of flood exposure on support for environmental policies. The study demonstrates how Bayesian spatial counterfactual imputation can uncover causal effects in settings where random assignment is impossible, and offers fresh evidence on the relationship between climate hazards can mobilize public support for mitigation. Beyond its substantive relevance for climate-risk perception, the approach is transferable to any setting where geographically referenced interventions interact with spatially correlated attitudes or behaviors.

Discrete Latent Variable Models For Time-Dependent Ranking Data

Authors:

Alfonso Russo[1*†], Alessio Farcomeni[2], Sabrina Giordano[1]
Antonello Maruotti[3]

[1] Università della Calabria
[2] Università degli Studi di Roma Tor Vergata
[3] Università LUMSA

[*] Corresponding author [†] Presenter

Contact: alfonso.russo@uniroma2.it

Keywords:

Discrete Latent Variables, Ranking Data, Markov Chain Monte Carlo

Abstract:

Ranking data arise whenever a collection of items, such as universities, institutions, teams, products, or countries are ordered according to criteria like quality, preference, or performance. While static ranking models are well established, there is growing interest in understanding how rankings evolve over time. Time dynamics are often driven by unobserved heterogeneity. The standard Plackett-Luce model provides a principled way to parsimoniously define probability distributions over rankings. Bayesian formulations allow for uncertainty quantification and can accommodate partial rankings. Recent extensions incorporate the temporal structure using continuous latent trajectories, but such approaches often require many parameters, may struggle to detect structural shifts, and rely on parametric distributional assumptions for the latent variables. We formulate a time-dependent Plackett-Luce model that is modulated by a discrete latent variable, which can flexibly evolve over time. Conditionally on the latent variable rankings are assumed to be independent of the past, and to follow a Plackett-Luce model with state-specific weight parameters. The proposed framework captures both persistent behaviour and distinct changes in the ranking mechanism. Other advantages include a reduced number of parameters, a natural segmentation of the time series which enhances interpretation, and the possibility to approximate the true underlying distribution arbitrarily well. Bayesian inference is carried out via a Markov Chain Monte Carlo scheme alternating between latent structure updates and conditional sampling of the emission parameters. This enables probabilistic conclusions about the latent process driving the observed rankings and accommodates partial or irregular observations in a coherent way. Overall, our framework offers a flexible and efficient approach for modelling dynamic rankings, combining the regime-switching capability of Markovian models with the structure and interpretability of the Plackett-Luce likelihood.

Session - Biomedical data analysis and systems biology

Organizer: Hans Kestler

Linear Classification Algorithms For Ordinal Classifier Cascades

Authors:

Ludwig Lausser[1*†]

[1] Technische Hochschule Ingolstadt

[*] Corresponding author [†] Presenter

Contact: Ludwig.Lausser@thi.de

Keywords:

Ordinal Classification, Linear Classification, Gene Expression Analysis

Abstract:

Ordinal classifier cascades are constraint multiclass classifier systems specialized in distinguishing ordinal categories such as stagings or gradings. Due to their structural properties, these models depend on a predefined class order. If the feature representation does not reflect this order, the corresponding ordinal classifier cascade will experience significantly reduced classification performance. They can therefore serve as indicators for ordinality. When used in exploratory screenings, they can even detect unknown ordinal relations among classes. Various base classifiers can be combined with ordinal classifier cascades, resulting in different classification outcomes. In a previous study, linear classification models performed best at identifying ordinal relationships among gene expression profiles. This work offers a more detailed analysis of that finding. It investigates how different training algorithms influence the performance of ordinal classifier cascades. An empirical study compares their classification accuracy and their ability to detect ordinal relations.

A Sparse Explainable Ensemble Classifier Leveraging Noise (And a Representation In The Decision Function Space) For High-Dimensional Data

Authors:

Annika Kestler[1*†], Hans A. Kestler[2]

[1] Ulm University
[2] University of Ulm, University Hospital Ulm

[*] Corresponding author [†] Presenter

Contact: annika.kestler@uni-ulm.de

Keywords:

decision tree, clustering, function space, explainability, ensemble

Abstract:

Sparse "if-then" decision rules are human-readable, hence can be considered explainable classifiers, opposed to so-called black-box models. A simple tree classifier trained on high-dimensional data (meaning each sample is high-dimensional; each dimension we denote as a feature), will only ever result in a tree with a maximum depth of "number of samples - 1", and is therefore both highly interpretable as well as it intrinsically performs a feature selection. Still, it has high variance and low bias; Hence, minor changes in the training data will potentially greatly affect the decision function, and the identified rules likely overfit the data. Therefore, we propose a classifier that leverages the high interpretability and feature selection properties while constructing a sparse ensemble decision function that is less susceptible to noise and exhibits a particular grouping in the decision function space. This is achieved by explicitly incorporating noise during the training. Here, multiple decision trees are trained on noisy datasets with varying degrees of noise, and their corresponding decision rules are extracted. We then cluster this subset of the decision function space to assess the function space and make statements on cluster-specific feature importance. We further identify class-specific minimal functions per cluster, which can result in functions relying only on proper subsets of the features relevant to the considered cluster. To obtain the final ensemble classifier, we perform a majority vote on all the identified class- and cluster-specific functions (allowing "cannot decide" outcomes).

Semantic Data Integration For The Reconstruction Of Gene Regulatory Networks

Authors:

Max Krüger[1*†], Ludwig Lausser[2]

[1] Technische Hochschule Ingolstadt, Germany
[2] Technische Hochschule Ingolstadt

[*] Corresponding author [†] Presenter

Contact: max.krueger@thi.de

Keywords:

Semantic data integration, Gene regulatory networks, Bayesian networks

Abstract:

Gene regulatory networks direct most of the intracellular molecular processes. They consist of multiple components that interact through complex schemes and dynamics. These interactions are often not directly observable due to the invasive nature of molecular measurements. They must be estimated or reconstructed from static molecular profiles, such as gene expression data. This data-driven modelling task is most challenging due to the high dimensionality and the low cardinality of the available data collections (n»m). In this work, we explore the possibility of simplifying the reconstruction of gene regulatory networks through semantic data integration. Specifically, we incorporate prior knowledge about molecular interactions into the data-driven training process of Bayesian networks. This approach narrows the hypothesis space to a focused set of potential candidate networks, thereby decreasing the risk of overfitting. We empirically evaluate our method on publicly available benchmark datasets and analyze the structural differences between the reconstructed networks.

Simulating a Boolean Network Of Hematopoietic Stem Cell Regulation On Neuromorphic Hardware

Authors:

Caya L. O. Hotstegs[1], Johann M. Kraus[1*†], Hans A. Kestler[1]

[1] Ulm University

[*] Corresponding author [†] Presenter

Contact: johann.kraus@uni-ulm.de

Keywords:

Boolean networks, Biomedical data analysis, Systems biology, Spiking neural networks, Neuromorphic computing

Abstract:

Spiking neural networks (SNNs) provide a biologically inspired and hardware-efficient alternative to traditional neural network models. They operate using discrete spikes instead of continuous signals. Their event-driven nature makes them especially suitable for execution on energy-efficient neuromorphic hardware such as SpiNNaker. In this work, we demonstrate how SNNs can model Boolean networks and present an implementation on neuromorphic hardware. Our approach simulates the deterministic logic of Boolean networks encoded with spike-based populations. This maintains the semantics of classical Boolean models while enabling integration with neuromorphic hardware. The system converts Boolean logic states into spike patterns. Each variable is represented by neuron populations, and logical functions are constructed using spiking subcircuits. The spiking signals travel through the network, activating or inhibiting other neurons, and ultimately reflecting a synchronous update of the variables. The implementation consists of several stages. First, logic expressions are parsed into structured abstract syntax trees (ASTs). These are then mapped onto a graph of neuron populations and connections, including helper and delay neurons. A global clock synchronizes information flow using spike generators to ensure step-by-step progression. This enables the simulation of complex Boolean networks and the study of their dynamics. The simulation backend is built on PyNN to ensure compatibility with existing neural modeling tools. To demonstrate our system in a real-world example, a Boolean model of hematopoietic stem cell regulation is implemented. Several input scenarios are tested to analyze network transitions between different attractors, depending on external signals. Our simulations demonstrate that logical models can be reliably implemented in spike-based systems. The spiking network accurately reproduced the attractor states and transitions of a Boolean model of a biological system. The timing architecture ensured that all logic operations were executed synchronously. Even with probabilistic inputs, the system followed the expected dynamics. This shows that spiking neural networks are not limited to biologically inspired tasks. They can also represent and solve abstract, rule-based systems. This opens new opportunities for using neuromorphic hardware in areas such as optimization, control, and modeling complex systems.

Session - Advances in preference learning

Organizer: Valeria Vitelli

Bayesian Rank-Clustering

Authors:

Michael Pearce[1*†], Elena A. Erosheva[2]

[1] Reed College
[2] University of Washington

[*] Corresponding author [†] Presenter

Contact: michaelpearce@reed.edu

Keywords:

ordinal comparisons, spike-and-slab, fusion priors, item indifference, learning to rank

Abstract:

Traditional statistical inference on ordinal comparison data results in an overall ranking of objects, e.g., from best to worst, with each object having a unique rank. However, ranks of some objects may not be statistically distinguishable. This could happen due to insufficient data or to the true underlying object qualities being equal. Because uncertainty communication in estimates of overall rankings is notoriously difficult, we take a different approach and allow groups of objects to have equal ranks or be rank-clustered in our model. Existing models related to rank-clustering are limited by their inability to handle a variety of ordinal data types, to quantify uncertainty, or by the need to pre-specify the number and size of potential rank-clusters. We solve these limitations through our proposed Bayesian Rank-Clustered Bradley-Terry-Luce model. We accommodate rank-clustering via parameter fusion by imposing a novel spike-and-slab prior on object-specific worth parameters in Bradley-Terry-Luce family of distributions for ordinal comparisons. We demonstrate rank-clustering on simulated and real datasets in surveys, elections, and sports analytics.

Flexible Models For Multiple Raters Data Via Bayesian Nonparametric Priors

Authors:

Giuseppe Mignemi[1*†], Ioanna Manolopoulou[2]

[1] Bocconi University
[2] University College London

[*] Corresponding author [†] Presenter

Contact: giuseppestamignemi@gmail.com

Keywords:

Bayesian nonparametric models, Bayesian hierarchical models, Bayesian mixture models, Rating models, Intraclass correlation coefficient

Abstract:

Rating procedure is crucial in many applied fields (e.g., educational, clinical, emergency). It implies that a rater (e.g., teacher, doctor) rates a subject (e.g., student, doctor) on a rating scale. Given raters' variability, several statistical methods have been proposed for assessing and improving the quality of ratings. Model estimation in the presence of heterogeneity has been one of the recent challenges in this research line. Consequently, several methods have been proposed to address this issue under a parametric multilevel modelling framework, in which strong distributional assumptions are made. We propose a more flexible model under the Bayesian nonparametric (BNP) framework, in which most of those assumptions are relaxed. By eliciting hierarchical discrete nonparametric priors, the model accommodates clusters among raters and subjects, naturally accounts for heterogeneity, and improves estimate accuracy. We propose a general BNP heteroscedastic framework to analyse continuous and coarse rating data and possible latent differences among subjects and raters. The estimated densities are used to make inferences about the rating process and the quality of the ratings. By exploiting a stick-breaking representation of the Dirichlet Process, a general class of Intraclass Correlation Coefficient (ICC) indices might be derived for these models. Our method allows us to identify latent similarities between subjects and raters independently and can be applied in precise education to improve personalised teaching programs or interventions. Theoretical results regarding the ICC are presented, along with computational strategies. Simulations and a real-world application are presented, and possible future directions are discussed.

Stability Post-Processing For Items Importance In Preference Learning Via The Bayesian Mallows Model

Authors:

Luca Coraggio[1*†], Valeria Vitelli[2]

[1] University of Naples Federico II
[2] University of Oslo

[*] Corresponding author [†] Presenter

Contact: luca.coraggio@unina.it

Keywords:

Bayesian inference, clustering, variable selection, rankings, high-dimensional data

Abstract:

Rank and preference data are becoming increasingly ubiquitous, stimulating continuous advances in preference learning methodologies and their wider adoption across different domains. The Lower-dimensional Bayesian Mallows Models with Mixtures (LowBM3) is a recent extension of the Bayesian Mallows Models, which was originally developed as a unifying Bayesian framework to estimate the Mallows model. LowBM3 extends the Bayesian Mallows Model to ultra-high-dimensional settings, allowing to estimate a clustering of the assessors, and the within-cluster sets of relevant items and their consensus rankings. In this paper, we propose a novel post-processing strategy for LowBM3, named stability post-processing. We validate our methodology through experimental analysis and demonstrate its superior performance in specific settings characterized by significant variability in absolute rankings across assessors.

The Clustered Mallows Model

Authors:

Luiza Piancastelli[1*†], Nial Friel[1]

[1] University College Dublin

[*] Corresponding author [†] Presenter

Contact: luiza.piancastelli@ucd.ie

Keywords:

Mallows model, Ranking data, Bayesian learning, Clustering, Rank aggregation

Abstract:

Rankings represent preferences that arise from situations where assessors arrange items, for example, in decreasing order of utility. Orderings of the item set are permutations (pi) that reflect strict preferences. However, strict preference relations can be unrealistic for real data. Common traits among items can justify equal ranks and there can also be different importance attribution to decisions that form pi. In large item sets, assessors might prioritise certain items, rank others low, and express indifference towards the remaining. Rank aggregation may involve decisive judgments in some parts and ambiguity in others. In this paper, we extend the famous Mallows (Biometrika 44:114-130, 1957) model (MM) to accommodate item indifference. Grouping similar items motivates the proposed Clustered Mallows Model (CMM), a MM counterpart for tied ranks with ties learned from the data. The CMM provides the flexibility to combine strictness and indifferences, describing rank collections as ordered clusters. CMM Bayesian inference is a doubly-intractable problem since the normalised model is unavailable. We overcome this with a version of the exchange algorithm (Murray et al. in Proceedings of the 22nd annual conference on uncertainty in artificial intelligence (UAI-06), 2006) and provide a pseudo-likelihood approximation as a computationally cheaper alternative. Analysis of two real-world ranking datasets is presented, showcasing the practical application of the CMM and highlighting scenarios where it offers advantages over alternative models.

Session - Complex environmental data

Organizers: Laura Sangalli and Francesco Lagona

Detecting Changes In Space-Varying Parameters In Seismic Point Processes

Authors:

Nicoletta D'Angelo[1*†]

[1] Università degli Studi di Palermo

[*] Corresponding author [†] Presenter

Contact: nicoletta.dangelo@unipa.it

Keywords:

local analyses, point process, spatial segmentation, spatial statistics, tessellation

Abstract:

Recent advances in local models for point processes have highlighted the need for flexible methodologies to account for the spatial heterogeneity of external covariates influencing process intensity. In this work, we introduce Tessellated Spatial Regression, a novel framework that extends segmented regression models to spatial point processes, with the aim of detecting abrupt changes in the effect of external covariates onto the process intensity. Our approach consists of two main steps. First, we apply a spatial segmentation algorithm to geographically weighted regression estimates, generating different tessellations that partition the study area into regions where model parameters can be assumed constant. Next, we fit log-linear Poisson models in which covariates interact with the tessellations, enabling region-specific parameter estimation and classical inferential procedures, such as hypothesis testing on regression coefficients. Unlike geographically weighted regression, our approach allows for discrete changes in regression coefficients, making it possible to capture abrupt spatial variations in the effect of real-valued spatial covariates. Furthermore, the method naturally addresses the problem of locating and quantifying the number of detected spatial changes. We validate our methodology through simulation studies and applications to two examples where a model with region-wise parameters seems appropriate and to an environmental dataset of earthquake occurrences in Greece.

Environmental Risk Assessment Via Nonhomogeneous Hidden Semi-Markov Models With Penalized Vector Auto-Regression

Authors:

Marco Mingione[1*†], Pierfrancesco Alaimo Di Loro[2], Francesco Lagona[1]
Antonello Maruotti[2]

[1] Roma Tre University
[2] LUMSA

[*] Corresponding author [†] Presenter

Contact: marco.mingione@uniroma3.it

Keywords:

Hidden semi-Markov models, Environmental risk, Air pollution, Penalized Vector Auto-Regression

Abstract:

Accurately assessing the risks associated with pollution exposure is a critical component of environmental research and public health policy. To this end, this study introduces a flexible statistical framework for modeling multivariate air pollution data via a nonhomogeneous hidden semi-Markov model with vector autoregression. The hidden process captures unobserved exposure conditions, while the vector autoregressive structure accounts for temporal autocorrelation and cross-pollutant dependencies. The model further allows time-varying environmental conditions to influence both the average levels of pollutant concentrations and the duration of exposure states. Model parameters are estimated via maximum likelihood using a tailored Expectation-Maximization (EM) algorithm, integrated with state-specific ℓ_1 regularization to control overfitting and automatically select relevant temporal lags. The proposal is tested on simulated data under different scenarios and then applied to daily concentrations of nitrogen (NO and NO_2) and particulate matter (PM_1, $PM_{2.5}$, PM_{10}) recorded from January 1, 2020, to December 31, 2022, at Danmarksplass, the busiest traffic intersection in Bergen (Norway). Environmental risk is assessed by a Shapley value-based decomposition that attributes marginal risk contributions. The results from the empirical analysis illustrate that air quality in Danmarksplass is generally clean, though temporary episodes of elevated pollutant concentrations occur during periods of intensified human activity and unfavorable meteorological conditions. Overall, the proposed model-based approach proves effective in capturing nonstationary, multivariate, and regime-switching behaviors in pollution data. By integrating dynamic risk measures, time-varying dependence structures from a hidden semi-Markov model, and Shapley value decomposition, it also enables interpretable, time-resolved analysis of inter-pollutant risk propagation. This supports more targeted interventions, with broad applicability beyond air quality monitoring.

Clustering Metabarcoding Data: a Model-Based Approach

Authors:

Luisa Ferrari[1*†], Maria Franco-Villoria[1], Garritt L. Page[2]
Alex Laini[3]

[1] University of Modena and Reggio Emilia
[2] Brigham Young University
[3] Università di Torino

[*] Corresponding author [†] Presenter

Contact: luisa.ferrari5@unibo.it

Keywords:

Metabarcoding, Bernoulli Mixture Model, Informative Priors

Abstract:

Analyzing species occurrence data is extremely relevant for monitoring biodiversity and designing conservation plans. The traditional method used to collect this type of data consists of capturing a sample using traps, followed by morphological identification by an expert. Obviously, this methodology is very time-consuming, and for this reason, the datasets collected in the field of ecology are usually quite small. Metabarcoding is a relatively novel technique based on genetic identification, which makes it possible to rapidly retrieve a large amount of information about the occurrence of multiple species from a single sample. The deployment of metabarcoding represents a complete paradigm shift in the study of ecological communities. Metabarcoding-generated datasets represent the future of research in this field, and therefore, new statistical tools specifically designed to address their challenges must be developed.In particular, these datasets consist of binary entries representing the presence or absence of a large number of species across multiple sampled sites. In this work, we focus on model-based clustering methods to explore ecological patterns within metabarcoding data. This can be useful to discover groups of species with similar occurrence patterns, as well as to identify sites with similar species composition-that is, biogeographical areas. Specifically, we consider the Bernoulli Mixture Model, which has already been successfully applied in the fields of image and text mining. We propose to extend the basic model to include environmental covariates collected at the different sampling locations. This allows us to investigate the relationship between the identified clusters and habitat characteristics. Additionally, we consider an informative prior on the number of clusters using the asymmetric Dirichlet prior recently proposed in the literature. To evaluate the soundness of our approach, we first conduct a simulation study comparing the performance of the current standard model in the literature-that is, the basic Bernoulli Mixture Model-with our extended version. Finally, we apply our model to a metabarcoding dataset recording the presence or absence of dung beetle species in the Northwestern Alps of Italy, in order to uncover potential clustering patterns and support ecologists in better understanding the ecological structure of the area.

Spatio-Temporal Regression With Pde Penalization: Mean And Quantile Estimation

Authors:

Eleonora Arnone[1*†], Laura Sangalli[2]

[1] Università di Torino
[2] Politecnico di Milano

[*] Corresponding author [†] Presenter

Contact: eleonora.arnone@unito.it

Keywords:

Spatio-temporal data, Partial differential equations, Quantile regression

Abstract:

In this presentation, we address the problem of modeling spatio-temporal data through flexible regression frameworks that account for anisotropy, non-stationarity, and complex domain geometries. We begin by considering spatio-temporal regression models with differential regularization, where Partial Differential Equations (PDEs) are embedded into the estimation process as penalization terms. This framework enables the inclusion of prior knowledge on the phenomenon under study, such as physical constraints or domain-induced heterogeneity. Through the use of PDE-based roughness penalties, the model adapts to irregular spatial designs and captures anisotropic effects. We further enhance this modeling approach by allowing the PDE operator to depend on unknown parameters, which are estimated directly from the data via parameter cascading. This estimation strategy provides a principled way to integrate incomplete or partially specified physical knowledge, offering greater modeling flexibility while preserving interpretability. Building on this foundation, we extend the methodology to the quantile regression setting. In many real-world applications, like environmental and urban studies, interest lies not only in the mean behavior of the system, but also in its extremes. Quantile regression allows us to model the conditional distribution of the response variable, capturing heteroskedasticity and skewness that are often observed in spatio-temporal data. We propose a semiparametric spatio-temporal quantile regression model where the quantile surface is estimated through PDE-regularized minimization of the pinball loss. We illustrate the effectiveness of the approach with applications in environmental monitoring.

Session - Unsupervised methods in data science with applications to finance and social sciences

Organizer: Antonio Balzanella

Model Selection For Mixture Hidden Markov Models: An Application To Clickstream Data

Authors:

Furio Urso[1*†], Antonino Abbruzzo[1], Marcello Chiodi[1]
Maria Francesca Cracolici[1]

[1] Department of Economics, Business and Statistics, University of Palermo

[*] Corresponding author [†] Presenter

Contact: furio.urso@unipa.it

Keywords:

Model selection, Clusters, Hidden States, Clickstream Data, Entropy-based scores, Information Criteria

Abstract:

Clickstream data, which capture the sequential interactions between users and websites, have become an invaluable asset for businesses aiming to optimise user experience and enhance marketing strategies. These data provide detailed information on user behaviour, including navigation paths and time spent on individual pages. However, they lack explicit information regarding users' underlying intentions - that is, the reasons driving their website browsing - which complicates the task of distinguishing and classifying diverse browsing behaviours. In particular, users exhibiting similar navigation patterns may pursue different, unknown objectives and therefore should be segmented distinctly. In this context, the Mixture Hidden Markov Model (MHMM) represents a powerful statistical tool capable of uncovering latent structures within the data and discerning meaningful differences among superficially similar behaviours. Consequently, it is well suited to address the hidden heterogeneity inherent in apparently analogous browsing patterns. By capturing both the diversity of user behaviours and the underlying dynamics of browsing patterns, MHMMs provide a comprehensive tool for understanding user interactions. However, their application to short categorical sequences has highlighted limitations in the accuracy of traditional model selection criteria, such as AIC and BIC, which often prove inadequate. Our study addresses this issue by introducing a novel model selection criterion grounded in an entropy-based framework (we named it BIC_H). BIC_H innovatively integrates cluster-level and state-level entropy measures into a unified penalisation scheme. By accounting for the two latent structures of MHMMs - subpopulations and hidden states - BIC_H ensures a more accurate and reliable evaluation of candidate models. The criterion minimises BIC_H to identify models with the optimal degree of class separation, addressing a longstanding challenge in MHMM applications. To evaluate its effectiveness, we conducted a Monte Carlo simulation study comparing BIC_H with established criteria such as AIC, BIC, and ssBIC across a range of sample sizes and sequence lengths. The results consistently demonstrated the superiority of BIC_H, particularly in scenarios characterised by short sequences and limited data. While traditional criteria often struggle to correctly identify the number of components and states,

BIC_H achieves better performance, offering a significant methodological advancement. The practical applicability of BIC_H was illustrated through its application to real clickstream data collected from the website of a Sicilian hospitality company. The dataset, representing user navigation across various sections of the site, was analysed using MHMMs guided by BIC_H. The analysis revealed three distinct user profiles: the Casual explorer/Potential partner, the Information seeker, and the Potential tourist. These profiles were differentiated based on navigation patterns, geographical location, device type, and access times. The findings provided actionable insights into user behaviour, highlighting the need for targeted improvements in website design and navigation flow. In conclusion, this study makes three critical contributions. First, it proposes BIC_H, a model selection criterion that enriches the methodological framework for MHMMs by incorporating entropy-based measures. Second, it demonstrates the effectiveness of MHMMs, supported by BIC_H, in extracting meaningful insights from complex behavioural data. Third, it provides a practical case study bridging the gap between theoretical innovation and real-world application.

Cluster-Based Prediction Under Missing Data: An Application To Green Funding Of Italian Smes

Authors:

Gianmarco Borrata[1*†], Antonio Balzanella[2], Raffaele Mattera[2]
Rosanna Verde[2]

[1] Università di Napoli "Federico II
[2] Università della Campania "Luigi Vanvitelli

[*] Corresponding author [†] Presenter

Contact: gianmarco.borrata@unina.it

Keywords:

Clusterwise regression, Missing values, Imputation

Abstract:

In this work, we introduce a novel framework for prediction in the presence of missing data, under the assumption of an underlying cluster structure in the data. A particularly challenging scenario arises when missing values are present in the test set, that is in the units for which predictions must be made. Traditional imputation techniques often aim solely at recovering values for the missing covariates without considering the prediction. Moreover, the traditional methods usually assume a homogeneous predictive relationship across the population, disregarding potential heterogeneity among subgroups or clusters. To address these issues, we propose an approach that integrates clusterwise regression with donor-based imputation. Specifically, we estimate cluster-specific predictive models that capture local relationships between predictors and the response variable. These models are then used to guide donor selection: for each test unit with missing covariates, donors are selected from the training set based on model-based similarity criteria within the same cluster. This ensures that the imputed values are coherent with the prediction objective within each local structure. We assess the performance of the proposed framework through a comprehensive set of experiments on both simulated and real-world datasets. Our results highlight that the use of clusterwise models significantly improves the alignment between imputation and prediction, especially when the importance of covariates varies across clusters. Furthermore, we analyze how the impact of missingness on the prediction varies depending on which covariates are missing, and how the predictive relevance of each covariate can differ across clusters. Finally, we apply our method to a dataset on green fund applications submitted by Italian SMEs, where we aim to predict the amount of funds received. The empirical analysis demonstrates that our approach outperforms standard imputation methods in terms of predictive accuracy, especially in the presence of heterogeneous relationships and structured missing data.

Aggregating Esg Scores: a Wasserstein Distance-Based Method

Authors:

Arianna Agosto[1*†], Paola Cerchiello[1], Antonio Balzanella[2]

[1] University of Pavia
[2] Universit'a degli Studi della Campania 'Luigi Vanvitelli'

[*] Corresponding author [†] Presenter

Contact: arianna.agosto@unipv.it

Keywords:

Wasserstein distance, Optimal transport theory, ESG indicators

Abstract:

This paper presents a novel methodology for aggregating Environmental, Social, and Governance (ESG) scores using optimal transport theory, specifically by means of the Wasserstein distance. The proposed approach addresses the well-known issue of divergence among ESG ratings provided by different agencies, which can hinder consistent investment decision- making and risk assessment. We propose the computation of the Wasserstein barycenter, so to produce a consensus ESG score that mediates the individual ratings. A bootstrap algorithm is then implemented to construct confidence intervals for the aggregated scores. Such intervals serve a dual function: they quantify statistical uncertainty and provide practical tools for stakeholders by accommodating more conservative or lenient perspectives in ESG evaluation. We show the application of our methodology to a sample of European companies for which scores assigned by three different agencies are available. Moreover, we explore how Wasserstein- based aggregation impacts the single ESG dimensions (E, S and G), with an application to SMEs. Our study shows that the Wasserstein-based indicator provides a coherent representation of companies' ESG profile, especially in capturing tail behavior, while offering greater granularity in the central range of scores. This increased sensitivity enhances interpretability and supports more informed investment decisions. The consensus rating does not merely average inputs but reflects the underlying distributional differences across providers, which adds transparency and robustness to ESG evaluations. In conclusion, the Wasserstein distance-based method proposed in this paper provides a statistically rigorous and conceptually transparent framework for ESG score aggregation. It offers a credible supplement to conven- tional ratings and serves as a valuable tool for investors, asset managers, and sustainability analysts seeking a comprehensive and reliable assessment of corporate ESG performance. The approach supports improved decision-making in sustainable finance by capturing both the central tendencies and the uncertainty inherent in ESG assessments.

Session - Advances in robust clustering

Organizers: Luis Angel Garcia Escudero

Robust And Interpretable Matrix-Variate Data Analysis

Authors:

Marcus Mayrhofer[1*†], Una Radojičić[1], Peter Filzmoser[1]

[1] TU Wien

[*] Corresponding author [†] Presenter

Contact: Marcus.Mayrhofer@tuwien.ac.at

Keywords:

Robust statistics, Covariance estimation, Matrix-valued data, Interpretability

Abstract:

In recent years, data complexity has rapidly increased, which often results in samples that are naturally represented by a matrix. Common examples include image data, multivariate functional data, and longitudinal data. Such data points are naturally arranged into matrices, however, they are often transformed into high-dimensional vectors (by stacking the rows or columns), leading to limitations for many multivariate data analysis procedures. We propose to analyze these matrix-variate observations in a framework that preserves the matrix structure, employing the semi-parametric family of matrix-variate elliptical distributions, which is parameterized by a mean matrix, rowwise and columnwise covariance matrix, as well as a density generating function. The mean as well as the row and column covariances are usually estimated using a maximum likelihood approach, which is sensitive to outliers, prompting the need for robust parameter estimation. We introduce the Matrix Minimum Covariance Determinant (MMCD) estimators to robustify estimation for matrix-variate data, which is the first high-breakdown estimator in this setting. We proved various important properties, such as invariance under linear matrix transformations, high breakdown point and efficiency, as well as consistency under matrix-variate elliptical distributions. Robust Mahalanobis distances based on the MMCD estimators enable reliable outlier detection and avoid the masking effect. Shapley values help explain why an observation is outlying by decomposing the squared Mahalanobis distance into contributions of individual rows, columns, or cells of an observation. We can extend the MMCD estimators to a setting with multiple groups by relying on the robust trimmed clustering approach. While the MMCD estimators only distinguish between regular observations and outliers by trimming a proportion of the observations, the trimmed clustering approach assigns each observation to one of k clusters or marks them as outliers.

A Probabilistic Branch-And-Bound Algorithm For Clusterwise Linear Regression

Authors:

Andrea Fois[1], Luca Insolia[2*†], Luca Consolini[1]
Fabrizio Laurini[3], Marco Locatelli[1], Marco Riani[3]

[1] Department of Engineering and Architecture, University of Parma
[2] Faculty of Science, University of Geneva
[3] Department of Economics and Management, University of Parma

[*] Corresponding author [†] Presenter

Contact: Luca.insolia@unige.ch

Keywords:

Gaussian Mixtures, Mixed-Integer Semidefinite Programming, Optimization

Abstract:

Clusterwise linear regression models combine regression with cluster analysis by assuming that the data are generated from a mixture of linear regression models, where the main goal is to minimize the overall loss by fitting a separate regression model within each cluster. These models have attracted a lot of attention in a variety of domains, such as medicine, social sciences, and agriculture, due to their flexibility in modeling heterogeneous populations. However, existing model-based methods often rely on heuristic iterative approaches based on resampling, which tend to suffer as the number of components and/or the dimensionality of the model increases. In this work, under Gaussian assumptions, we consider a new reformulation of the clusterwise linear regression problem that makes it suitable to a branch-and-bound based approach, and propose a probabilistic branch-and-bound algorithm called `pclustreg` which is tailored for this problem. Under very mild conditions, we show that `pclustreg` provides, with high probability, solutions that in terms of log-likelihood are at least as good as the ones obtained by an oracle estimator that relies on information on the true (unknown) cluster assignments. Moreover, by limiting the number of nodes expanded during the branching phase, the proposed `pclustreg` algorithm can also be used as a computationally lean heuristic to find suitable solutions. By exploiting the properties of branch-and-bound methods, even when the algorithm execution is stopped before convergence, the solutions found by `pclustreg` during the search remain competitive with the ones provided by state-of-the-art methods. We illustrate the advantages offered by `pclustreg` through extensive numerical experiments on both synthetic and real-world datasets. These results indicate that when `pclustreg` is used as a heuristic, it typically provides a better trade-off between computational efficiency and model accuracy compared to existing heuristic methods, and it often achieves better solutions while reducing computing times.

Cellwise Outliers In Heterogeneous Populations: a Fuzzy Clustering Approach

Authors:

Giorgia Zaccaria[1*†], Lorenzo Benzakour[1], Francesca Greselin[1]
Luis Angel García-Escudero[2], Agustín Mayo Iscar[2]

[1] University of Milano-Bicocca
[2] University of Valladolid

[*] Corresponding author [†] Presenter

Contact: giorgia.zaccaria@unimib.it

Keywords:

cellwise contamination, fuzzy assignments, constrained optimization, EM algorithm, missing data

Abstract:

Real data often contain outliers, which are values that deviate from the pattern followed by the majority of the data. Outliers typically refer to entire cases or rows of a data matrix (casewise or rowwise outliers). In recent years, a novel paradigm has been introduced to account for contamination in individual cells of a data matrix. These can be handled similarly to the casewise outliers, that is by removing the corresponding information (i.e., rows in the casewise, cells in the cellwise) from parameter estimation. However, when only cells, rather than entire observations, are considered contaminated, the reliable cells within an observation can be retained and used to gather information about the contaminated ones. In this work, we introduce a robust fuzzy clustering approach for the detection of cellwise outliers, which are particularly harmful in heterogeneous populations. Unlike the existing robust fuzzy clustering methodologies, which primarily address casewise contamination, our proposal relaxes the spherical assumption for the clusters while retaining a constraint on the eigenvalue ratio. The latter controls the allowable differences among the scatter matrices and prevents issues such as degeneracies. The proposal is estimated using an Expectation-Maximization algorithm, which includes an additional step for detecting a fixed proportion of outlying cells per variable. In the E-step, the contaminated cells are treated as missing information and are therefore imputed. Consequently, the parameters are estimated on the completed data in the M-step. The performance of the proposed methodology is illustrated through two real data applications, along with guidance for selecting the tuning parameters on which the model depends.

Session - Variable selection in complex settings

Organizer: Silvia Bacci

Variable Selection In Latent Regression Irt Models Via Knockoffs: An Application To International Large-Scale Assess- Ment In Education

Authors:

Zilong Xie[1], Yunxiao Chen[2*†], Matthias von Davier[3]
Haolei Weng[4]

[1] The Chinese University of Hong Kong
[2] London School of Economics and Political Science
[3] Boston College
[4] Michigan State University

[*] Corresponding author [†] Presenter

Contact: Y.Chen186@lse.ac.uk

Keywords:

international large-scale assessment, missing data, knockoffs

Abstract:

International large-scale assessments (ILSAs) play an important role in educational research and policy making. They collect valuable data on education quality and performance development across many education systems, giving countries the opportunity to share techniques, organisational structures, and policies that have proven efficient and successful. To gain insights from ILSA data, we identify non-cognitive variables associated with students' academic performance. This problem has three analytical challenges: (a) academic performance is measured by cognitive items under a matrix sampling design; (b) there are many missing values in the non-cognitive variables; and (c) multiple comparisons due to a large number of non-cognitive variables. We consider an application to the Programme for International Student Assessment, aiming to identify non-cognitive variables associated with students' performance in science. We formulate it as a variable selection problem under a general latent variable model framework and further propose a knockoff method that conducts variable selection with a controlled error rate for false selections.

Variable Selection Via Knockoffs For Clustered Data

Authors:

Silvia Bacci[1], Emanuela Dreassi[1], Leonardo Grilli[1]
Carla Rampichini[2*†]

[1] Deparment of Statistics, Computer Science, Applications 'G.Parenti', University of Florence
[2] Dipartimento di Statistica, informatica, applicazioni 'G. Parenti' Università di Firenze

[*] Corresponding author [†] Presenter

Contact: carla.rampichini@unifi.it

Keywords:

hierarchical data, multilevel models, repeated measures

Abstract:

The selection of relevant predictors affecting a response is a fundamental issue in assessing a statistical model. It is particularly challenging when numerous predictors are available. Indeed, different selection strategies may lead to different results with the risk of including in the model variables with null effects or, on the opposite, excluding variables with a non-null effect. The knockoffs approach has the advantage of controlling for the false discovery rate (FDR, the proportion of variables wrongly declared non-null). We extend the knockoffs method for selecting predictors in the case of clustered data (cross-section or repeated measures). To our knowledge, the literature on this topic is null. In the setting of clustered data, variable selection is more complex since some predictors are measured at the observation level (level 1), whereas other predictors are measured at the cluster level (level 2), so their values are constant within clusters by design. A solution is to carry out variable selection separately at the two levels. To this end, we propose a two-step approach: (i) decompose each level 1 predictor into level 2 and level 1 components by replacing it with %SB: tolto the e messo its its cluster mean and the deviation from the cluster mean; (ii) perform variable selection separately at the two levels, where the level 1 data matrix includes the deviations from the cluster means, and the level 2 data matrix includes the cluster means of level 1 predictors and the level 2 predictors. To evaluate the performance of the proposed method, we perform a simulation study with a continuous response and several continuous predictors at level 2 (clusters) and level 1 (observations). The study shows satisfactory results in terms of FDR and power. All variable selection methods fail when applied to the original data matrix with both level 1 and level 2 predictors. In contrast, all methods perform better when applied to the level 1 and level 2 data matrices separately. Moreover, the sequential knockoffs method performs substantially better than the Lasso. We apply the proposed method to a challenging case study whose data set has a clustered, specifically repeated measures, structure, a setting not accounted for by traditional knockoff methods. The analysis model includes random effects to account for the correlation within repeated measures of the same subject (cluster). Our proposal to implement the knockoffs method in a clustered data framework is feasible, flexible, and effective.

Taming Complexity: Variable Selection In Mixed-Effects Location-Scale And Location-Shift Models For Ordinal Data

Authors:

Moritz Berger[1], Maria Iannario[2*†]

[1] University of Bonn
[2] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: maria.iannario@unina.it

Keywords:

Ordinal Regression, Mixed-Effects Models, Variable Selection, Location-Scale Models, LASSO Regularization

Abstract:

Understanding ordinal responses in the presence of heterogeneity and clustering is a central challenge in modern statistical modeling. Classical cumulative models, such as the proportional odds model, often fail to capture variations in response dispersion across subpopulations or hierarchical groupings. This paper addresses these limitations by advancing model-based solutions that explicitly account for varying dispersion through two alternative approaches: the location-scale model and the location-shift model, each enriched with mixed-effects components. These models offer flexible structures to separately capture between-cluster and within-cluster variability, which is particularly relevant when analyzing multi-level or cross-country data, such as those collected in large-scale surveys. A key methodological contribution of this work lies in its approach to variable selection under complex modeling conditions. Both the location-scale and location-shift models introduce two sets of predictors: one influencing the central tendency (location) of the ordinal response, and the other driving variability (dispersion). The inclusion of multiple covariates in both components can lead to model over-parameterization and interpretation challenges. To overcome this, we propose a penalized likelihood framework that incorporates LASSO-type regularization, including fused penalties for ordinal covariates. This allows the model to automatically identify the most relevant variables for each model component while collapsing redundant or similar levels in ordinal predictors. The methodological framework is evaluated using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), focusing on self-assessed health status across 27 countries. The dataset is divided into training, validation, and test sets to rigorously assess model fit and predictive performance. We demonstrate that models with random intercepts accounting for country-specific effects significantly outperform standard models, and that accounting for dispersion leads to more accurate and interpretable results. Among the competing models, the penalized location-scale model with country-specific random effects emerges as the best-fitting structure. It successfully identifies a subset of predictors with significant influence on both the mean and variability of self-rated health. The penalized estimation framework facilitates these insights by reducing complexity and emphasizing meaningful variation. Overall, the study underscores the importance of using flexible cumulative models to disentangle systematic shifts and dispersion effects in ordinal data, especially in hierarchical settings. The proposed variable selection strategy not only improves model parsimony but also enhances interpretability and robustness. These contributions are relevant for a wide range of applications in the social sciences, public health, and survey method-

ology, where ordinal responses are common and heterogeneity is expected. Future directions include extending the approach to longitudinal designs and exploring alternative penalization techniques to further refine model flexibility.

Session - Sports analytics

Organizer: Michel van de Velden

Rewriting The Rules: Can a Draft System Close The Premier League's Competitive Divide?

Authors:

Benjamin Holmes[1*†]

[1] University of Liverpool

[*] Corresponding author [†] Presenter

Contact: b.holmes@liverpool.ac.uk

Keywords:

Sports analytics, Simulation, Optimisation, Competitive balance

Abstract:

This paper explores whether a draft system, similar to those in North American sports, could improve competitive balance in the Premier League. The league has long been dominated by a small group of financially powerful clubs, while newly promoted teams often struggle to survive. A well-designed draft could help redistribute young talent, increase parity, and potentially become a major commercial event in itself. The proposed draft would be limited to players aged 21 or under who have not played more than 450 minutes of league football for their current club. This prevents manipulation through brief substitute appearances or excessive U21 usage in cups. The draft would be structured over multiple rounds, with selection order based on reverse league standings. Each team would receive five additional draft slots, and to prevent financial imbalance, a league-subsidised wage system would ensure poorer clubs aren't priced out of top prospects. To assess the impact, we repurpose an existing squad optimisation model to simulate team selections, followed by a forecasting model to simulate the league under two scenarios: with and without the draft. Key outcomes include the points gap between top and bottom clubs, the survival rate of promoted teams, and increased minutes for U21 players. All modelling infrastructure is in place; it is now a case of defining the draft pool and running simulations. Previous applications of these models show that optimised squads consistently outperform standard ones, suggesting that a draft mechanism could significantly improve parity. While hypothetical, this study provides a data-driven framework for evaluating policy changes in football. The draft may also enhance youth development by offering increased playing time and pathways at a wider range of clubs. Commercially, a Premier League draft could evolve into a globally significant event, drawing attention similar to major transfer windows or international tournaments. Its success would depend on careful integration with existing systems and avoiding exploitation by wealthier clubs, but the potential benefits to competition, development, and global engagement are substantial.

Unlocking Prescriptive Training: Causal Machine Learning For Actionable Athlete Guidance

Authors:

Thomas Servotte[1], Tom Van Deuren[2*†], Tim Verdonck[1]

[1] Department of Mathematics, University of Antwerp
[2] IDLab, Department of Computer Science, University of Antwerp-imec

[*] Corresponding author [†] Presenter

Contact: tom.vandeuren@uantwerpen.be

Keywords:

sports performance optimization, athlete monitoring, precision training, causal machine learning, prescriptive analytics

Abstract:

Effective training process management in sports like soccer, cycling, and running aims to guide athletes toward desired performance outcomes while maintaining health. While athlete monitoring technologies provide abundant data, translating this into optimal, personalized training actions remains a key challenge. Current data analytics often describe past events or predict future possibilities but frequently fall short of prescribing specific interventions to achieve a targeted goal. To move toward such prescriptive guidance, a deeper understanding of the underlying cause-and-effect relationships between training variables, athlete responses, and outcomes is essential, moving beyond reliance on observed correlations. Causal Machine Learning (CML) provides a robust framework with powerful methods to uncover these crucial causal links and enable such understanding. In this talk, we will discuss how CML can be leveraged to enhance precision training, facilitating a shift from data-informed observation to data-driven prescription. We will illustrate this with examples from professional soccer. Here, our research applied the potential outcomes framework for time-varying treatments to estimate the individualized effects of different training regimens on Rating of Perceived Exertion (RPE). We utilized models like the Counterfactual Recurrent Network (CRN) to address the bias introduced by time-dependent confounders. This enables causal inferences about how training modifications impact player responses, thereby providing more informed ways to adjust training load and achieve desired effort levels. In cycling and running, causal modeling principles are similarly applied to develop performance models and individual training advice. This involves using techniques, such as Directed Acyclic Graph (DAG) learning, to address limitations in models like the Fitness-Fatigue Model (FFM). While the FFM is conceptually useful for its interpretability, studies have highlighted significant flaws; for example, it often struggles with poor identifiability of its fitness and fatigue parameters and can be prone to overfitting, where the fatigue component may not improve predictive accuracy. These issues can limit its reliability for truly personalized training prescription. Our CML approach seeks to address such challenges by not only aiming for more personalized models but also by causally determining how objective markers such as Heart Rate Variability (HRV) and sleep quality truly influence athlete adaptation and fatigue. This allows for these internal response data to be meaningfully integrated with training load, leading to more robust, causally-informed indicators for guiding the training process. The presentation will highlight how this CML-driven approach, utilizing the increasing availability of data from wearables and physiological monitoring, can lead to more precise and adaptive training plans. The ultimate

goal is to equip coaches and athletes with tools that enhance performance and support injury prevention, driven by a CML-fostered understanding of how individual responses to training are shaped by the interplay of various internal and external factors, leading to truly personalized adaptation and readiness strategies.

Disentangling Successful Football Actions: A Network-Based Approach

Authors:

Roberto Rondinelli[1*†], Lucio Palazzo[2], Riccardo Ievoli[3]
Giancarlo Ragozini[1]

[1] University of Naples Federico II
[2] University of Naples L'Orientale
[3] University of Ferrara

[*] Corresponding author [†] Presenter

Contact: roberto.rondinelli@unina.it

Keywords:

Passing networks, Passing path, Offensive Actions, Clustering

Abstract:

Recent advances in football analytics have highlighted the potential of passing networks as a complementary tool to traditional match statistics for evaluating team performance and modelling match outcomes. However, most existing approaches rely on aggregated representations of these networks, typically constructed over entire matches or predefined time intervals (e.g., halves or 15-minute segments). Such aggregations often neglect the temporal and spatial dynamics of individual offensive actions. We propose a novel framework in which passing networks are constructed at the level of individual offensive actions, each defined as a sequence of passes initiated by a team in possession and terminated by a defensive intervention, a shot, or a loss of possession. This granularity allows for a more detailed examination of the relationship between network structure and outcomes of offensive actions. From each action-specific network, we extract a set of topological features such as network density, centrality measures, and clustering coefficients, which can be meaningfully associated with the success of an offensive action. To explore this relationship, we group offensive actions based on their structural characteristics and observed results. This approach enables us to identify distinct profiles of offensive play and to characterise the network patterns most frequently associated with successful attacking behaviour. Our methodology is illustrated through a real-world dataset involving matches from the 2015-2016 season of the Italian Serie A. The framework can be readily extended to other competitions, offering a generalizable tool for tactical analysis and performance evaluation in football and other team sports.

Identifying Playing Styles In Football Through Topic Modelling

Authors:

Michel van de Velden[1*†], Vanja Misuric-Ramljak[2]

[1] Erasmus Universtity Rotterdam
[2] Statistics Netherlands (CBS)

[*] Corresponding author [†] Presenter

Contact: vandevelden@ese.eur.nl

Keywords:

Cluster analysis, Sport analytics, Topic Modelling

Abstract:

In the competitive and constantly evolving world of modern football, data analytics has become increasingly popular over recent years. Football clubs analyze large amounts of data in an attempt to improve their performance and gain a competitive advantage over rivals. Several attempts have been made in formulating, detecting, and measuring team-based indicators. That is, indicators that do not focus on individual players' performances, but that concern appraisal of team performances. One team-based indicator popular with football analysts and managers, is a so-called playing style. Analysts at all levels of the game regularly use the term playing style to better understand the complexity of football matches and team tactics. A formal definition, let alone a proper quantification, of a playing style is typically not provided. In this paper, we introduce a method for quantifying a team's playing style based on match event data. In particular, we define playing styles based on the location and patterns of a team's consecutive actions with the ball. Using our method, a team's playing style is represented by a "style" vector, that summarizes the playing style in a way that is both interpretable and suitable for further data analysis. Characterizing playing style from match event data is challenging as such data are complex and high-dimensional. The match event data involve player locations, their movements, and the actions they perform. We deal with these challenges by first constructing ball movement patterns, which are sequences of coordinated actions involving the ball. Using these ball movement patterns, we identify playing styles by adapting methodology from the field of text analytics. Specifically, we apply Latent Dirichlet Allocation (LDA) to ball movement patterns to obtain distributions over such ball movement patterns with similar structure; that is, the playing styles. We apply our methodology to a publicly available data set, and illustrate how the resulting playing styles can be used in practice.

Session - Statistical modelling of financial data

Organizer: Carmela Iorio

A Data-Driven Fragmented Autocorrelation Approach For Time Series Clustering

Authors:

Jorge Caiado[1*†], Nuno Crato[1]

[1] ISEG/Universidade de Lisboa and CEMAPRE

[*] Corresponding author [†] Presenter

Contact: jcaiado@iseg.ulisboa.pt

Keywords:

Time series clustering, Autocorrelation function (ACF), Partial autocorrelation (PACF), Fragmented methods, Distance metric, Economic indicators

Abstract:

Time series clustering relies heavily on the choice of an appropriate distance metric to capture underlying patterns and structures. While traditional methods often employ full autocorrelation functions (ACF) or periodograms, these can be inefficient when irrelevant lags or frequencies obscure discriminative features. Recent advancements, such as the fragmented periodogram and fragmented autocorrelation methods, address this by focusing on key frequencies or lags. However, these approaches assume prior knowledge of the time series structure, limiting their applicability in real-world scenarios where such information is unavailable. In this paper, we propose a novel data-driven fragmentation procedure that automatically identifies significant autocorrelations and partial autocorrelations (PACF) to enhance time series clustering. By applying a statistical significance threshold, our method filters out noise and retains only the most informative lags, improving clustering accuracy. We evaluate our approach through extensive simulations involving linear time series models and demonstrate its effectiveness on real-world economic and financial datasets, including multi-country indicators such as GDP growth, inflation, military expenditure, and fertility rates. Our results show that the proposed metric outperforms conventional methods in distinguishing between different generating processes, particularly when the underlying dynamics are unknown. This work contributes to the broader literature on feature-based time series clustering by providing a robust, automated framework for selecting discriminative features in the time domain.

Clustering Financial Time Series By Good And Bad Realized Volatility Decomposition

Authors:

Raffaele Mattera[1*†], Germana Scepi[2]

[1] University of Campania "Luigi Vanvitelli
[2] Department of Economics and Statistics, University of Naples "Federico II

[*] Corresponding author [†] Presenter

Contact: raffaele.mattera@unicampania.it

Keywords:

cluster analysis, high-frequency data, weighted hierarchical clustering

Abstract:

Financial time series are often clustered based on conditional volatility, estimated from GARCH models. However, realized measures based on high-frequency data provide a more accurate estimation of the latent volatility process. In this paper, we assess the similarity of realized volatility dynamics using an autoregressive metric and the decomposition of volatility into good and bad components. In particular, we introduce a novel weighted algorithm for improving the hierarchical clustering ap- proach and apply it to the U.S. stocks traded in the Dow Jones Industrial Average (DJIA) index.

Generalized Multivariate Markov Chains

Authors:

Carolin Vasconcelos[1], Bruno Damasio[1*†]

[1] NOVA IMS, Universidade Nova de Lisboa

[*] Corresponding author [†] Presenter

Contact: bdamasio@novaims.unl.pt

Keywords:

Generalized Markov chains, Mixture transition distribution, Stochastic process

Abstract:

Multivariate Markov chains (MMC) have a wide range of applications across various fields. However, the availability of packages for estimating and applying these models is limited. Most existing methods rely on algorithms and software that are either not widely accessible or are only applicable in specific contexts. In addition to the lack of software implementation, previous work on MMC models has primarily focused on improving estimation techniques and/or enhancing model parsimony. This study aims to address both of these gaps. Firstly, we propose a new generalization of the MMC model that incorporates exogenous variables. Specifically, our model includes the effects of past MMC values and those of pre-determined or exogenous covariates by introducing a non-homogeneous Markov chain structure. We also address the problem of statistical inference. Secondly, we developed a novel R package that implements this generalization, along with the MTD model for MMC and the MTD-probit model. To evaluate the model's type I and type II error rate, we conducted a Monte Carlo simulation study. The results demonstrated that our model consistently identified the presence of a non-homogeneous Markov chain. Furthermore, an empirical application illustrated the relevance of the proposed model by estimating the transition probability matrix across different values of the exogenous variables. Ignoring the effect of exogenous variables in MMC means that we would not detect the probabilities' changes according to the covariates' values. In this setting, one would have a limited view of the studied process. This approach not only clarifies the influence of specific variables on a process but also provides a broadly accessible software implementation.

Does Sustainability Impact Tail Risk Measurement? Evidence From a Novel Text-Based Esg Indicator

Authors:

Alessandra Amendola[1*†], Vincenzo Candila[1], Shahram Dehghan Jabarabadi[2]
Peter Winker[3]

[1] University of Salerno
[2] University of Padua
[3] Justus Liebig University Giessen

[*] Corresponding author [†] Presenter

Contact: alamendola@unisa.it

Keywords:

ESG indicator, Textual Analysis, Text Classification, Tail Risk Measurement

Abstract:

Over the last two decades, there has been growing interest in Environmental, Social, and Governance (ESG) topics. In the current framework, while there is no general ESG index that measures overall interest in ESG issues, there are various firm-specific ESG ratings, typically provided by third-party rating agencies. However, rating agencies may suffer from several drawbacks, including non-homogeneous grades for the same companies, reports influenced by the size of the company under investigation, and infrequent reporting. To address these issues, this paper proposes a novel general ESG indicator based on news articles scraped from an online news aggregator that covers a wide range of disciplines and topics, offering a multi-dimensional database aligned with ESG factors. The proposed indicator is entirely self-contained and independent of external factors, taking advantage of the textual analysis of scraped news articles. The new ESG indicator is evaluated through both graphical analysis and in a tail risk context. The graphical approach involves thoroughly examining the index graph alongside relevant ESG-related events. Incorporating the ESG indicator as additional information in the estimation of tail risk measures allows us to explore the extent to which ESG-related metrics influence market risk. This sheds light on the financial relevance of ESG factors in predicting market conditions.

Session - Advances in directional statistics

Organizers: Stefania Fensore and Marco Di Marzio

Rounding Errors In Circular Data

Authors:

Charles Taylor[1*†], Stefania Fensore[2], Marco Di Marzio[2]

[1] University of Leeds
[2] University of Chieti-Pescara

[*] Corresponding author [†] Presenter

Contact: C.C.Taylor@leeds.ac.uk

Keywords:

deconvolution, digit preference, rounding, crime data

Abstract:

Data are often recorded imprecisely. This may be due to instrument error, in which case we can treat the errors as being i.i.d.., and seek to recover the density function of the error-free distribution. Data which include times can be considered as circular. There may be periodic behaviour which arise from the time of day, day of the week, or day of the year, and these may also be nested. Time stamped data will often be rounded – to the nearest minute, nearest five minutes, nearest quarter of an hour etc., and yet – if one considers only the minute past the hour, these should be (approximately) uniformly distributed, even of there is a moderate trend on a larger scale. In this case of recorded digit preference, the errors will no longer have the same distribution. We give examples of data which exhibit rounding characteristics and investigate approaches, including regularized deconvolution, to estimate the underlying density.

Robust Estimation In Multivariate Torus Data

Authors:

Claudio Agostinelli[1*†], Luca Greco[2], Giovanni Saraceno[3]

[1] University of Trento
[2] University Giustino Fortunato, Benevento
[3] University of Padova

[*] Corresponding author [†] Presenter

Contact: claudio.agostinelli@unitn.it

Keywords:

Circular data, Expectation-Maximization algorithm, Influence Function, Outliers, Pearson residual, Ramachandran plot, Wrapped Elliptically Symmetric models

Abstract:

We consider robust estimation of wrapped models to multivariate circular data that are points on the surface of a p-torus based on the weighted likelihood methodology. Robust model fitting is achieved by a set of weighted likelihood estimating equations, based on the computation of data dependent weights aimed to down-weight anomalous values, such as unexpected directions that do not share the main pattern of the bulk of the data. Model fitting is based on a data augmentation approach and achieved according to a suitable modification of the EM or Classification EM algorithm Asymptotic properties and robustness features of the estimators under study have been studied, whereas their finite sample behavior has been investigated by Monte Carlo numerical experiment and real data examples.

Conditional Von Mises Bayesian Networks

Authors:

Anna Gottard[1*†], Agnese Panzera[1]

[1] DiSIA - University of Florence

[*] Corresponding author [†] Presenter

Contact: anna.gottard@unifi.it

Keywords:

Circular data, Conditional independence, Dihedral angles

Abstract:

Directed acyclic graphical models, or Bayesian networks, use a directed acyclic graph to represent the conditional independence relationships between a set of random variables. We introduce a novel class of Bayesian networks specifically designed for circular or angular variables, utilizing the properties of the von Mises distribution. We illustrate our proposal by applying these models to study the conditional independencies within a sequence of angles that characterize the structure of a glycopeptide.

Session - Advanced clustering methods for complex data II

Organizer: Marta Nai Ruscone

A Gaussian Mixture Model Approach For Clustering And Cellwise Outlier Detection

Authors:

Francesca Greselin[1*†], Luis Angel García-Escudero[2], Agustín Mayo Iscar[2]
Giorgia Zaccaria[1]

[1] University of Milano-Bicocca
[2] University of Valladolid

[*] Corresponding author [†] Presenter

Contact: francesca.greselin@unimib.it

Keywords:

Robustness, Model-based clustering, Cellwise contamination, Missing data, EM algorithm, Imputation

Abstract:

Real-world applications may be affected by outlying values. In the model-based clustering literature, several methodologies have been proposed to detect units that deviate from the majority of the data (rowwise outliers) and trim them from the parameter estimates. However, the discarded observations can encompass valuable information in some observed features. Following the more recent cellwise contamination paradigm, we introduce a Gaussian mixture model for cellwise outlier detection. The proposal is estimated via an Expectation-Maximization (EM) algorithm with an additional step for flagging the contaminated cells of a data matrix and then imputing - instead of discarding - them before the parameter estimation. This procedure adheres to the spirit of the EM algorithm by treating the contaminated cells as missing values. We analyze the performance of the proposed model in comparison with other existing methodologies through a simulation study with different scenarios and illustrate its potential use for clustering, outlier detection, and imputation on three real data sets. Additional applications include socio-economic studies, environmental analysis, healthcare, and any domain where the aim is to cluster data affected by missing information and outlying values within features.

Simultaneous Clustering And Reduction Of Curves

Authors:

Roberto Rocci[1*†], Stefano Antonio Gattone[2]

[1] Sapienza University of Rome
[2] University G. d'Annunzio

[*] Corresponding author [†] Presenter

Contact: roberto.rocci@uniroma1.it

Keywords:

Functional data analysis, Unsupervised classification, Dimensional reduction, Penalised maximum likelihood

Abstract:

We propose a new model-based method for performing clustering and dimensional reduction of functional data simultaneously. The approach assumes that the observed functional data are distributed as a finite mixture of Gaussian processes. The differences among the components, in terms of means and covariances, are represented in a functional subspace of reduced dimension. Inference is drawn conditionally on the points at which the curves are evaluated, using a penalised maximum likelihood approach. The penalty term is introduced to account for the functional nature of the data. This enables us to obtain smooth estimates of the centroids. We present an EM-type algorithm for computing these estimates. The calibration of the penalty is data-driven by using cross-validation. The effectiveness of the proposal is demonstrated through applications involving real and simulated data.

On Decision Making In Cluster Analysis With Focus On Variables Of Mixed Type

Authors:

Christian Hennig[1*†]

[1] Dipartimento di Scienze Statistiche "Paolo Fortunati", Universita di Bologna

[*] Corresponding author [†] Presenter

Contact: christian.hennig@unibo.it

Keywords:

cluster analysis, mixed type data, dissimilarity design, local independence, data preprocessing

Abstract:

A systematic framework guiding the decisions required when using cluster analysis in practice will be presented. This includes issues such as connecting the aim of analysis to the choice of appropriate methodology, decisions at the preprocessing stage: standardisation, transformation, dimension reduction, choice of distance measure, decisions regarding outliers and number of clusters, and use of validation tools such as stability assessment, testing, and visualisation. It is important in practice that the researchers are aware of the constructive role of their decisions when clustering data, which goes beyond just finding "true" clusters that supposedly exist independently of such decisions. Actually, various different legitimate clusterings may exist in the same data set, and the researchers' impact is crucial for the clustering result and for understanding its meaning. A special focus will be put on dealing with variables of mixed type, discussing the issue of aggregating different kinds of information, or alternatively using them in complementary ways. A standard example for mixed type data are data comprising continuous and categorical variables, potentially including also ordinal variables. Aggregation can be done via defining a joint distance measure, but this requires a decision regarding how a not numerically meaningful difference between categories is aggregated with the numerical differences on continuous variables. Another approach is model-based clustering, often using local independence as a principle for defining clusters on variables of mixed type. Alternatively, one set of variables can be used for defining constraints on clusters computed on another variable type. One set of variables can also be used for interpretation and validation of a clustering based on another set of variables.

A Novel Multi-View Mixture Model Framework For Longitudinal Clustering With Application To Anca-Associated Vasculitis

Authors:

Shen Jia[1*†], James Ng[1], Mark Little[2]
David Selby[3]

[1] School of Computer Science and Statistics
[2] School of Medicine
[3] German Research Centre for Artificial Intelligence (DFKI)

[*] Corresponding author [†] Presenter

Contact: SHJIA@TCD.IE

Keywords:

Longitudinal Clustering, Irregular time series, Multi-view Clustering

Abstract:

Effectively modeling complex, irregularly sampled longitudinal data is critical for gaining insights into disease progression, identifying patient subgroups, and improving risk prediction in clinical research. In this work, we introduce a novel two-view mixture model that jointly incorporates static baseline variables and dynamic longitudinal biomarker trajectories within a unified probabilistic clustering framework. The temporal component is modeled using Neural Ordinary Differential Equations (Neural ODEs), which allow for continuous-time latent trajectory inference, even with sparse and unevenly spaced measurements. Each temporal cluster is represented by its own Neural ODE, with distinct parameters that capture its unique latent dynamics. Meanwhile, the static baseline features-potentially of mixed data types-are dimensionally reduced and contribute to clustering through a separate probabilistic component. The full model is trained using an Expectation-Maximization (EM) algorithm, with a sparsity-inducing log-penalty to prevent overfitting and promote interpretable, parsimonious subgroup discovery. We validate the proposed framework through simulation studies and apply it to a real-world clinical dataset from an Irish cohort of patients with ANCA-associated vasculitis. The model identifies clinically meaningful subgroups characterized by distinct creatinine trajectories and baseline characteristics. These subgroups exhibit differing relapse risks, demonstrating the clinical relevance of the inferred latent structure and the added value of jointly modeling temporal and static data. Overall, this work presents a scalable and interpretable probabilistic modeling approach for dynamic patient stratification. The framework is particularly well-suited for longitudinal biomedical datasets where the combination of baseline covariates and temporal biomarkers is critical to uncovering disease heterogeneity and informing personalized care strategies.

Session - Nonparametric estimation of latent variable models

Organizer: Matthieu Marbac-Lourdelle

Hierarchical Mixtures Of Latent Trait Analyzers For Clustering Three Way Binary Data

Authors:

Dalila Failli[1*†], Bruno Arpino[2], Maria Francesca Marino[3]

[1] Università degli Studi di Perugia
[2] Università di Padova
[3] Università degli Studi di Firenze

[*] Corresponding author [†] Presenter

Contact: dalilafailli2@gmail.com

Keywords:

Model-based clustering, Finite mixtures, Latent variables, Variational EM algorithm, NPML

Abstract:

Data analysis often involves hierarchical structures, such as students nested within schools or individuals residing in the same countries. In such cases, the data form three-way datasets, where rows represent first-level units, columns represent variables, and layers correspond to second-level units. This structure introduces complex dependencies, and unobserved heterogeneity between layers must be taken into account to ensure accurate and reliable analysis. The Mixture of Latent Trait Analyzers allows to deal with multivariate binary (categorical) items and identify clusters of units with shared unobserved characteristics. This is done via a finite mixture specification; the residual heterogeneity in the way units respond to the different items is captured through a unit-specific, continuous, latent trait in the model. To deal with hierarchical data, we extend the MLTA by modeling the prior mixture component probabilities as functions of both unit-specific covariates and second-level-specific random effects. These latter are intended to capture within-layer dependencies, thus explicitly accounting for the three-way data structure. Furthermore, we propose to leave their distribution unspecified and estimate it from the data via a Non-Parametric Maximum Likelihood approach. This results in the estimation of a discrete distribution that is totally free from strict parametric assumptions. Thanks to the discreteness of such a distribution, we are also able to cluster second-level units. Within each of such clusters (referred to as blocks), the MLTA specification still allows to partition first-level units, thanks to the finite mixture specification. Overall, we obtain a Hierarchical Mixture of Latent Trait Analyzers (HMLTA) that enables: (i) a hierarchical clustering of first- and second-level units, (ii) accounting for the residual dependence within first-level units clusters through a continuous latent trait, and (iii) measuring the effect of covariates on first-level unit clusters. Parameter estimation is performed by extending a double EM algorithm, which is based on a variational approximation of the model log-likelihood function. This method is computationally efficient and straightforward to implement, making it especially useful for handling large datasets. A simulation study is conducted to evaluate parameter estimation and clustering accuracy across different scenarios. The proposed model is applied to data from the European Social Survey (ESS) on digital skills. In detail, we focus on a three-way binary data matrix representing

the mastery of various digital technologies (variables) by residents (first-level units) in different countries (second-level units). The proposal identifies clusters of respondents based on their digital skills, and, on the top of that, blocks of countries based on the baseline attitudes to digital technology of their residents. The model also examines how demographic factors influence individuals' levels of digitalization and accounts for potential residual heterogeneity in their digital skill mastery.

Importance Sampling For Online Variational Learning In State-Space Models

Authors:

Pierre Gloaguen[1*†], Mathis Chagneux[2], Sylvain Le Corff[3]
Jimmy Olsson[4], Mathias Muller[4]

[1] Université Bretagne Sud
[2] Telecom Paris
[3] Sorbonne Université
[4] KTH

[*] Corresponding author [†] Presenter

Contact: pierre.gloaguen@univ-ubs.fr

Keywords:

Variational inference, Importance sampling, State-space models, Stochastic optimization

Abstract:

We address online variational estimation in state-space models. We focus on learning the smoothing distribution, i.e. the joint distribution of the latent states given the observations, by using a variational approach together with importance sampling. The variational distribution are approximated by distributions whose moments are outputs of neural networks. We propose a new algorithm for efficiently computing the gradient of the evidence lower bound (ELBO) gradient in the context of streaming data where observations arrive sequentially. Our contributions also include a computationally efficient online ELBO estimator, demonstrated performance in offline and truly online settings, and adaptability for computing general expectations under joint smoothing distributions.

Non-Parametric Multi-Partitions Clustering

Authors:

Marie du Roy de Chaumaray[1], Vincent Vandewalle[2*†]

[1] Université Rennes 2, IRMAR
[2] Université Côte d'Azur

[*] Corresponding author [†] Presenter

Contact: vincent.vandewalle@univ-cotedazur.fr

Keywords:

model based-clustering, non-parametric models, latent class model

Abstract:

In the framework of model-based clustering, a model, called multi-partitions clustering, allowing several latent class variables has been proposed. This model assumes that the distribution of the observed data can be factorized into several independent blocks of variables, each block following its own mixture model. In this presentation, we assume that each block follows a non-parametric latent class model, *i.e.* independence of the variables in each component of the mixture with no parametric assumption on their class conditional distribution. The purpose is to deduce, from the observation of a sample, the number of blocks, the partition of the variables into the blocks and the number of components in each block, which characterise the proposed model. By following recent literature on model and variable selection in non-parametric mixture models, we propose to discretize the data into bins. This permits to apply the classical multi-partition clustering procedure for parametric multinomials, which are based on a penalized likelihood method (e.g. BIC). The consistency of the procedure is obtained and an efficient optimization is proposed. The performances of the model are investigated on simulated data.

Latent Variable Models For Species Detection: Propagating Uncertainty From Deep Features To Ecological Inference

Authors:

Marie-Pierre Etienne[1*†], Célian Monchy[2], Olivier Gimenez[2]

[1] CREST - ENSAI
[2] CEFE

[*] Corresponding author [†] Presenter

Contact: marie-pierre.etienne@ensai.fr

Keywords:

Deep features, Latent Variable, Uncertainty propagation, Species Distribution Model

Abstract:

Non-invasive sensors, such as camera traps and acoustic recorders, are increasingly used for wildlife monitoring. Given the large volume of data they generate, machine learning algorithms are now widely employed to detect the presence of target species. However, such algorithm-based detections are often treated as equivalent to direct field observations, despite introducing additional sources of uncertainty. Recent approaches have proposed using the classification score of neural networks as a probability, enabling Monte Carlo simulations to capture variability. However, this method critically depends on the quality of the network's calibration. To address this limitation, in the general context of species distribution model, we propose a new framework that leverages automated feature extraction-such as embeddings from the final layers of convolutional neural networks-to characterize images. The presence or absence of the target species is then modeled as a latent variable, explicitly accounting for detection uncertainty. Crucially, the introduction of latent variables enables an integration of machine learning outputs into hierarchical statistical models. This approach bridges state-of-the-art image analysis with established methods in ecological inference, effectively combining the strengths of both fields. We illustrate the potentiel of this approach on simulation data and the study of the presence of Lynx in France.

Session - Bayesian approaches in model-based clustering

Organizer: Gertraud Malsiner-Walli

Dependent Dirichlet-Multinomial Processes With Random Number Of Components

Authors:

Andrea Cremaschi[1*†], Beatrice Franzolini[2]

[1] IE University
[2] Bocconi University

[*] Corresponding author [†] Presenter

Contact: andrea.cremaschi@ie.edu

Keywords:

Grouped data, Dependent processes, Clustering, Mixture of finite mixtures, Partial exchangeability

Abstract:

Over the past two decades, Bayesian nonparametrics has expanded to include flexible dependent prior distributions for mixture models, extending beyond univariate species sampling processes to effectively capture dependencies in grouped data under partial exchangeability. While most research has focused on nonparametric priors with almost surely infinite support points, much less attention has been given to almost surely finite-dimensional dependent mixture models under partial exchangeability, despite their strong theoretical properties and promising performance in the exchangeable case. In this work, we explore alternative strategies for defining a multivariate extension of the finite Dirichlet-Multinomial process and its counterpart incorporating a prior on the number of components. Specifically, we introduce a class of flexible dependent Dirichlet-Multinomial processes based on Generalised Wishart unnormalised weights. We analyse their theoretical properties and demonstrate that, unlike existing alternatives, the proposed prior can achieve any desired level of dependence for any fixed number of components. Additionally, our approach allows for efficient posterior computation without the need for costly variable augmentation schemes. We further demonstrate the practical advantages of our model through extensive simulation studies and the analysis of two real datasets: a benchmark dataset and a case study on sex-specific gene expression differences in the human brain, highlighting its flexibility and computational efficiency in capturing complex dependence structures.

Clips - Finding Cluster Distributions Behind Data

Authors:

Gertraud Malsiner-Walli[1], Sylvia Frühwirth-Schnatter[1], Bettina Grün[1*†]

[1] Vienna University of Economics and Business

[*] Corresponding author [†] Presenter

Contact: Bettina.Gruen@wu.ac.at

Keywords:

Bayesian analysis, label switching, mixture of finite mixtures, model-based clustering, telescoping sampler

Abstract:

When using finite mixture models in model-based clustering, one usually assumes that each mixture component represents a different cluster distribution. Based on these cluster distributions, the latent groups in the data generation process are characterized and differentiated. In addition, the mixture model allows to assign (new) observations to the groups and thus perform clustering on the available data, but also for new observations. A Bayesian analysis allows to estimate the number of clusters by fitting a mixture of finite mixtures using priors which induce a sparse solution regarding the number of filled components. We suggest the post-processing procedure CliPS (Clustering in the Parameter Space) when fitting Bayesian mixture models in the context of model-based clustering to (1) assess the quality of the clustering solution and (2) identify the cluster distributions. The procedure relies on the point process representation of a mixture model. Assuming that a suitable cluster solution requires the clusters to be differentiable with respect to a low-dimensional functional of the component-specific parameters of the mixture, the component-specific MCMC draws are mapped to and clustered in this space, exploiting that, while data distributions usually overlap, the posterior of these functionals should be more and more separated when sample size increases. We outline the procedure, considering in particular the case when fitting a mixture of finite mixtures with the telescoping sampler focusing on the filled components, and illustrate the use of the procedure on multivariate data when fitting a Bayesian mixture of finite mixture model with suitable priors for clustering.

Tree-Structured Mixtures For Spatial Prior Specification

Authors:

Clara Grazian[1*†], Gianluca Mastrantonio[2]

[1] University of Sydney
[2] Politecnico di Torino

[*] Corresponding author [†] Presenter

Contact: clara.grazian@sydney.edu.au

Keywords:

spanning trees, mixture models, Bayesian clustering

Abstract:

In spatial statistical modeling, the specification of prior distributions is crucial for capturing spatial dependencies and supporting robust inference, especially in data-sparse regions. We propose a novel class of spatial priors based on distance-driven mixture models with a spanning tree structure. This approach integrates spatial distances directly into the model hierarchy via a tree-constrained clustering framework, enabling flexible representation of spatial heterogeneity without requiring strong parametric assumptions. The resulting tree-structured mixtures naturally group spatial locations by response similarity, yielding adaptive, interpretable priors.

Session - Differential privacy and robust classification

Organizers: Claudio Agostinelli and Anand Vidyashankar

Randomized Smart Subset Selection

Authors:

Nicholas Rios[1*†], Zhaoxue Tong[2]

[1] George Mason University
[2] Florida State University

[*] Corresponding author [†] Presenter

Contact: nrios4@gmu.edu

Keywords:

Subset Selection, Regularization, Subsampling, Generalized Linear Model, Binary Classification

Abstract:

In many applications, the number of available covariates far exceeds the number of available data points. This makes fitting a Generalized Linear Model (GLM) for binary classification infeasible via traditional methods without reducing the number of model covariates. This reduction can be done through regularization or subset selection, but this is a challenging problem, as the number of possible sets of active covariates is quite large. Many methods have been proposed for subset selection, but existing methods may have difficulty at screening for important effects when the number of active effects is very small. Furthermore, many existing subset selection or regularization methods have a large number of false positive results. We propose a novel Randomized Smart Subset Selection (RS3) algorithm which assigns importance weights to each predictor based on a randomized blocking scheme and then efficiently screens for active covariates. RS3 is then augmented with False Positive Control (FPC) based on subsampling. The RS3 method is flexible, and can be adapted to a wide array of generalized linear models. Empirical results and examples demonstrate the effectiveness of RS3 at finding active predictors and fitting models with high accuracy in binary classification problems.

Adaptive Estimation Under Differential Privacy Constraints

Authors:

Lasse Vuursteen[1*†]

[1] Duke University

[*] Corresponding author [†] Presenter

Contact: lasse.vuursteen@duke.edu

Keywords:

adaptation, privacy, minimax

Abstract:

Estimation guarantees in nonparametric models typically depend on underlying function classes (or hyperparameters) that are seldom known in practice. Adaptive estimators provide simultaneous near-optimal performance across multiple such function classes. In this talk, I will discuss recent work with co-authors Tony Cai and Abhinav Chakraborty, in which we study adaptation under differential privacy constraints. Differential privacy fundamentally limits the information that can be revealed about each individual datum by each data holder. We develop a general theory for adaptation under differential privacy in the context of estimating linear functionals of a density. Our framework characterizes the difficulty of private adaptation problems through a specific "between-class modulus of continuity" that exactly describes the optimal achievable performance for private estimators that must adapt across two or more function classes. Our theory reveals and quantifies the extent to which adaptation between specific function classes suffers as a consequence of imposing differential privacy constraints.

Privacy-Aware Neymanpearson Classification Via DiverGences

Authors:

Pramita Bagchi[1*†], Anand Vidyashankar[2], Fengnan Deng[2]

[1] George Washington University
[2] George Mason University

[*] Corresponding author [†] Presenter

Contact: pbagchi@gmu.edu

Keywords:

Classification, Differential privacy, Hellinger distance

Abstract:

The Neyman-Pearson approach addresses classification with asymmetric errors by formulating it as a constrained hypothesis testing problem. In domains such as medicine and finance, data sharing is severely restricted due to privacy concerns, limiting the direct implementation of traditional methods. While several ad hoc solutions, including synthetic data generation, have been proposed to address privacy restrictions, these typically lack rigorous theoretical guarantees. To overcome this limitation, we propose a principled divergence-based Neyman-Pearson classification framework integrated with recently developed divergence-based differential privacy constraints (e.g., Hellinger DP). We derive oracle inequalities that explicitly characterize the excess risk in terms of the divergence measure and the privacy cost, thus providing theoretical assurances on how privacy affects classification performance.

Session - Advances in statistical learning and modeling

Organizer: Roberta Siciliano

Integrated Quadratic Distance As An Adaptation Criterion For Adaptive Importance Sampling

Authors:

Alessandro Santonicola[1*†]

[1] AICOR - Institute for Artificial Intelligence - University of Bremen

[*] Corresponding author [†] Presenter

Contact: ale_san@uni-bremen.de

Keywords:

Bayesian Inference, Adaptive Importance Sampling, Divergence Functions, Proper Divergence Functions, Integrated Quadratic Distance

Abstract:

Bayesian Inference provides us with a solid framework for estimating uncertainty that is all-pervasive in many different scientific fields. During the years, along with the improving performances attained by computers, computational methods have imposed themselves as the most efficient way of employing Bayesian Inference. We concentrate our efforts on Adaptive Importance Sampling (AIS) and, in particular, the aim of this work is to design a novel adaptation criterion that enjoys certain statistical properties. Hence, k-proper divergence functions are introduced and we recognize that two of the most popular adaptation criteria for AIS, i.e. forward and backwards Kullback-Leibler divergence satisfy the proper condition. However, those criteria can not be fully trusted as they do not tell us whether all the modes of the posterior distribution have been explored or not, as different local optima may produce the same results. In order to provide an alternative to the golden standard we hereby use the integrated quadratic distance, a k-proper divergence function, which is also a metric, as an adaptation criterion for the AIS framework. The main contribution of this work lies in the re-definition of the integrated quadratic distance for the AIS schemes. Once defined, this new criterion is tested by using Gaussian mixtures as proposal functions according to two different procedures: AIS and AMIS (Adaptive Multiple Importance Sampling), against the other two proposed divergence functions. We show that the results produced by using the new criterion are comparable in terms of estimated evidence to the golden standard. Thus, the premises put in motion by this work are encouraging and show that a deeper investigation on the usage of the integrated quadratic distance for AIS schemes could potentially lead to new and interesting results.

Enhance Physics-Informed Neural Networks Performance For Solving Richards Equation By Deep Learning Optimization

Authors:

Salvatore Cuomo[1*†], Francesco Piccialli[1], Roberta Siciliano[1]
Alessandro Bottino[1]

[1] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: salvatore.cuomo@unina.it

Keywords:

Scientific Machine Learning, Deep Learning, Optimization

Abstract:

Scientific Machine Learning (SciML) has emerged as a transformative paradigm, combining data-driven methods with domain-specific knowledge to tackle complex physical problems. This work focuses on optimizing Physics-Informed Neural Networks (PINNs) to solve the Richards equation, a nonlinear partial differential equation governing unsaturated flow in porous media:

$$\frac{\partial \theta(h)}{\partial t} = \nabla \cdot [K(h)\nabla h] + S(h),$$

where $\theta(h)$ is the volumetric water content, $K(h)$ the hydraulic conductivity, h the pressure head, and $S(h)$ a source/sink term. The proposed framework introduces advanced deep learning optimization techniques to enhance the training efficiency, accuracy, and generalization performance of PINNs. Additionally, explainable SciML models are developed to integrate hydrological domain knowledge, ensuring both interpretability and computational reliability. This synergistic approach bridges physical modeling and deep learning, delivering robust, optimized, and interpretable solvers for subsurface flow applications.

Human-Guided Learning For Interpretable Detection Of Online Misogyny

Authors:

Emiliano del Gobbo[1*†], Alex Cucco[2], Lara Fontanella[2]
Elisa Ignazzi[3], Sara Fontanella[4]

[1] Department of Economics, Management and Territory, University of Foggia
[2] Department of Socio-Economic, Managerial, and Statistical Studies, G. d'Annunzio University
[3] Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University
[4] National Heart and Lung Institute, Imperial College London

[*] Corresponding author [†] Presenter

Contact: emiliano.delgobbo@unifg.it

Keywords:

explanation-guided learning, machine learning, misogyny

Abstract:

Women are disproportionately subjected to online harassment, with sexualized hostility representing a prevalent form of gender-based violence. These behaviors are fundamentally entrenched in misogyny, defined as a cultural inclination that sustains animosity or disdain against women exclusively due to their gender. The efficient identification of misogynistic content is essential to reduce harm, protect users' rights, and avert the mainstream of discriminatory language in digital spaces. This work examines the incorporation of human expertise into machine learning systems to enhance the accuracy, interpretability, and ethical alignment of misogyny detection algorithms. We employ an explanation-guided learning (EGL) methodology that utilizes human-annotated rationales to inform model training and improve transparency. Our dataset consists of 13,500 user-generated comments sourced from Twitter, Instagram, and Facebook. Each comment was separately assessed by mixed-gender pairs of annotators during three evaluation rounds. Instead of awarding solely binary labels, annotators pinpointed individual text spans that contributed to the sexist classification, thus offering contextual and semantically nuanced input. Findings demonstrate that models utilizing EGL and span-level human annotations attain superior classification accuracy relative to conventional approaches. The integration of human-supplied explanations allows the algorithm to produce interpretable outputs, elucidating the rationale behind each prediction. This feature is crucial for cultivating trust, responsibility, and adherence to ethical standards in automated content moderation. In conclusion, our findings highlight the essential need of human-in-the-loop solutions in the advancement of socially responsible AI. By integrating human judgment and contextual awareness into the training process, EGL boosts technical performance while aligning model behavior with human values. This research promotes a transition to inclusive, explanation-aware systems in combating online sexism, highlighting the significance of human-centered design in AI-based moderation tools.

Statistical Learning For Large-Scale Few-Shot Classification

Authors:

Semhar Michael[1*†], Andrew Simpson[1], Yana Melnykov[2]

[1] South Dakota State University
[2] The University of Alabama

[*] Corresponding author [†] Presenter

Contact: Semhar.Michael@sdstate.edu

Keywords:

Large-scale classification, Few-shot learning , Covariance clustering, Discriminant analysis

Abstract:

In large-scale few-shot classification problems, the data is characterized by a large number of classes with only few observations per class, posing significant challenges for classical statistical and machine learning approaches. Given the nature of the few-shot learning framework, classical statistical methods make strong assumptions about the data-generating process. A standard solution in such settings is to assume a shared covariance matrix across classes, as in linear discriminant analysis (LDA), to ensure stable and non-singular estimates of the covariance matrix. However, as the number of classes increases, this assumption becomes overly restrictive. Assuming a class-specific covariance matrix gives unstable and singular estimates. We propose a finite mixture model-based approach that relaxes the assumption of LDA and provides a stable estimate of the covariance matrix, enabling a robust classification in few-shot settings. This talk will focus on the development of a few-shot classification model for both singular and non-singular covariance cases, as well as establishing a general framework. In addition, we will discuss extensions to data with skewness. Through simulation studies, we demonstrate the effectiveness of the proposed method. We also illustrate its utility in a real-world application involving forensic source classification, where the number of potential sources is large and the observations per source are sparse.

Session - Analysis of multiblock data

Organizer: Katrijn Van Deun

Exploring Common And Specific Observation-Structures For Several Blocks Of Variables

Authors:

Stéphanie Bougeard[1*†], Jean Michel Galharret[2], Mohamed Hanafi[2]

[1] Anses
[2] Oniris

[*] Corresponding author [†] Presenter

Contact: stephanie.bougeard@anses.fr

Keywords:

Multiblock exploratory method, multiblock method, clustering, dimension reduction

Abstract:

In the age of big data and data fusion, scientists - biologists in particular - are looking for statistical methods that will enable them to study complex links between variables - from different sources and measured on the same observations - in reduced dimension spaces. For this purpose, multiblock methods are relevant tools. The context is one of high-dimensional exploration with unsupervised component-based multiblock methods. To go further in the exploration and interpretation of data, some multiblock methods also make it possible to quantify common (C), partially common (PC), specific (S) and residual (R) information shares of each block. These original methods have been applied, for instance, in the fields of chemometrics, image analysis or 'multi-omics'. Multiblock methods devoted to quantify C/PC/S/R information shares of each block can be classified into two groups. (i) The first one is based on standard methods, which characterize their components according to their C/PC/S/R status. These include sparse-SCA and DISCO-SCA methods. The assignment of such statuses is also possible on the basis of canonical factorization applied to CPCA/SCA, MCOA, MFA, STATIS or CCSWA methods. (ii) The second group of methods - generally associated with an additive criterion - searches for C/PC/S/R status components within the blocks of variables. These methods include PO-PLS, OnPLS, JIVE, COBE, SLIDE and D-GCCA. All these methods have never been compared, although some partial comparisons exist. Moreover, these comparisons are not associated with simulated structures that can be interpreted from observations. Our first aim is to compare the performance of the above-mentioned multiblock methods on simulated data with controlled observation structures and on clear performance indices. Our second aim is to interpret these shares with respect to blocks, observation-structure (e.g., in clusters) and variables. Indeed, each block can contain both C/PC/S/R information, each of which is associated with different structures of the same observations (e.g., a common structure in three clusters, superimposed on a partially common structure in two clusters), associated with different variables in each block (e.g., some variables in a block contribute to the common share, others to the specific share).

Alternative Definitions Of Effects/Contributions In Path Analysis With Multidimensional Blocks

Authors:

Tormod Næs[1*†], Rosaria Romano[2], Oliver Tomic[3]
Age Smilde[4], Kristian Hovde Liland[3]

[1] Nofima AS, Norway; Dept. of Food Science, Faculty of Sciences, University of Copenhagen, Fredriksberg, Denmark
[2] Department of Economics and Statistics, University of Naples Federico II, Italy
[3] Faculty of Science and Technology, Norwegian University of Life Sciences, Norway
[4] Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands

[*] Corresponding author [†] Presenter

Contact: tormod.naes@nofima.no

Keywords:

path analysis, direct/indirect effects, multiblock regression

Abstract:

Path analysis is a statistical technique used to describe the directed dependencies among a set of variables. It is an extension of multiple regression and is often used to test hypotheses about the relationships between variables. In path analysis, a node represents a variable, and the paths between nodes represent the hypothesized causal relationships. Each path has an associated coefficient that quantifies the strength and direction of the relationship. When path analysis involves blocks of variables, it corresponds to structural equation modeling, which allows for the analysis of complex relationships among observed and latent variables. In univariate path analysis, the effects of one node on another are well-defined. However, when nodes are multivariate, meaning they consist of multiple variables or dimensions, the problem becomes more complex and less intuitive. Previous attempts to address this issue have not provided a clear and comprehensible solution. This study introduces a novel definition of effects (called contributions to avoid confusion), grounded in simple orthogonalization techniques and extended to flexible regression models. The aim is to establish an intuitive framework for defining path effects in multidimensional blocks, which remains applicable in unidimensional scenarios. This approach seeks to enhance the clarity and utility of path analysis in more complex, multivariate settings. The methodology involves three distinct regression models, incorporating an input block, an output block, and multiple blocks influencing the output block. This leads to the formulation of several key contributions: total contribution, unique contribution, interaction, and additional contribution. Each of these contributions is carefully defined to capture the intricate relationships between the blocks. The study presents findings from both simulations and real-world data, highlighting the practical implications of the proposed definitions. Special attention is given to cases where rank deficiency poses challenges, demonstrating how the new framework can address these issues effectively. A comprehensive definition of path effects/contributions for multidimensional blocks is provided, along with an analysis of its strengths and weaknesses. This new approach offers a transparent and intuitive solution to a previously complex problem, paving the way for more effective path analysis in multivariate contexts.

Stacked Domain Learning: A Theory-Guided Approach To Multidomain Data Modeling

Authors:

Zino Brystowski[1*†]

[1] Leiden University

[*] Corresponding author [†] Presenter

Contact: brystowskizd@vuw.leidenuniv.nl

Keywords:

Stacking, Multidisciplinary, Theory-driven Modeling, Predictive modeling

Abstract:

Stacked Domain Learning (SDL) is a modeling framework developed to address the complexity of multidisciplinary data by integrating and evaluating theory-based models from different disciplines. Coherent sets of predictors, representing discipline-specific theoretical models, are defined as domains. In a modeling procedure based on the predictive modeling technique Stacking, base models are first trained separately on each domain to predict a specific outcome of interest. These models generate cross-validated (i.e., out-of-sample) predictions, which are then used as inputs to a meta-model that integrates information across domains. The use of lasso penalization in the meta-model further reduces overfitting and performs feature selection on the domain-level predictions, resulting in effective domain selection. This penalized meta-model enhances interpretability through parsimony and highlights the domains that contribute most strongly to outcome prediction. In this way, SDL supports the testing, comparison, and integration of discipline-specific theories, enabling the development of more comprehensive interdisciplinary models to guide theory development. The introduction of the framework is accompanied by results from a small simulation study and an empirical example.

Session - Analysis of teaching and research activity of higher education institutions

Organizers: Paweł Lula and Miladin Stefanovic

Future-Oriented Insights On The Role Of Ai In Higher Education In The Light Of Scientific Publications.

Authors:

Ildikó Dén-Nagy[1], Paweł Lula[2], Anna Drabina[2]
Katarzyna Wójcik[2*†], Norbert Magyar[1]

[1] Budapest University of Economics and Business
[2] Krakow University of Economics

[*] Corresponding author [†] Presenter

Contact: wojcikk@uek.krakow.pl

Keywords:

bibliometrics, scientometrics, Latent Dirichlet Allocation method, BERTopic, futures studies

Abstract:

Our presentation aims to develop and propose a methodology for extracting insights regarding artificial intelligence (AI) and the future of higher education from scientific papers published on the field of futures studies and to use proposed methods and tools for analysis of abstracts of research papers to identify and classify main views and trends. First, the presentation outlines a systematic approach that enables researchers to identify relevant literature in a replicable and systematic manner, than various quantitative methods will be used to extract themes and findings related to the selected topic. The authors will present a framework for the identification of relevant publications, including a discussion of suitable academic databases such as Scopus and OpenAlex, along with criteria for inclusion. Then, the presentation will focus on techniques to extract and analyze key topics and insights related to AI usage in higher education. This includes the application of machine learning-based topic modeling methods (such as LDA, BERTopic, and classification models) to detect trends, thematic clusters, and emerging concerns across publications. Identified opinions will be aggregated due to time and characteristics of the author team. The study is designed to serve a dual purpose. On the one hand, it aims to provide a practical and replicable analytical solution for methodology experts who seek to examine emerging scientific topics through a structured, data-driven approach. On the other hand, the framework is also intended to support the work of futures researchers by offering them insights into how a particular topic is represented in academic discourse. The study allows foresight practitioners to explore the scope, direction, and intensity of scholarly discussions, which can contribute meaningfully to scenario development or designing research projects. Ultimately, the methodology seeks to enhance the foresight capacity of educational and research institutions while promoting international comparative analyses.

Success Factors For Obtaining Horizon Europe Grants By European Higher Education Institutions

Authors:

Miladin Stefanovic[1], Anna Drabina[2*†], Katarzyna Wójcik[2]
Paweł Lula[2]

[1] University of Kragujevac
[2] Krakow University of Economics

[*] Corresponding author [†] Presenter

Contact: drabinaa@uek.krakow.pl

Keywords:

Horizon Europe Programme, success factors of research grants, international projects, international consortium, graph and text analysis

Abstract:

The Horizon Europe program is the next framework program created by the European Commission to support European Union countries (but not only) in the development of innovation and research, to make Europe more competitive in comparison to the rest of the world. In the current perspective for the years 2021-2027, an amount of euro95.5 billion has been allocated for this purpose. HORIZON EUROPE projects belong to the main research undertakings in Europe. They cover all fields of science, allow for cooperation between different types of institutions and enhance international cooperation and knowledge transfer. The presentation will be focused on the analysis of: Identification of the main project topic, Consortium composition, Distribution of fundings, Identification of the main success factors for obtaining the HE project. The analysis of projects' topics will be performed with the use of the Latent Dirichlet Allocation method or the BERTopic framework. Also, the information about keywords (defined by authors and automatically identified by transformer models) will be studied. The authors are also going to present the distribution of topics over time and over countries and institutions. For consortium composition analysis, the number and the type of partners and the degree of internationalization will be taken into account. With the use of the graph clustering method, main groups of consortia will be identified (using embedding-based approach). The next stage of analysis will include the distribution of fundings over time, project topics and consortium structure. For performing the above analysis, the Cordis database will be used. All programs used for analysis will be prepared in Python and R language. The authors are convinced that the results of the project will help in identification of the main success factors for obtaining international research grants.

Sentiment Analysis Of Study Programmes' Evaluation Reports Prepared By Polish Accreditation Committee

Authors:

Anna Drabina[1], Paweł Konkol[1*†], Katarzyna Wójcik[1]

[1] Krakow University of Economics

[*] Corresponding author [†] Presenter

Contact: konkolp@uek.krakow.pl

Keywords:

sentiment analysis, Natural Language Processing, topic modelling, study programme evaluation

Abstract:

The Polish Accreditation Committee is an independent institution established to enhance the quality of Polish higher education, especially its compliance with quality standards reflecting the European and global best practices. The aim of this organisation is to ensure a high position of Polish higher education graduates on the labour market and to increase the competitiveness of Polish higher education institutions. Once a year the Polish Accreditation Committee publishes a list of actual fields of study that are going to be evaluated in the nearest future. The assessment consists of examination of all formal documentation constituting particular field of study as well as visitation in the institution. The result of the evaluation is the descriptive report summarized by overall grade that could be positive, conditional or negative. The aim of our research is sentiment analysis of study programmes' evaluation reports prepared by Polish Accreditation Committee. All their reports are publicly available and could be obtained and examined. The reports are written in Polish. From the perspective of our work, the descriptions of meeting detailed criteria for program evaluation and education quality standards are particularly interesting. We are going to analyse them using Natural Language Processing techniques focusing on sentiment analysis. The correlation between report sentiment and its overall assessment is to be examined. Additionally, topic modelling analysis will be performed to observe the most relevant issues arising in reports. For the purpose of research, scripts in Python language will be used to obtain Polish Accreditation Committee reports, extract relevant texts from reports and perform sentiment analysis and topic modelling. For sentiment analysis VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool, will be used. For topic modelling Latent Dirichlet Allocation (LDA) method and BERTopic technique will be applied.

Selected Counting Processes As a Tool For Modeling The Dynamics Of Scientific Paper Citations

Authors:

Przemyslaw Jasko[1], Paweł Lula[1*]

[1] Krakow University of Economics

[*] Corresponding author [†] Presenter

Contact: pawel.lula@uek.krakow.pl

Keywords:

counting processes, inhomogeneous mixed Poisson process, inhomogeneous negative binomial process, paper citations dynamics, count variable regression, lasso regression estimator

Abstract:

Counting processes such as inhomogeneous Poisson process and its more flexible extension Cox process (or doubly stochastic Poisson process), of which examples are inhomogeneous mixed Poisson process and inhomogeneous negative binomial process, can be used to model scientific papers citations dynamics. Inhomogeneous Poisson process is determined by the parameter called intensity (or instantaneous rate) denoted $\lambda(t)$. In inhomogeneous Poisson process intensity parameter is a function of time, which can also be dependent on covariates x_{it}: $\lambda(t; x_{it}) = \beta * g(t) * h(x_{it})$, where in the considered case of citations dynamics modelling, time t is an age of the ith paper (and is equal to a time passed from the publication moment of the paper, to the considered moment), β is a base rate, $g(t)$ is a decay function (parametric, e.g. exponential decay $g(t) = e^{-\delta t}$ or power-law decay $g(t) = t^{-\alpha}$; or nonparametric), and h(.) captures covariates affecting citation likelihood, related to paper-specific features (e.g., number of references, journal impact factor, author influence metrics, topical information). Symbol x_{it} above represents a covariate vector, with values recorded when the ith publication age was t. A mixed Poisson process is a stochastic process that generalizes the standard Poisson process by introducing randomness into its intensity parameter. It can be viewed as a mixture of Poisson processes with different intensities, where the intensity λ(t) is a random variable (in our case with gamma distribution) introducing randomness into its rate parameter. In turn, a negative binomial process, extends the inhomogeneous mixed Poisson process by replacing the gamma mixing distribution with a gamma process over time. Thus, for negative binomial process, so called cumulated intensity $\Lambda(t) = \int_0^t \lambda(s)$ ds, follows a gamma distribution. We provide the distribution of variable $N(t_{i,final})$, which represents the number of citations observed up to the last moment of observation in the conducted analysis, when the ith publication is of age $t_{i,final}$. This distribution implicated by the considered counting processes, which are inhomogeneous (intensity depending on the time and some weakly exogenous covariates). We show that for negative binomial process, distribution of random variable $N(t_{i,final})$, (conditional on the covariates values observed over time, from the moment of publication until ith paper reaches age of $t_{i,final}$ in the last moment of observation), is negative binomial with parameters which can be derived from the underlying process definition. Thus, and taking in mind that we have citations data (taken from Scopus for papers published by Polish Authors in years 2000-2024) representing only the state in the final moment of observation, we built negative binomial regression model, with the set of possible covariates such as authors scientometric indices, publication topic structure (derived from Latent Dirichlet Allocation), publication type, paper language, authors number, funding type, etc.

To specify the regressors for the negative binomial regression model, we use lasso regression estimation, for which the loss function hyperparameter (multiplying L_1 norm of parameters vector) value, was selected to minimize negative binomial GLM deviance, for validation sets considered in V-fold crossvalidation procedure.

Solecited Sessions

Session - Multidimensional scaling and related methods

Multidimensional Scaling Utilizing Self-Similarity

Authors:

Akinori Okada[1*†], Tadashi Imaizumi[2]

[1] Rikkyo University
[2] Tama University

[*] Corresponding author [†] Presenter

Contact: okada@tama.ac.jp

Keywords:

diagonal element, multidimensional scaling, self-similarity, visualization

Abstract:

A procedure of multidimensional scaling which utilizes the diagonal elements or self-similarities of a similarity matrix in deriving a configuration representing similarity relationships between objects is introduced. Most of the multidimensional scaling procedures based on the distance model do not utilize the diagonal elements, although those based on the vector model usually do. The vector model uses the same style in dealing with both the diagonal elements and the off-diagonal elements (for example, inner product). However, the diagonal elements seem to represent different aspects of relationships between objects and can have different sorts of information from the one the off-diagonal elements have. It is desirable to deal with the diagonal elements differently from the off-diagonal elements. The multidimensional scaling based on the distance model usually fits the distance between two points in an object configuration to the similarity between the two corresponding objects. It is reasonable to think that larger diagonal element of a brand switching matrix tells that the corresponding brand is less similar to the other brand. The present multidimensional scaling fits the sum of the distances between the points representing the object corresponding to a diagonal element and the points representing other objects to the diagonal element. The overall goodness of fit of the obtained configuration to the diagonal elements and the similarities between two objects is defined by the sum of two terms: one is the goodness of fit for the diagonal elements and the other is the goodness of fit for the off-diagonal elements. An algorithm to maximize the overall goodness is developed. The procedure is applied to several sets of data.

Tracking Preference Evolution With Dynamic Multidimensional Unfolding

Authors:

Giuseppe Gismondi[1*†], Marco Cardillo[2], Alfonso Piscitelli[2]

[1] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy
[2] Department of Agricultural Science, University of Naples Federico II, Naples, Italy

[*] Corresponding author [†] Presenter

Contact: giuseppe.gismondi@unina.it

Keywords:

Unfolding, Dynamic MDS, Longitudinal preferences

Abstract:

This paper introduces a novel approach to dynamic multidimensional scaling (MDS) by incorporating the unfolding paradigm for the analysis and visualization of longitudinal preference data. Traditional MDS methods often struggle with temporal datasets, particularly when within-set dissimilarities are either unavailable or not well-defined, which can lead to degenerate or uninterpretable solutions. To address these limitations, we propose an innovative methodology that extends the classical unfolding framework into a dynamic context. By employing data augmentation strategies based on Kemeny-equivalent dissimilarities, our approach reconstructs the necessary dissimilarity matrices, enabling the application of non-metric MDS techniques to preference rankings observed over multiple time points. A central feature of our method is the assumption of temporal stability within one set (typically the items), which aids in disentangling changes in judges' preferences from shifts in the item space, enhancing interpretability. This assumption aligns with practical experimental setups, such as those in sensory evaluation and training contexts, where item characteristics remain fixed. Our approach generates a unified spatial configuration that highlights temporal changes in preferences at both individual and aggregate levels. We validate the methodology through applications on simulated datasets with controlled temporal dynamics and two real-world case studies: a sensory evaluation of olive oils before and after a training seminar, and interregional hospital mobility data across five years. In both scenarios, the method successfully differentiates between evolving and stable preference patterns. These results demonstrate the method's versatility, interpretability, and practical utility for dynamic preference analysis in various research domains. Moreover, the framework provides a foundation for future extensions that may accommodate more complex temporal dynamics, including simultaneous evolution in both items and judges.

Clustering Multivariate Trajectories Of Neighbourhood Change: Exploring Self-Organizing Maps And Alternatives

Authors:

Lizbeth Burgos-Ochoa[1*†], Katrijn Van Deun[1]

[1] Department Methodology and Statistics, Tilburg University

[*] Corresponding author [†] Presenter

Contact: l.burgosochoa@tilburguniversity.edu

Keywords:

Neighbourhood Change, Self-Organising Maps, Multidimensional Scaling, Spatial Classification

Abstract:

Understanding patterns of neighbourhood change over time is a primary objective in neighbourhood research and policymaking. However, classifying neighbourhood trajectories based on multiple (socio-demographic) indicators poses both conceptual and methodological challenges. This study explores data-driven approaches for identifying typologies of neighbourhood change using multivariate longitudinal data from Statistics Netherlands. The data include open-access indicators on housing, demographics, income, and social composition at the neighbourhood level, collected across multiple years. We address the question of how to meaningfully cluster trajectories of neighbourhood change when the joint changes of several variables define each trajectory. Building on prior work that combined Self-Organising Maps (SOMs) with k-means clustering. These SOM-based projections are then used to construct trajectories for each neighbourhood, which are clustered into typologies based on shared patterns over time. Furthermore, as a methodological contribution, we compare SOM-based trajectory construction with an alternative approach, i.e., Multidimensional Scaling (MDS). In our comparison, we assess trade-offs in interpretability, flexibility, and sensitivity to nonlinear structure. We identified a diversity of neighbourhood change patterns, including trajectories of socioeconomic stability, decline, and upward transition. The resulting typologies offer interpretable groupings of neighbourhoods that have experienced similar multivariate developments over time. This work contributes to the field of spatial classification by adapting unsupervised learning techniques for longitudinal, multivariate data. We discuss the strengths and limitations of SOMs relative to another popular dimensionality-reduction method and offer recommendations for future research.

Session - Latent variable models and dimensionality reduction methods for complex data I

Extending Landmarking To Mixture Cure Models With Longitudinal Covariates

Authors:

Marta Cipriani[1*†], Marco Alfò[1], Mirko Signorelli[2]

[1] Sapienza University of Rome
[2] Leiden University

[*] Corresponding author [†] Presenter

Contact: marta.cipriani@uniroma1.it

Keywords:

dynamic prediction, mixture cure models, landmarking, longitudinal data

Abstract:

Dynamic prediction models represent an essential class of models for personalized medicine, providing real-time updates on prognosis based on evolving patient information. Several methods, such as time-dependent Cox and joint models, have been developed to accommodate longitudinal covariates. Among these, the landmarking approach has gained popularity due to its flexibility and conceptual simplicity. However, its integration into cure models remains underexplored. In the context of cure models, and specifically mixture cure models (MCM), current applications of landmarking rely exclusively on traditional summarization techniques for time-varying covariates, notably based on the last observation carried forward (LOCF) approach. In this work, we propose a novel dynamic prediction framework that extends model-based landmarking to MCMs, with a focus on the analysis of survival in a sample of liver transplant patients. Cure models are particularly relevant in this setting, as a significant portion of transplant recipients may be considered "cured" and no longer at risk for the event of interest. Our framework separates prediction into two components: (1) the incidence component, which estimates the probability of being uncured using logistic regression and baseline covariates, and (2) the latency component, which estimates post-landmark survival among uncured individuals through a Cox proportional hazards model that incorporates summaries of longitudinal data trajectories. To improve upon traditional approaches, we introduce a model-based landmarking strategy within the cure model framework. Specifically, we model the longitudinal trajectories of patient covariates up to the landmark time (in this case, transplant) using linear mixed-effects models (LMMs) or multivariate generalized linear mixed-effects models (MGLMMs). These models allow for the estimation of individual-specific random effects, which provide a compact and informative summary of the patient's covariates' trajectory and are then used as (fixed) predictors in the cure model. We validate our method through extensive simulation studies, varying cure fractions, sample sizes, and longitudinal designs (either balanced or unbalanced). The results show consistently improved predictive performance, in terms of both area under the curve (AUC) and Brier score, when using model-based landmarking over traditional LOCF. Additionally, we apply our approach to a real-world cohort from the UNOS registry, defining the landmark at transplant and predicting post-transplant liver-related survival, while censoring unrelated deaths.

A Penalized Likelihood Approach To Dif Detection

Authors:

Michela Battauz[1*†]

[1] University of Udine

[*] Corresponding author [†] Presenter

Contact: michela.battauz@uniud.it

Keywords:

Differential Item Functioning, Item Response Theory, Latent Variables

Abstract:

Differential item functioning (DIF) occurs when the probability of giving a certain response to an item depends on the characteristics of the subjects, beyond the latent construct the underlies the responses. The detection of DIF can be performed by fitting an Item Response Theory (IRT) model separately to the responses given by different groups of subjects and comparing the item parameter estimates. However, it is first necessary to convert them to a common metric due to the constraints required to identify the model. Since the estimation of the scale conversion is affected by DIF items, an iterative procedure, which alternates between the estimation of the scale conversion and the detection of DIF, is usually applied. This talk proposes a novel likelihood-based method that simultaneously estimates scale conversion and detects items with DIF. A profile likelihood is defined by treating the item parameters on a common metric as nuisance parameters and the scale conversion coefficients as parameters of interest. The selection of items with DIF is performed employing a lasso penalty. Furthermore, covariates can be included to explain differences in item functioning across groups. One example of such variables is the position of the item within the test. The method can be used to detect DIF across multiple groups who are administered either the same test or different test forms, provided that the forms are linked through common items. The proposal is applied to the TIMSS data to detect differences among countries and the effect of item position. The performance is also illustrated through simulation studies.

Trajectory Reconstruction In Muon Scattering Tomography Using Two-Component Mixture Modelling

Authors:

Marta Ferrari[1], Alessandra R. Brazzale[1], Giovanna Menardi[1*†]

[1] Department of Statistical Sciences, University of Padua

[*] Corresponding author [†] Presenter

Contact: giovanna.menardi@unipd.it

Keywords:

Mixture Model, Muon Tomography, Scattering, Stochastic EM

Abstract:

Muons are elementary particles naturally produced in the upper atmosphere when cosmic rays interact with atomic nuclei. Muon tomography is a non-invasive imaging technique which takes advantage of their natural presence and remarkable ability to penetrate matter to investigate the interior of complex or otherwise inaccessible structures. As muons traverse materials, their trajectories are deflected in proportion to the local density, allowing for the reconstruction of three-dimensional images of the so-called scattering density within the object. This technique has been applied in diverse fields, including archaeology, geology, engineering, medicine, and security. In this work, we tackle the problem of reconstructing the internal composition of a volume using muon tomography. Within this context, it is common to assume that the scattering angles and displacements of muons passing through the volume follow a zero-mean Gaussian distribution. A major challenge arises from the fact that muon paths inside the volume are unobservable; only their entry and exit positions and directions are recorded by suitable detectors. Reconstructing the muon trajectories is then framed as a missing data problem, for which the Expectation-Maximization (EM) algorithm is a widely used method to perform maximum likelihood estimation. Since the true scattering distribution exhibits heavier tails than a Gaussian, we adopt here a more flexible model by extending the standard formulation to a mixture of two zero-mean Gaussian distributions with proportional variances. This extension introduces a twofold latent structure in the data, related both to the unknown muon trajectories and to the unobserved mixture components. These complexities make the standard Expectation step of the EM algorithm computationally infeasible. To address this, we develop a stochastic version of the EM algorithm (SEM), which approximates the intractable expectations using simulation techniques based on an Independent Metropolis-Hastings sampler. The proposed methodology is illustrated on a set of data simulated within an experiment aimed at evaluating the wear level in the inner walls of an insulating tube. The goal is to infer the material composition of the internal layers by estimating the scattering density within each voxel, a three-dimensional pixel representing a small volume element. Our results show that our estimation algorithm converges more quickly than state-of-the-art methods, although each iteration of SEM is computationally more expensive due to the presence of an additional stochastic step. Furthermore, it produces more accurate reconstructions compared to the state-of-the-art method. The results also underline the importance of using fine voxel resolution to avoid averaging heterogeneous regions, which may lead to underestimation of scattering density. A sufficiently large number of muon events is essential to ensure reliable estimates, despite the higher computational burden.

Session - Item response theory and scale validation

Leveraging Social Network Analysis For Semantic Differential Scale: An Application To Survey Data

Authors:

Ilaria Primerano[1], Maria Carmela Catone[2*†]

[1] National Research Council, Institute for Research on Population and Social Policies (CNR-IRPPS)
[2] Dept of Social and Politica Studies, Univeristy of Salerno

[*] Corresponding author [†] Presenter

Contact: mcatone@unisa.it

Keywords:

Semantic Differential, Social Network Analysis, Students' perception, Distance learning, Survey Responses

Abstract:

Advancing methodological strategies for analyzing subjective perceptions is crucial in social research, particularly when dealing with large-scale digital transitions. This study proposes a quantitative approach that leverages Social Network Analysis (SNA) methods to extract a network of adjectives from Semantic Differential scales. The approach is applied to data collected from an online survey realized during the pandemic among undergraduate students at a Southern Italian university, providing insights into the dimensional structure of their evaluations. The results highlight a positive perception of distance learning services, particularly in terms of activities conducted, use of digital platforms, online interactions, and self-study attitudes. Additionally, the study demonstrates that the perception of these services varies depending on students' prior digital skills.

A Hybrid Latent-Class Item Response Model For Detecting Measurement Non-Invariance In Mental Health Survey Data

Authors:

Gabriel Wallin[1*†], Qi Huang[2]

[1] Lancaster University
[2] Purdue University

[*] Corresponding author [†] Presenter

Contact: g.wallin@lancaster.ac.uk

Keywords:

Latent Trait Model, Latent Class Model, Measurement Non-Invariance

Abstract:

Measurement non-invariance occurs when the psychometric properties of a scale or questionnaire differ across subgroups, undermining the validity of comparisons between groups. At the item level, this manifests as differential item functioning (DIF), which arises when item response probabilities differ across subgroups after conditioning on the latent trait. DIF makes inferences about the latent variable(s) of interest more challenging, as the observed responses are influenced not only by the respondent's latent trait level but also by their group membership. There is an extensive literature on DIF detection, all with the aim of increasing the measurement quality. Traditional detection methods generally use a latent-variable framework and require both known group labels and a set of "anchor" items assumed free of DIF. More recent work allows detection when only one of these two pieces of information is available, but there is little on the case where neither subgroup labels nor anchor items are known a priori. A very recent approach addresses this in a DIF analysis framework where item-specific DIF effects are introduced, and the comparison groups are modelled as latent classes. A regularised estimator is used to simultaneously identify the latent classes and the DIF items. Although this framework has shown great promise, it is limited in a number of ways: 1) It is restricted to binary item responses and cannot accommodate ordinal responses, which are common in mental health and social science applications, 2) it can only detect uniform DIF (when the baseline probability of endorsing an item differs systematically across groups, regardless of the individual's level on the latent trait) but not non-uniform DIF (when the relationship between the latent trait and the item response differs across groups), 3) it does not provide any uncertainty quantification of the detected DIF effects. In this presentation, we introduce a framework that fills the gap. Our proposed model accommodates ordinal item responses, motivated by an application to a mental health survey which uses a Likert scale. Each respondent is assigned probabilistically to one of K+1 unobserved latent classes. Within each class, a single latent trait governs response probabilities via a proportional-odds model. Class-specific intercept shifts capture uniform DIF, while class-specific slope adjustments capture non-uniform DIF. To identify which items exhibit DIF, we impose an LASSO-type penalty on both types of DIF-effect parameters in the marginal likelihood, under the assumption that most items are DIF-free. We develop an efficient EM algorithm to solve the resulting optimisation problem, and implement a straightforward post-estimation procedure for constructing confidence intervals for the estimated DIF effects, providing uncertainty quantification not available in previous regularised DIF frameworks. We illustrate the method using data from a widely used mental health questionnaire and evaluate its performance in extensive

simulations. By handling ordinal responses without requiring predefined comparison groups or anchor items, and by accommodating both uniform and non-uniform DIF with corresponding confidence intervals, the proposed framework offers a principled strategy for assessing measurement invariance in mental health scales.

Novel Estimation Methods For Regularized Large-Scale Multidimensional Item Response Theory Models

Authors:

Travis Yang[1*†], Wilco Emons[1], Yves Rosseel[2]
Katrijn Van Deun[1]

[1] Tilburg University
[2] Ghent University

[*] Corresponding author [†] Presenter

Contact: p.yang@tilburguniversity.edu

Keywords:

Multidimensional Item Response Theory, Item Factor Analysis, Large-scale Application, Regularized Methods, Joint Maximum Likelihood Estimation, the MM Algorithm, Cardinality Approach, l1 Penalty, AO-ADMM

Abstract:

Multidimensional Item Response Theory (MIRT), also known as item factor analysis, presents substantial computational and statistical challenges in large-scale applications, particularly when the latent trait space is high-dimensional. To be more specific, traditional estimation methodsmost notably marginal maximum likelihood estimation (MMLE) combined with the expectation-maximization (EM) algorithm-require iterative integration over the latent trait distribution. This becomes computationally intensive when the number of items is large, and increasingly intractable as the number of latent dimensions grows. To overcome these limitations, Chen et al. (2019) introduced a joint maximum likelihood estimation (JMLE) framework. Their approach is promising in terms of computational efficiency, but it lacks interpretability due to the use of a post hoc rotation technique that does not yield sparse loading patterns. Building on this framework, our study retains the computational advantages of JMLE while addressing the interpretability issue by directly inducing sparsity in the loading matrix. Specifically, we propose two strategies: a cardinality-constrained approach and variable selection via an *ell* 1 penalty. While the *ell* 1 penalty is widely used in the literature, the cardinality-constrained method offers greater intuitiveness and user control, allowing researchers to explicitly specify the number of nonzero loadings (Adachi and Trendafilov, 2016). Despite its appeal, its application in large-scale MIRT context remains underexplored. This study addresses this gap by developing a novel estimation framework based on the Minimization-Majorization (MM) algorithm, enabling efficient implementation of cardinality-constrained regularization in large-scale models under orthogonal factor structures. Furthermore, we extend our approach to propose a more flexible algorithm that incorporates an *ell* 1 penalty, accommodating both orthogonal and nonorthogonal factor structures. This method is implemented within the alternating optimization-alternating direction method of multipliers (AO-ADMM) framework (Huang et al., 2016), offering enhanced flexibility. Simulation studies show that both proposed methods accurately recover the true sparse structure of the loading matrix, effectively distinguishing between zero and nonzero elements. In addition to their precision, both approaches demonstrate strong computational efficiency, even in large-scale scenarios-for instance, when the number of items reaches 500 and the number of underlying factors increases to 10. The practical advantages in interpretability are further illustrated through a real data application. Taken together, these

findings highlight the potential of the proposed methods to serve as scalable and interpretable solutions for large-scale MIRT analysis. References: Adachi, K., and Trendafilov, N. T. (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. Computational Statistics, 31, 1403-1427. Chen, Y., Li, X., and Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. Psychometrika, 84(1), 124-146. Huang, K., Sidiropoulos, N. D., and Liavas, A. P. (2016). A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. IEEE Transactions on Signal Processing, 64(19), 5052-5065.

Advanced clustering techniques for mixed-type and imbalanced data

Association-Based Spectral Clustering For Mixed Data

Authors:

Alfonso Iodice D'Enza[1*†], Cristina Tortora[2], Francesco Palumbo[1]

[1] Università degli Studi di Napoli Federico II
[2] San Jose State University

[*] Corresponding author [†] Presenter

Contact: iodicede@unina.it

Keywords:

mixed data, association-based mistaken, spectral clustering

Abstract:

In distance-based clustering, such as partitioning, hierarchical or spectral methods, observations are compared via pairwise distances defined in a metric or dissimilarity space. For continuous variables, this space is typically $\mathbb{R}^p$, with standard metrics such as Euclidean or Mahalanobis distance. For categorical variables, observations cannot be directly embedded in $\mathbb{R}^p$ in the same way, but dissimilarities can still be computed. In the case of mixed data, hybrid or custom distance functions (e.g., Gower's distance or weighted combinations) are used, allowing each observation to be compared through a scalar dissimilarity-thus enabling clustering within a unified distance-based framework. This paper focuses on spectral clustering (SC) for mixed-type data, with particular attention to the definition of the distance matrix. Several approaches have been proposed to address this challenge. Some discretize continuous variables and apply a dimension-reduction variant of SC. Others retain native data types, using Euclidean distance for continuous variables, the matching coefficient for categorical ones, and a tuning algorithm to balance their contributions. Still others define distances based on information-theoretic principles, capturing both inter- and intra-variable relationships. More recently, a unified framework has been introduced based on association-based distances for categorical data. In this approach, dissimilarity between categories is weighted according to the divergence between the conditional distributions of the remaining variables, given each category. When the conditional distributions are similar, the mismatch is down-weighted, indicating a lower degree of dissimilarity. We extend, in the context of spectral clustering, the association-based distance to the mixed-data setting, accounting for both the interdependencies among variables of the same type and the interactions between blocks of homogeneous variables.

Fast And Flexible Convex Clustering With General Weights

Authors:

Marco Stefanucci[1*†], Mauro Bernardi[2], Antonio Canale[2]

[1] University of Rome - Tor Vergata
[2] University of Padua

[*] Corresponding author [†] Presenter

Contact: marco.stefanucci@uniroma2.it

Keywords:

Convex Clustering, Adaptive methods, Fusion penalties, ADMM

Abstract:

The effectiveness of convex clustering heavily relies on the choice of weights used to penalize distances between cluster centroids. These weights are commonly determined by functions of pairwise distances between data points, specified sparsely to promote localized clustering, or designed using a combination of both approaches. In this paper, we introduce a computationally efficient and adaptable algorithm for convex clustering that accommodates a wide range of weight structures, using the Alternating Direction Method of Multipliers (ADMM). Our method is especially advantageous when the weight matrix is sparse, leading to significant computational savings. We also explore the dual formulation of convex clustering under general weight schemes, from which we derive essential tools for implementation, including a principled approach to selecting the regularization parameter to yield a meaningful number of clusters, and a novel expression for the model's effective degrees of freedom that can be directly incorporated into standard information criteria for model selection. We demonstrate the accuracy and practical benefits of our approach through simulations on synthetic data and an application to real-world data.

Polynomial Manifold Clustering

Authors:

Emil Lambert[1*†], Daniyal Kazempour[2], Peer Kröger[1]

[1] Christian-Albrechts Universität zu Kiel
[2] Christian-Albrechts-Universität zu Kiel

[*] Corresponding author [†] Presenter

Contact: emil.lambert@email.uni-kiel.de

Keywords:

Non-linear Manifold Clustering, Parameter Space Clustering, Subspace Clustering, Polynomial Manifold, Hough Transform

Abstract:

Manifold clustering aims to reveal partitions in datasets that concentrate around lower-dimensional manifolds. A notable approach adopts Hough Transform, mapping each data point to a function corresponding to all linear manifolds containing that point. Regions in the resulting parameter space where many functions are similar in value or intersect indicate dense manifolds in the data space. Later developments proposed the use of geometric representations—such as the Hesse Normal Form for hyperplanes—mapping k-tuple samples to single points in parameter space. Facilitating distance-based identification of dense regions, thus enabling the use of standard clustering methods. A key feature of this family of methods is that they make no assumption of locality in data space: points lying on the same manifold can be arbitrarily distant from one another. This allows the detection of clusters that are intersecting, disconnected, or fragmented in data space-situations that challenge many other manifold clustering methods. We extend this framework from linear to *polynomial manifolds* by constructing a parameter space consisting of the coefficients of implicit polynomials. For a polynomial of total degree k in d variables,

$$P(x_1, \ldots, x_d) = \sum_{i_1+\cdots+i_d \leq k} c_{i_1,\ldots,i_d} x_1^{i_1} \cdots x_d^{i_d},$$

we seek clusters of data points lying near the zero set of such polynomials—that is, subsets approximately satisfying $|P(x)| < \varepsilon$, where $\varepsilon > 0$ accounts for jitter. To construct this space, we adopt the established strategy of sampling point tuples of size $N = \binom{d+k}{k}$, the minimal number required to determine a unique degree-k polynomial. Each such tuple defines a parameter vector, and dense regions in this space correspond to polynomial structures in the data. The resulting parameter space is high-dimensional and semantically heterogeneous: constant, linear, and higher-order coefficients vary in scale and carry distinct geometric meaning, making the application of traditional clustering algorithms a non-trivial task. Addressing this complexity, we (1) adapt techniques known from traditional Hough Transform applications to focus computation on meaningful regions, (2) introduce suitable distance measures, and (3) propose a hierarchical clustering scheme that organizes polynomials by their homogeneous components, starting with highest-degree terms and progressively subdividing clusters based on lower-degree coefficients.

Session - Local modeling and advanced inference for spatial and functional data

When Standard Calibration Metrics Fail In Evaluating Classifier Calibration: A Simulation Study

Authors:

Ndeye Awa Dieye[1*†], Giorgio Russolillo[1], Ndèye Niang[1]

[1] CNAM

[*] Corresponding author [†] Presenter

Contact: ndeye-awa.dieye@lecnam.net

Keywords:

calibration metrics, supervised classification, true underlying probability, loss scores

Abstract:

Right evaluation of classifiers is essential to provide useful support to decision-making. A classifier generally produces two outputs: the predicted class and the predicted probability of the event. To assess its quality, both discrimination - the ability to differentiate between classes - and calibration - how close the predicted probabilities are to the true ones - are considered. While one can directly evaluate the quality of classification by comparing true and predicted class, assessing the reliability of predicted probabilities (calibration) is a more challenging task. A key issue is that true probabilities are unknown in real data. As a result, calibration metrics rely on comparing predicted values to empirical event frequencies. Many metrics have been proposed, but they often lead to different conclusions, with no clear consensus on which to choose. In this work, we simulate datasets with controlled True Distribution Shapes (TDS) and we compare the performance of several classifiers in terms of calibration. The knowledge of the true probabilities allows us to compare them to the predicted ones in order to measure the true calibration of the classifiers. We observe how well several classifiers preserve true probability distributions using Kolmogorov-Smirnov tests. Moreover, we measure the true reliability of the predicted probabilities using the mean squared error between the predicted and true probabilities as a reference metric. Finally, we assess how well standard calibration metrics (Expected Calibration Error, Brier Score, Log-Loss) reflect the true quality of the classifier's predicted probabilities. Our results show that logistic regression, SVM, and neural networks tend to better preserve the TDS. In contrast, random forests and naïve bayes tend to distort it, even after recalibration. Metrics like ECE often fail to detect these distortions and may give a false sense of good calibration. In comparison, loss-based metrics, especially when decomposed into epistemic and refinement components, provide more reliable results. Post-hoc calibration does not fully overcome the limitations of some models or metrics. When accurate probability estimation is the goal, we recommend relying on loss-based metrics to choose the best classifier when evaluating calibration.

New Insights Into Volleyball Setter Evaluation Through Spatial-Outcome Clustering

Authors:

Martina Narcisi[1*†], Garritt L. Page[2], Gilbert W. Fellingham[2]

[1] University of Bologna
[2] Brigham Young University, Department of Statistics, Provo, USA

[*] Corresponding author [†] Presenter

Contact: martina.narcisi2@unibo.it

Keywords:

ordinal Product Partition Model, spatial clustering, ranking, volleyball

Abstract:

The accurate evaluation of individual player performance in team sports is a complex task due to the intricate interactions among players and the multiple factors contributing to team success. This complexity is particularly evident in volleyball when assessing the performance of setters, whose role as playmakers is neither purely offensive nor defensive, but rather strategic and coordinative. The setter's ability to optimally distribute sets under varying game conditions directly influences the team's offensive potential, yet traditional evaluation metrics often fail to adequately capture this contribution. Conventional methods tend to rely heavily on the final outcome of the attack, attributing setter performance largely to the hitter's success, and thus potentially overlooking the setter's skill in positioning and decision-making. In this work, we propose a methodological framework specifically designed to isolate and quantify the setter's individual contribution, taking into account both the spatial location of the set and its subsequent outcome. Our approach leverages detailed data from the 2018 Men's World Championship, encompassing the exact court position of each set and whether the attack following the set resulted in a successful point (kill) or not. The modeling strategy employs an ordinal Product Partition Model (PPMx), which facilitates the grouping of sets into clusters based on their spatial characteristics and binary outcomes. This clustering allows for the derivation of cluster-specific probabilities that reflect the likelihood of a successful attack from each set location. Following the clustering stage, we calculate a Setter Score (SS) for each individual setter. This metric is constructed to integrate the difficulty of the set location with the corresponding attack outcome, thereby providing a more nuanced assessment of setter performance. Specifically, the Setter Score rewards those setters who are able to facilitate successful attacks even when setting from less favorable positions on the court, recognizing their technical skill, adaptability, and tactical awareness. Conversely, setters who achieve success predominantly from optimal positions receive comparatively lower scores, as their task is inherently less challenging. Importantly, the Setter Score allows not only for the evaluation of individual performance but also for the construction of a ranking among setters participating in the tournament. By analyzing the distribution of Setter Scores across all players, we identify those setters who consistently enable successful attacks across a range of positional challenges. These individuals are positioned at the upper end of the ranking and emerge as candidates for the best setters of the championship. This ranking framework provides a more comprehensive and objective tool for comparing setter performance, offering valuable insights for coaches, analysts, and researchers. Furthermore, the proposed methodology is flexible and generalizable, allowing adaptation to other sports con-

texts where spatial and categorical outcome data are available. Overall, this work contributes to the growing field of performance analytics by addressing the unique challenges of setter evaluation in volleyball, emphasizing the critical but often underappreciated aspects of their role that traditional metrics fail to capture.

Unisound: a Sustainable Design Based Active Noise Cancellation Device For Workplaces

Authors:

Domenico Persia[1], Vito Santarcangelo[2], Angelo Romano[2]
Giulio Setzu[2], Angelo Lamacchia[3], Alessandro D'Alcantara[2]
Michele Di Lecce[2], Saverio Gianluca Crisafulli[2], Massimiliano Giacalone[3*†]
Gianfranco Piscopo[4], Sergio Vitullo[2]

[1] Uniservice Srl
[2] iInformatica Srl
[3] Università della Campania "Luigi Vanvitelli"
[4] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: maxgiacit@yahoo.it

Keywords:

sound pollution, sustainable design, noise-cancellation approach

Abstract:

Noise pollution in workplaces and public environments is an increasing concern due to its impact on comfort, concentration, and overall well-being. This study presents the development of an active noise cancellation (ANC) device designed to be inserted into the soil of indoor ornamental plants, enabling the integration of acoustic control within sustainable interior design elements. The device "UNISOUND" consists of a vertical rod-like structure housing microphones, speakers, and environmental sensors, capable of detecting ambient noise and emitting anti-phase sound waves to suppress it. A widened base ensures stability in the soil, while the vertical body features LED indicators and sound-emitting openings to support acoustic and visual functionality. By being embedded in indoor greenery, the device achieves a discreet and camouflaged presence, seamlessly blending with the environment without disrupting its visual or spatial harmony. The work shows tests conducted in controlled environments show a perceptible noise reduction suggesting promising applications for a new class of hybrid acoustic solutions that merge sustainable design principles with embedded noise-cancellation technology.

Session - Model-Based clustering and representation learning for structured and incomplete data

Understanding Students' Paths In Italian Higher Education: A Bayesian Network Approach

Authors:

Marta Campagnoli[1*†], Marta Magnani[1], Silvia Salini[1]

[1] University of Milan

[*] Corresponding author [†] Presenter

Contact: marta.campagnoli@unimi.it

Keywords:

data visualization, graphical models, administrative data, what-if scenario

Abstract:

Over the past decade, university enrollment patterns in Italy have undergone significant transformations. The COVID-19 pandemic, the rapid expansion of distance learning opportunities, and broader social and technological changes have reshaped both the higher education landscape and students' expectations. Today, higher education is a dynamic and evolving phenomenon, shaped by an increasing variety of institutional formats, disciplinary offerings, and individual aspirations. In this context, understanding how the university student population is changing-and what drives students in their educational choices-has become a central question for both researchers and policymakers. This paper proposes a flexible methodological framework to investigate the determinants of university enrollment in Italy, with a specific focus on the choice between different institutional paths such as traditional universities vs. distance learning institutions. The analysis is based on Bayesian Networks, a class of probabilistic graphical models that allow the representation and interpretation of complex dependencies among categorical variables in high-dimensional data. We apply this framework to the Italian National Registry of University Students (ANS - Anagrafe Nazionale degli Studenti e dei Laureati), an administrative dataset covering academic years from 2010/11 to 2023/24, which includes detailed demographic, educational, and institutional information at the individual level. Using Bayesian Networks, we aim to identify the most relevant combinations of factors-such as gender, age group, geographical origin, and type of secondary school-that influence the likelihood of enrolling in specific types of universities or academic areas. A distinctive strength of Bayesian Networks lies in their ability to support both forward and inverse reasoning. This enables what-if sensitivity scenarios: for instance, given a particular student profile, the model can estimate the probability of choosing a certain type of institution or field of study. Conversely, given a target outcome (e.g., enrolling in a distance learning university), the model can identify the most typical characteristics of students likely to make that choice. This framework offers a rigorous and interpretable approach for visualizing complex enrollment dynamics and contributes to a better understanding of the paths characterizing students' decisions in an increasingly diversified higher education system.
Acknowledges financial support within the 'Fund for Departments of Excellence academic funding' provided by the Ministero dell'Università e della Ricerca (MUR), established by Stability Law, namely 'Legge di Stabilità n.232/2016, 2017' - Project of the Department of Economics, Management, and Quantitative Methods, University of Milan.

Subjective Perceptions Of The Financial Situation Of Polish Households And Their Savings

Authors:

Małgorzata Grzywińska-Rąpca[1], Aneta Ptak-Chmielewska[2*†]

[1] University of Warmia and Mazury in Olsztyn
[2] Warsaw School of Economics

[*] Corresponding author [†] Presenter

Contact: aptak@sgh.waw.pl

Keywords:

Households, Subjective assessment of financial situation, Household savings, Financial security, Structural modelling

Abstract:

Studies on subjective perceptions of the financial situation play a role in shaping the level and structure of household savings. The aim of the study was (1) to analyse the relationship between the savings of Polish households and their perceived financial situation, (2) to determine the impact of accumulated savings on the sense of financial security and the overall assessment of financial stability among Polish households. As savings play a key role in households' sense of financial security, the study focused on presenting the economic context of household savings and showing trends in their level and distribution, taking into account the main forms of accumulating financial resources. Households' perceptions of their own financial situation, in the form of subjective assessments, depend not only on actual financial resources but also on individual expectations and aspirations. Structural modelling based on data sets from the Household Budget Survey (GUS, 2022) conducted among 30,432 households aimed to show that savings are an important, but not the only element influencing the subjective sense of financial stability of Polish households, and that the diversity of socio-economic factors shapes the financial feelings and assessments of individual socio-economic groups.

Enhancing Sentiment Detection In Social Media Using Llms And Embedding-Based Clustering

Authors:

Domenica Fioredistella Iezzi[1*†], Roberto Monte[1]

[1] Tor Vergata University

[*] Corresponding author [†] Presenter

Contact: stella.iezzi@uniroma2.it

Keywords:

Sentiment Analysis, Deep Learning, Embedding, Text Clustering

Abstract:

Large Language Models (LLMs) have transformed the field of natural language processing (NLP), achieving state-of-the-art performance across a wide range of linguistic tasks, including sentiment analysis. Unlike traditional machine learning methods that rely on manual feature engineering and task-specific datasets, LLMs such as ChatGPT, LLaMA, and BERT are pretrained on vast textual corpora and can capture nuanced semantic and contextual relationships across domains. This makes them particularly well-suited for extracting sentiment signals from informal, noisy, and unstructured data, such as user-generated content on social media platforms. In this study, we adopt a multi-step pipeline to analyze Twitter (now X) content related to two automotive brands: Toyota Camry and FIAT 500 (L500). The analysis begins by generating dense sentence embeddings using Sentence-BERT. To uncover latent semantic structures, we apply Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, followed by HDBSCAN, k-means and k-medoids to identify coherent topic clusters within the corpus. These clusters provide a thematically organized foundation for interpreting user discourse. Subsequently, we perform content classification to distinguish among tweet types (e.g., user experiences, complaints, advertising) and finally estimate sentiment using multiple transformer-based models. Specifically, we compare the outputs of CardiffNLP's Twitter-RoBERTa, BERT, ChatGPT, and LLaMA, evaluating their ability to detect sentiment polarity, subtle emotional gradations, and consistency with human interpretations. We apply this approach to two datasets: approximately 163,000 English-language tweets about the Toyota Camry (2009-2020), and 20,387 tweets about the FIAT 500L (2012-2018). We aim to utilize the signals as components of a predictive state space model for the sales volumes. To achieve this, we evaluate and compare the quality and coherence of sentiment signals generated by various LLMs, examining how these signals vary across semantic clusters, time periods, and brand contexts. This layered methodology demonstrates how embedding-based clustering, combined with advanced sentiment models, can improve the interpretability and robustness of sentiment analysis in real-world, domain-specific social media data.

Session - Latent structures and interpretability

Modeling And Estimating Skewed And Heavy-Tailed Populations Via Unsupervised Mixture Models

Authors:

Marco Bee[1*†], Flavio Santi[1]

[1] University of Trento

[*] Corresponding author [†] Presenter

Contact: marco.bee@unitn.it

Keywords:

Pareto tail, EM algorithm, Mixture distribution, Classification

Abstract:

In many fields of application, the population distribution is skewed and heavy-tailed. In some of these cases, it may be difficult to find a single probability model for all the observations, especially if one needs to fit the tail with a high degree of precision. We develop an unsupervised mixture model for non-negative, skewed and heavy-tailed data, such as losses in actuarial and risk management applications. The mixture has a lognormal component, which is usually appropriate for the body of the distribution, and a Pareto-type tail, aimed at accommodating the largest observations, since the lognormal tail often decays too fast. We show that maximum likelihood estimation can be performed by means of the EM algorithm and that the model is quite flexible in fitting observations from different data-generating processes. Simulation experiments and a real-data application to automobiles claims suggest that the approach is equivalent in terms of goodness-of-fit, but easier to estimate, with respect to two existing distributions with similar features. In general, when compared to competing models, the main advantages are an easy interpretation and a reliable estimation procedure characterized by a low computational burden. Moreover, an important by-product of the EM algorithm is the estimate of the posterior probability of the observations, which can be employed for classification and cluster analysis purposes. In terms of flexibility, the static lognormal-GPD mixture yields excellent results when used for fitting data in mis-specified setups, and allows the investigator to estimate risk measures quite precisely. In particular, both in simulation and in real-data setups, Value-at-Risk measures computed via the static lognormal-GPD model are less variable than via similar composite models. Finally, the distribution can be easily generalized: in particular, setting up and estimating mixtures with different component densities is almost immediate. To ease the implementation of the methodology, we have also developed the {R} package {lognGPD}, containing the codes for simulating and estimating the lognormal-GPD distribution proposed in this paper, which is available at {https://github.com/marco-bee/lognGPD}.

Unravelling Latent Cognitive Dissonance In E-Commerce: A Profile-Based Analytical Framework

Authors:

Furio Urso[1], Nicola Argentino[1], Antonino Abbruzzo[1]
Reza Mohammadi[2], Kevin Pak[2], Maria Francesca Cracolici[1*†]

[1] Department of Economics, Business and Statistics, University of Palermo
[2] Faculty of Economics and Business, Section Business Analytics, Amsterdam Business School

[*] Corresponding author [†] Presenter

Contact: mariafrancesca.cracolici@unipa.it

Keywords:

Browsing behaviour, Traceable and anonymous users, Mixture hidden Markov model, Logistic regression

Abstract:

The digitalisation of the tourism industry has enabled companies to access a wealth of information about potential customers exploring their e-commerce websites. This development has heightened the need to leverage clickstream data - capturing every interaction within a website - to gain deeper insights into consumer behaviour. As shown by the marketing literature, a cornerstone for analysing consumer behaviour is the theory of cognitive dissonance. Cognitive dissonance arises from a state of psychological discomfort caused by conflicting choices, ideas, or beliefs. Consumers may experience cognitive dissonance during any phase of the customer journey-pre-purchase, purchase, or post-purchase. A critical challenge in studying online consumer behaviour lies in addressing the cognitive dissonance that occurs during the online pre-purchase phase, which can significantly influence purchase likelihood. Specifically, at the onset of browsing, consumers may experience discomfort due to information overload or low engagement. Conversely, toward the end of their browsing activity, they might face indecision when evaluating alternative options. It is evident that different experiences of cognitive dissonance influence browsing behaviour and, subsequently, the probability of purchase. It is important to note that cognitive dissonance manifests in various ways, shaped by individual characteristics and technical attributes of the website. This complexity makes it a latent and challenging construct to measure a priori during the browsing phase. Our study adopts a two-stage approach to address these challenges. First, we identify distinct customer browsing profiles that account for variations in cognitive dissonance. Second, we investigate how these profiles influence purchase probability. In the first stage, a Mixture Hidden Markov Model (MHMM) is applied to both traceable and anonymous users, uncovering six distinct browsing profiles. The second stage focuses on traceable users who have multiple visits to the website. These visits, often associated with different browsing profiles, are analysed to assess how varying behaviours impact purchase likelihood. For this purpose, a logistic regression analysis is employed. Empirical findings reveal that early-stage profiles, characterised by information overload or casual browsing, reduce the likelihood of purchase. In contrast, late-stage profiles, marked by focused exploration and indecision among alternatives, significantly increase purchase probability. From a practical standpoint, our findings offer actionable insights for companies aiming to design tailored strategies that address cognitive dissonance at each stage of the customer journey, ultimately steering consumers towards purchase. For early-stage users, personalised content and

targeted incentives can alleviate information overload and encourage engagement. For late-stage users, streamlined pathways and tailored offers can mitigate indecision and facilitate purchase completion. Additionally, the proposed framework empowers businesses to use browsing profiles to directly engage traceable users, enabling interventions that effectively guide them towards completing their purchase.

A Sparse And Interpretable Post-Clustering Logistic Regression For Modelling Higher Education Dropouts

Authors:

Andrea Nigri[1], Massimo Bilancia[2*†], Barbara Cafarelli[1]
Samuele Magro[2]

[1] University of Foggia
[2] University of Bari Aldo Moro

[*] Corresponding author [†] Presenter

Contact: massimo.bilancia@uniba.it

Keywords:

Higher Education Attrition, Interpretability, Logistic Regression with Cluster-Specific Interactions, Group-LASSO

Abstract:

University student attrition presents a significant global challenge for tertiary education systems. While machine learning methodologies can achieve substantial predictive accuracy on specific datasets, their practical implementation by policymakers for the unsupervised identification and characterization of student subgroups at heightened risk of dropout remains notably constrained. This paper introduces a specialized logistic regression model, specifically tailored for the analysis of university dropout. Logistic regression maintains its status as a cornerstone among robust statistical models, largely owing to the straightforward interpretability of its parameters as odds ratios. Our methodology extends this conventional framework by integrating population heterogeneity. This is accomplished via a preliminary clustering algorithm designed to delineate latent student subgroups, each exhibiting distinct dropout propensities. These propensities are subsequently modeled using cluster-specific effects. We offer a comprehensive interpretation of the model parameters within this expanded framework and augment interpretability further by enforcing sparsity through a customized variant of the LASSO algorithm. To illustrate the practical utility of the proposed methodology, we provide an extensive case study grounded in the Italian university system, demonstrating the systematic application of all developed tools.

Session - Recursive partitioning and related methods

Posterior Inference For Shapley Values Through Bayesian Horseshoe Estimation Of Tree-Based Prediction Rule Ensembles

Authors:

Giorgio Spadaccini[1*†], Marjolein Fokkema[1], Mark van de Wiel[2]

[1] Leiden University
[2] Vrije Universiteit Amsterdam

[*] Corresponding author [†] Presenter

Contact: g.spadaccini@fsw.leidenuniv.nl

Keywords:

Bayesian uncertainty quantification, Tree ensembles, Shapley values, Interactions, Nonlinearity

Abstract:

In many scientific research settings, Machine Learning (ML) is gaining increasing popularity in hypothesis-free discovery of risk (or protective) factors and groups. ML is strong at discovering nonlinearities and interactions, but this power of ML is compromised by a lack of methods to reliably infer such effects on a local level. This is needed as the high complexity of both reality and ML models imply that the influence of risk factors strongly varies across subjects and their unique combination of features. While local measures of feature attributions can be combined with ML models such as tree ensembles, uncertainty quantifications for these measures remain only partially available and oftentimes unsatisfactory. We propose RuleSHAP, a framework for using rule-based, hypothesis-free discovery that combines sparse Bayesian regression, tweaked tree ensembles and Shapley values to carry out a one-step procedure that both detects and tests complex patterns at the individual level. We compare our model with linear regression and with state-of-the-art tree ensemble models on simulated data. Moreover, we apply our machinery to data from an epidemiological cohort to detect and infer several complex effects for high cholesterol level and blood pressure. We illustrate Shapley values and their uncertainty quantifications for the most important features, allowing to infer what is relevant and for which subgroups of the population. From these experiments, we conclude that the combination of rule-based prediction, the horseshoe prior and our derived Shapley values provides a powerful and interpretable method that combines the flexibility of tree ensembles with statistical inference.

Weighted Logistic Oblique Tree For Regression

Authors:

Andrea Carta[1*†], Luca Frigau[1]

[1] University of Cagliari, Department of Economics and Business Sciences

[*] Corresponding author [†] Presenter

Contact: andrea.carta88@unica.it

Keywords:

GLM, Decision Tree, Classification, Regression, Oblique Tree

Abstract:

Decision trees are among the most popular nonparametric models for supervised learning, thanks to their simplicity and interpretability. However, traditional decision trees rely on axis-parallel splits and often fail to capture complex linear relationships among variables. Oblique decision trees address this limitation by allowing splits along hyperplanes defined by linear combinations of predictors. Yet, identifying the optimal oblique split at each node is computationally challenging. Thus, heuristic methods are usually applied. In this work, we propose a novel method for regression tasks called Weighted Oblique Logistic Tree for Regression (WOLTReg). At each node, WOLTReg applies a weighted logistic regression classifier to identify the best oblique split. A key innovation lies in the use of weights derived from the response variable, which improves the model's ability to preserve information lost during the dichotomization process required to apply the classifier. Specifically, the response variable is transformed into a binary target across multiple quantile thresholds, and each observation is weighted according to the absolute scale value of the target variable. This allows the model to emphasize observations in the extreme tails of the distribution, improving impurity reduction. WOLTReg also includes a variable selection step that restricts the candidate features to the most correlated with the response, reducing overfitting and computational cost. We evaluate our method on 60 real-world benchmark datasets, comparing its performance with other oblique tree methods and standard decision trees. The results show that WOLTReg consistently ranks among the best methods in terms of predictive accuracy, particularly when a small number of features is used to compute the separating hyperplane.

A Powerful Random Forest Featuring Linear Extensions (RaFFLE)

Authors:

Jakob Raymaekers[1], Peter Rousseeuw[2], Thomas Servotte[1*†]
Tim Verdonck[1], Ruicong Yao[2]

[1] University of Antwerp
[2] KU Leuven

[*] Corresponding author [†] Presenter

Contact: Thomas.Servotte@uantwerpen.be

Keywords:

Algorithm, Consistency, Linear Model Trees, Ensemble, Machine Learning, Regression

Abstract:

Random forests are a cornerstone of regression modeling, prized for their scalability and ability to uncover complex nonlinear patterns. However, their typical base learners-CART decision trees-struggle at modeling linear relationships due to their piecewise-constant nature. We introduce RaFFLE (Random Forest Featuring Linear Extensions), an ensemble framework designed to overcome this limitation by leveraging PILOT trees (Piecewise Linear Organic Trees) as base learners. PILOT trees blend the simplicity and speed of classical decision trees with the adaptability of linear models. They dynamically select among five node-level regression types: constant, univariate linear, piecewise constant, broken-linear, or fully piecewise linear. Model selection is guided by Bayesian Information Criterion (BIC), ensuring a rigorous balance between predictive fit and model complexity. This organic fitting reduces the need for post-hoc pruning and maintains computational efficiency akin to CART. Integrating PILOT into a forest requires stimulating diversity among trees-an essential property for ensemble strength. We introduce two key enhancements: (1)Tunable regularization via α: We modify the BIC's penalty structure, interpolating between over-regularized ($\alpha = 1$) and more flexible ($\alpha = 0$) fits, thus enabling control over the variance-bias trade-off in each tree. (2) Node-level feature subsampling: Beyond standard tree-level feature bagging, RaFFLE randomly samples feature subsets at each node. Combined with bootstrap sampling, this further enriches the diversity of tree-specific fits. We formally establish two theoretical properties: (1) consistency under general data-generating conditions, and (2) accelerated convergence rates when the true underlying model is linear. Specifically, we prove that RaFFLE achieves near-optimal error decay $O\left(\frac{\log n}{n^{\delta}}\right)$ when set appropriately. To assess practical performance, we benchmark RaFFLE across 136 diverse regression datasets drawn from UCI and PMLB repositories. Results show consistent superiority over CART trees, classical random forests, linear methods (Lasso/Ridge), and even state-of-the-art gradient-boosted frameworks such as XGBoost-across both linear and nonlinear problem settings. These gains come without substantial computational cost, positioning RaFFLE as a powerful "best-of-both-worlds" method. In conclusion, RaFFLE brings together the interpretability and speed of tree-based methods with the expressive power of linear models, making it a versatile and theoretically sound tool for regression tasks. Its capacity to flexibly adapt to data structure, while providing performance guarantees and empirical advantages, marks a significant step forward in ensemble learning.

"Can You Explain That?" E2tree, Shap, And Lime For Interpretable Random Forests

Authors:

Agostino Gnasso[1*†], Massimo Aria[1]

[1] Department of Economics and Statistics, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: agostino.gnasso@unina.it

Keywords:

E2Tree, SHAP, LIME, Machine Learning, Classification

Abstract:

An explainability methodology for Random Forest (RF) models, called Explainable Ensemble Trees (E2Tree), has been recently introduced. E2Tree methodology enhances interpretability by transforming the decision structure of RF into a simplified, explainable tree while preserving accuracy. It achieves this by leveraging the co-occurrence of observations in RF decision paths to create a globally interpretable representation of the model's decision process. In this study, we apply three distinct explainability techniques, E2Tree, Shapley Additive Explanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME), to analyze the decision-making process of RF models. While SHAP and LIME offer global and local perspectives, respectively, particular emphasis is placed on E2Tree for its ability to summarize the ensemble's decision logic into a single, interpretable structure. We illustrate these methods using the German Credit dataset, focusing on how variables such as loan duration, credit amount, housing status, and financial liquidity influence credit risk assessment. Acknowledgments: this research has been financed by the following research projects: - PRIN 2022 SCIK- HEALTH (Project Code: 2022825Y5E - CUP: E53D23006110006); - PRIN 2022 PNRR The value of scientific production for patient care in Academic Health Science Centres (Project Code: P2022RF38Y - CUP: E53D23016650001).

Session - Statistical models for sequential, functional, and structured data

Clustering Dolphin Signature Whistles With Dirichlet Process Mixtures

Authors:

Gianluca Mastrantonio[1*†], Giovanna Jona Lasinio[2], Alessio Pollice[3]

[1] Politecnico di Torino
[2] Sapienza, Università di Roma
[3] Università degli Studi di Bari Aldo Moro

[*] Corresponding author [†] Presenter

Contact: mastrantonio.gluca@gmail.com

Keywords:

Bayesian Inference, Sounds, Dirichlet process

Abstract:

Bottlenose dolphins (Tursiops truncatus) produce individually distinctive sounds, called "signature whistles," which differ from each other in their stereotyped shapes (i.e., frequency modulation patterns). The ability to identify individual dolphins through their signature whistles has practical applications in monitoring population dynamics, tracking movement patterns, and assessing the health of wild populations. Indeed, signature whistles exhibit a high degree of individual variability, adjusted in frequency and duration in response to specific contextual elements. This variability makes classifying these sounds and assigning them to specific individuals much more challenging. Hence, new classification methods are needed to make the process less time-consuming and more rigorous. Even though we know which dolphin produces which sound, we do not want to use this information since we are also interested in defining how similar the sounds of different dolphins are. From the spectrogram of each recorded sound, we extract the signature whistle, which is then represented as a change in frequency over time through a spline. Taking inspiration from Size-and-Shape modeling, we decomposed the information of the signature whistles into the shape of the sound, it size/length, and position on the frequency axis. Clustering of the whistles is then performed based on these three components, where the shape of the sound is represented as an Ornstein-Uhlenbeck process. To estimate both the number of latent clusters and the model parameters, the model is formalized using a Dirichlet process mixture model. The model accurately describes the shape of the animal's whistle and identifies which animals produce similar sounds. A cross-validation experiment is used to assess the model's ability to cluster new sounds and to evaluate its strengths and weaknesses.

A State-Restricted Hidden Markov Model For Authorship Attribution Of The Deutero-Pauline And Pastoral Epistle

Authors:

Josiah Leinbach[1], Xuwen Zhu[2*†], Shuchismita Sarkar[1]

[1] Bowling Green State University
[2] The University of Alabama

[*] Corresponding author [†] Presenter

Contact: xzhu20@crimson.ua.edu

Keywords:

hidden Markov model, EM algorithm, stylometry, authorship attribution, Pauline Epistle

Abstract:

The New Testament contains thirteen epistles attributed to the Apostle Paul, all of which were traditionally accepted as authentically Pauline by early Christian theologians. Since the 19th century, however, many scholars have questioned Paul's authorship of certain epistles due to differences in vocabulary and writing style compared to the undisputed Pauline epistles. In particular, two clusters of epistles, known as the Deutero-Pauline Epistles (Ephesians, Colossians, and 2 Thessalonians) and the Pastoral Epistles (1 Timothy, 2 Timothy, and Titus) have been subject to the most doubt. This article presents a novel state-restricted hidden Markov model that constructs a constrained state space for the undisputed Pauline epistles and an unrestricted state space for other epistles. The model jointly analyzes all thirteen epistles and some additional biblical texts, studying transitions between parts of speech to classify sentences based on Pauline and non-Pauline style detection. Then, informed by New Testament scholarship, the result of the model is interpreted and the possibility of Pauline authorship for the Deutero-Pauline and Pastoral Epistles has been examined.This paper adopts the "Informed Canon" approach which applies reasonable, minimal, and justifiable priors, reducing the risk of producing inexplicable results. The Undisputed Seven epistles will serve as the foundational Informed Canon in this study. A hidden Markov model (HMM), a state-of-the-art method for analyzing sequence data, was initially developed for speech recognition. In this study, the HMM jointly models all sentences in the Pauline corpus, including the Undisputed Seven, Deutero-Pauline Epistles (DPE), and Pastoral Epistles (PE). The model-based HMM assumes that the style of each sentence depends only on the preceding sentence, enabling the classification of sentences into distinct stylistic categories (hidden states).

Active Learning For Sequential Classification With Partial Labels

Authors:

Christian Capezza[1*†], Antonio Lepore[1], Kamran Paynabar[2]

[1] University of Naples Federico II, Department of Industrial Engineering
[2] Georgia Tech, School of Industrial and Systems Engineering

[*] Corresponding author [†] Presenter

Contact: christian.capezza@unina.it

Keywords:

Statistical Process Monitoring, Hidden Markov Model, Sequential Data Analysis, Imbalanced Classification

Abstract:

In many classification tasks involving sequential or streaming data, acquiring labeled observations is costly or time-consuming, making it infeasible to fully label all instances. This limitation is particularly evident in industrial process monitoring, where determining whether the state of a process is in control or out of control often requires expert intervention or specialized testing procedures. To address this challenge, we propose a stream-based active learning framework to support classification in scenarios where only a limited number of labels can be obtained. The method integrates partially hidden Markov models (pHMMs), effectively capturing temporal dependencies while combining labeled and unlabeled data for probabilistic classification. At each time step, the active learning framework evaluates incoming observations and decides whether to request a label, using a novel dual criterion that balances exploitation (refining classification boundaries) and exploration (identifying new or unknown process states). An online fitting strategy is developed for updating the pHMM over time, including a robust initialization procedure specifically tailored for highly imbalanced classification settings, common in quality monitoring applications where most data reflect nominal operation. The proposed active learning framework also incorporates a model selection mechanism to dynamically determine the number of latent states in the process. The proposed framework's performance is evaluated through an extensive simulation study and applied to a case study involving resistance spot welding in the automotive industry. In this case study, process profiles are continuously collected, and labels are selectively obtained through ultrasonic inspection. Acknowledgements: The research activity of C. Capezza and A. Lepore was supported by Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 5 Componente 2, Investimento 1.3-D.D. 1551.11-10-2022, PE00000004 within the Extended Partnership MICS (Made in Italy - Circular and Sustainable). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Matrix Variate Hidden Markov Models With Skewed Emissions

Authors:

Michael Gallaugher[1*†], Xuwen Zhu[2]

[1] Baylor University
[2] University of Alabama

[*] Corresponding author [†] Presenter

Contact: Michael_Gallaugher@baylor.edu

Keywords:

Hidden Markov Models, Three-way data, Time Series

Abstract:

Data collected today have increasingly become more complex and cannot be analyzed using regular statistical methods. Matrix variate time series data is one such example where the observations in the time series are matrices. Herein, we introduce a set of three hidden Markov models using skewed matrix variate emission distributions for modeling matrix variate time series data. Compared to the hidden Markov model with matrix variate normal emissions, the proposed models present greater flexibility and are capable of modeling skewness in time series data. Simulated and real data on Texas university salaries will be used for illustration.

Session - Clustering and community detection for structured and complex data

Extending The Boosted-Oriented Probabilistic Clustering To The Unit Hypersphere: A Textual Data Perspective

Authors:

Rebecca Rivieccio[1*†], Roberta Siciliano[2]

[1] Department of Physics, University of Naples Federico II, Naples, Italy
[2] Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy

[*] Corresponding author [†] Presenter

Contact: rebecca.rivieccio@unina.it

Keywords:

Boosting, Clustering, Text Mining, Directional data

Abstract:

Clustering techniques play a crucial role in uncovering patterns and structures within complex and high-dimensional datasets. In this study, we propose an extension of the boosted-oriented probabilistic clustering algorithm, originally developed for time series data, to effectively handle directional data. The proposed formulation addresses the unique challenges posed by data constrained to the unit hypersphere, such as those found in text mining applications where documents are represented as normalized vectors. The proposed approach incorporates the core principles of boosting to iteratively refine cluster assignments, emphasizing the most representative and impactful observations in the dataset. Unlike traditional boosting, where the focus is on the hardest-to-classify observations, this algorithm inverts the paradigm by assigning greater weight to objects with higher cluster membership probability. The method is applied to a textual dataset represented as a document-term matrix, where vectors are normalized to unit length to eliminate the influence of text length. Distances between documents are computed using cosine similarity, a widely adopted metric for effectively analyzing sparse and high-dimensional data. The algorithm produces a soft clustering output, represented by a probabilistic membership matrix, which allows each document to belong to multiple clusters to varying degrees. This characteristic proves particularly valuable in text mining, where documents often overlap in content or thematic structure. The method is also compared with other clustering techniques commonly used for directional data, in order to explore differences in behavior and resulting document groupings. The obtained clusters reveal meaningful structures, some of which align with predefined categories, while others suggest new, potentially insightful patterns. Overall, this study presents a novel, non-parametric probabilistic clustering approach for directional data that does not require a fuzziness parameter, with potential applications in various domains involving unit-normed vector representations.

Hybrid Single Linkage Clustering

Authors:

Ilaria Mozzetta[1*†], Maurizio Vichi[1], Tiziano Iannaccio[1]

[1] Sapienza University of Rome

[*] Corresponding author [†] Presenter

Contact: ilaria.mozzetta@uniroma1.it

Keywords:

Hybrid clustering, Single linkage, Non hierarchical clustering

Abstract:

Traditional clustering methods, whether hierarchical or non-hierarchical, are fundamental techniques in multivariate statistics that are widely used in data analysis. Nonetheless, these approaches have drawn significant criticism, particularly regarding their limitations when applied to large or complex datasets. Hierarchical clustering methods, though intuitive, lack a clear objective function and are inflexible, potentially leading to suboptimal solutions. Non-hierarchical methods are typically computationally efficient, but highly sensitive to initial conditions and frequently require pre-specification of the number of clusters, which is often unknown. To try to overcome some of these limitations, a novel hybrid method that combines hierarchical and non-hierarchical approaches is proposed. It is named Hybrid Single linkage, briefly HSL, because the methodology resembles the single linkage hierarchical clustering, trying to solve the well-known limitation of the chaining effect. The HSL methodology begins by partitioning the data using a new non-hierarchical algorithm that maximizes the between-cluster distance, based on the single linkage criterion, which defines the distance between clusters as the minimum distance between points in different groups. This step results in an initial partition that tries to identify optimally separated clusters. The method then constructs a hierarchical structure, continually maximizing the inter-cluster distance at each level to ensure consistency in the objective function throughout the process. A key advantage of HSL is the ability to identify an optimal partition and a parsimonious dendrogram around this partition, optimizing an objective function. This flexibility makes HSL adaptive to different data analyses. Simulated data sets together with benchmarks are used to show the features of HSL.

Density-Based Community Detection Combining Structure And Attribute Information

Authors:

Sara Geremia[1*†], Domenico De Stefano[1]

[1] University of Trieste

[*] Corresponding author [†] Presenter

Contact: sara.geremia@phd.units.it

Keywords:

Community detection, Attributed networks, Leader influence, Homophily

Abstract:

In social networks, community structure arises from a complex interplay between structural features and actor attributes. This study aims to enhance density-based community detection by integrating both topological and attribute information within a unified framework. Evaluations on simulated and real-world datasets demonstrate that incorporating attribute information improves detection performance in networks with moderate to high mixing. Results suggest that the optimal combination of topological and attribute information depends on the network structure and the type of attribute data, with a balanced approach often yielding the best results. The study also highlights the importance of choosing the topology similarity measure, with the neighbor set similarity approach proving to be more robust.

Session - Bayesian learning for structured and functional data

Local Conformal Prediction For Non-Parametric Uncertainty Bands In Functional Ordinary Kriging

Authors:

Anna De Magistris[1*†], Elvira Romano[1], Gerardo Toraldo[1]

[1] dipartimento di Matematica e Fisica Luigi Vanvitelli

[*] Corresponding author [†] Presenter

Contact: anna.demagistris@unicampania.it

Keywords:

local conformal prediction, functional ordinary kriging, spatio-functional data

Abstract:

The prediction of spatial functional data, functions observed at spatial locations, represents a key aspect in many fields, including ecology, medicine and geosciences. A major challenge lies in providing reliable uncertainty quantification for predictions at unobserved sites. Traditional methods, such as Functional Ordinary Kriging (FOK) and Functional Universal Kriging (FUK), offer accurate spatial predictions; however, they often depend on strong assumptions and require computationally intensive resampling to construct prediction bands. Assessing uncertainty in Functional Ordinary Kriging (FOK) involves evaluating the variability in the predicted functions at unsampled locations. The process can include analyzing prediction variance, constructing confidence bands, using cross-validation, and applying simulation methods. Techniques such as resampling methods have been introduced to estimate the uncertainty in predicted curves, allowing for confidence bands for functional predictions. However, these are computationally intensive, especially for large datasets, due to the need for multiple resampling iterations. In addition, they may lead to biased estimates if the sample size is small or not representative of the population. In this work, we propose a Local Spatial Conformal Prediction (LSCP) method, which constructs prediction bands with finite-sample coverage guarantees without requiring strict distributional assumptions on residuals. LSCP adapts to spatial heterogeneity by selecting an adaptive neighborhood around each prediction site and modulates uncertainty along the functional domain using a spatial varibility measure. Unlike previous conformal approaches based on fixed spatial kernels, our method allows local adaptation in both space and function shape, enhancing flexibility and interpretability. Extensive simulations and a real-world application to the prediction of vegetation cycles in *Fire Rings* demonstrate that LSCP achieves accurate coverage and competitive band widths compared to existing kriging-based techniques. Our findings indicate that LSCP is a robust and computationally efficient alternative for uncertainty quantification in spatial functional data analysis.

Modelling Longitudinal Health-Related Constructs: A Latent Variable Approach

Authors:

Niccolò Cao[1*†], Silvia Cagnone[1], Livio Finos[2]
Marilina Amabile[3]

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna
[2] Department of Statistical Sciences, University of Padova
[3] IRCCS Istituto Ortopedico Rizzoli

[*] Corresponding author [†] Presenter

Contact: niccolo.cao@unibo.it

Keywords:

Underlying Variable Approach, Ordinal data, Full information maximum likelihood, Limited information estimation methods

Abstract:

Questionnaires are increasingly used tools in clinical and medical research to collect self-reported data on patients' perceived health outcomes. In particular, Patient-Reported Outcome Measures (PROMs) have been specifically designed to assess health-related constructs - such as health status and treatment effects - based on patients' responses to ordinal items. When questionnaires are administered at multiple time points, latent variable models (LVMs) offer a variety of methods to gain insights from the data. For instance, first-order latent growth models are commonly employed to analyse changes across measurement occasions, treating each item separately. Alternatively, longitudinal confirmatory factor analysis (CFA) models offer a powerful framework for studying the measurement properties of a set of items over time, such as assessing whether the items consistently measure the same latent construct across time (i.e., measurement invariance). Within this framework, a common approach to dealing with ordinal items is the Underlying Variable Approach (UVA), which assumes continuous latent variables underlying the observed categorical responses. A longitudinal CFA model for ordinal data, under the UVA, incorporates the following structures: (i) an auxiliary model, linking the ordinal response to an underlying continuous response; and (ii) a measurement model, relating the underlying continuous responses to latent factors at different occasions. CFA can be extended by including a structural model to account for the temporal dynamics of the latent variables, for example through a second-order latent growth model with random effects. A standard approach for fitting LVMs with categorical items is full-information maximum likelihood (FIML), which is theoretically robust and effectively handles data affected by different types of missing data mechanisms. However, FIML can be computationally unfeasible, as it requires the evaluation of high-dimensional integrals, especially in models with many items and/or many latent variables. In contrast, limited information methods, such as diagonally weighted least squares (DWLS), rely on first- and second-order sample statistics and represent a computationally faster solution. While limited information methods can handle missing completely at random data, they need to be integrated with additional methodologies (e.g., multiple imputation or estimator corrections) to appropriately address data that are missing at random. Methodological advancements involving longitudinal latent variable modelling will be presented, along with an example based on longitudinal PROMs data.

Model-Based Clustering Of Functional Data Via Random Projection Ensembles

Authors:

Matteo Mori[1*†], Laura Anderlucci[1]

[1] University of Bologna

[*] Corresponding author [†] Presenter

Contact: matteo.mori8@unibo.it

Keywords:

clustering, functional data, ensemble, random projections

Abstract:

Clustering functional data poses significant challenges due to its inherently infinite-dimensional nature. A common approach to mitigate this issue is to represent functions within a finite-dimensional space through basis function decomposition, where each curve is expressed as a linear combination of basis functions. However, achieving an accurate representation often results in high-dimensional embeddings, making traditional clustering methods less effective. In this work, we propose a dimensionality reduction technique based on Random Projections (RPs) applied to the basis coefficients. RPs randomly map the original high-dimensional data onto a lower-dimensional subspace using a projection matrix. This approach is theoretically supported by the Johnson-Lindenstrauss Lemma, which ensures that pairwise distances between observations are approximately preserved under random projection. A major limitation of RPs in clustering contexts is their variability-different projections may or may not highlight the underlying cluster structure. To address this, we adopt an ensemble approach, combining results from multiple projections to improve both stability and clustering accuracy. Specifically, we apply Gaussian Mixture Models (GMMs) to the projected basis coefficients. Among the various projections, those showing better cluster separation-measured using the Kullback-Leibler divergence-are selected and aggregated. An ensemble approach is then employed to integrate these multiple clustering results into a final partition. The proposed method involves the tuning of three main hyperparameters: the dimensionality of the projected space, the number of random projections, and the ensemble size. Simulation experiments were performed to guide the selection of optimal parameter values. The method's performance is assessed using synthetic data and a real-world application.

Adaptive Density Estimation With Application To Image Segmentation

Authors:

Raul Zanatta[1*†], Giovanna Menardi[1]

[1] Dipartimento di Scienze Statistiche, Università degli Studi di Padova

[*] Corresponding author [†] Presenter

Contact: raul.zanatta00@gmail.com

Keywords:

adaptive smoothing, modal clustering, image segmentation

Abstract:

Image segmentation is the task of partitioning a digital image into meaningful regions based on visual characteristics such as intensity or texture. In grayscale images-the focus of this work-this involves grouping pixels according to their light intensity. Among clustering techniques, nonparametric methods are particularly well-suited to segmentation, as they can detect clusters of arbitrary shape and do not require the number of segments to be specified in advance. One of the most widely used approaches in this class is the mean shift algorithm, a gradient ascent procedure based on kernel density estimation, which has gained popularity in image segmentation tasks due to its ability to identify modes in the intensity distribution. In this work, we propose the use of a nonparametric clustering method we originally developed for the joint task of adaptive density estimation and mode detection, to the context of grayscale image segmentation. Unlike traditional segmentation methods based on simplified probabilistic models, our approach leverages adaptive smoothing to better accommodate images with strong local heterogeneity-both in the distribution of intensity values and in the relative sizes of image regions. Specifically, it can capture large, homogeneous areas as well as smaller, more nuanced structures, without assuming prior knowledge of their number or shape. This makes it particularly effective in real-world images, where grayscale intensity may vary substantially: some segments may consist of nearly uniform tones, while others may include gradual transitions or textural variation. The method relies on an EM-style iterative algorithm, where each iteration includes two maximization steps: one to update the assignment of pixels to regions (mode estimation), and the other to refine the local smoothing parameters in a fully data-driven way. This dual maximization structure enhances the flexibility of the procedure, allowing it to adapt simultaneously to the geometry and the scale of the underlying image structure.

Session - Dimensionality reduction and latent structures in high-dimensional data

Comparison Of Group Lasso Methods For Finite Mixtures Of Linear Regression Models

Authors:

Ana Moreira[1*†], Susana Faria[1]

[1] Departamento de Matemática, CMAT, Universidade do Minho

[*] Corresponding author [†] Presenter

Contact: anaafmoreira98@hotmail.com

Keywords:

Group lasso, Mixtures of linear regression models, Penalized maximum likelihood estimation, Simulation study

Abstract:

Finite Mixture Regression (FMR) models provide a flexible and powerful tool for analysing data that arise from heterogeneous populations, where the relationship between the dependent variable and explanatory variables may differ across latent subpopulations. In practical applications, these models often involve a large number of explanatory variables, making variable selection a critical aspect of model building. However, traditional variable selection methods, such as all-subset selection methods, are computationally intensive. To address this, various penalty-based methods have been developed. Among these, grouped variable selection methods are particularly attractive when dealing with categorical data, as they preserve the inherent group structure of the predictors. In this study, we investigate the problem of variable selection in mixtures of linear regression models in low and high-dimensional settings. Specifically, we compare the performance of three penalization methods: the Least Absolute Shrinkage and Selection Operator (LASSO), the Group LASSO and the Group Adaptive LASSO. An extensive simulation study is conducted to evaluate how different data structures and settings influence the performance of these methods in selecting relevant explanatory variables. In particular, we explore scenarios where the number of groups increases with the sample size, as well as scenarios where the number of groups exceeds the sample size. Overall, the Group Adaptive LASSO consistently outperforms the other approaches across various scenarios. Based on these findings, we strongly recommend the use of Group Adaptive LASSO for variable selection in low- and high-dimensional finite mixture regression models - an expected result, given that Group Adaptive LASSO enjoys the oracle property.

Odk-Means: A Simultaneous Approach To Clustering And Outliers Detection

Authors:

Tiziano Iannaccio[1*†], Maurizio Vichi[1]

[1] Sapienza Università di Roma

[*] Corresponding author [†] Presenter

Contact: tiziano.iannaccio@uniroma1.it

Keywords:

Outliers, Clustering, Pseudo-isolation

Abstract:

The K-means algorithm is widely used for clustering due to its simplicity and efficiency. However, its performance can be significantly degraded by the presence of outliers, which are common in real-world datasets. These outliers can distort the resulting centroids, leading to suboptimal clustering outcomes. To address this issue, many researchers have attempted to mitigate or completely eliminate the influence of such anomalies, considering some possible scenarios in which these atypical observations occur. Furthermore, current methods only identify items far away from their cluster's centroid, thus failing to recognize internal outliers. Internal outliers are data points that, while closer than average to the centroid or central tendency of the data, are still anomalous because they do not conform to the expected distribution or patterns of the majority of the data. Such units might lie within the main cluster but exhibit unique properties or values that make them different from typical points. Although internal outliers affect the precision of centroids estimates only marginally, they can corrupt training data in classification tasks. Since these outliers are close to their centroids, they might be treated as representative of the typical data distribution, thereby skewing decision boundaries and increasing the risk of false negatives. The need for an outlier detection technique that not only manages the exceptions mentioned above but can also be seamlessly integrated into most clustering and classification procedures leads to the proposal of ODK-means. This algorithm is based on the concept of pseudo-isolation (i.e., a measure of isolation that does not depend on the immediately nearest neighbors). This novel approach eliminates the need for preprocessing steps required by current techniques, making it superior in terms of overall efficiency. Furthermore, ODK-means applies a consistent detection criterion across all types of anomalies, including the previously mentioned exceptions. ODK-means enhances the robustness and accuracy of the clustering process and can be combined with dimensionality reduction techniques such as PCA and FA to create a versatile and comprehensive clustering framework. This integration not only improves the algorithm's ability to handle high-dimensional data but also enhances the interpretability and visualization of clustering results. Empirical evaluations on both real and simulated datasets demonstrate the effectiveness of this novel approach in delivering superior clustering outcomes in the presence of outliers while providing meaningful insights through dimensionality reduction.

Bayesian Multi-Study Biclustering

Authors:

Leonardo Genesin[1*†], Giovanni Parmigiani[2], Giovanna Menardi[1]

[1] University of Padua
[2] Department of Data Science, Dana Farber Cancer Institute

[*] Corresponding author [†] Presenter

Contact: leonardo.genesin@phd.unipd.it

Keywords:

Biclustering, Factor Analysis, Spike-and-Slab prior, Gibbs Sampling, Multi-Study Analysis

Abstract:

Latent structures in high-dimensional data can vary significantly in terms of sparsity, complexity, and scope. These structures may range from global patterns that appear consistently across all observations, such as those captured by traditional Factor Models, to more localized and heterogeneous patterns relevant only to specific subsets of variables and samples, as identified by methods like Sparse Factor Analysis or Biclustering. Identifying such latent patterns is a key challenge in many scientific domains, including genomics, where datasets are both large and complex. The growing availability of data from multiple related studies provides a powerful opportunity to improve the reliability of the discovery of such complex latent structures by leveraging shared information across studies. However, this also introduces additional heterogeneity due to study-specific effects, which can obscure the identification of replicable, biologically meaningful signals. Multi-study models, therefore, have been proposed to distinguish shared latent structures from study-specific ones. We propose a flexible Bayesian Multi-Study Spike-and-Slab model that jointly analyzes multiple datasets within a unified framework. Our model enables the discovery of latent structures that span a wide range of sparsity levels, from dense factors to highly sparse, with respect to both variables and samples, biclusters, while also accommodating both shared and study-specific patterns. By employing spike-and-slab priors, the model encourages adaptive sparsity and facilitates interpretable structure recovery. Parameter estimation is performed via an efficient Gibbs sampling algorithm, which enables scalable inference in high-dimensional settings. We demonstrate the utility of our approach through a simulation study and apply it to single-cell transcriptomic data, to uncover both consistent and context-dependent transcriptional structures.

Principal Covariate Regression With Nuclear Norm Penalty

Authors:

Kaiwen Liu[1*†], Mark de Rooij[1], Wouter Weeda[1]

[1] Leiden University

[*] Corresponding author [†] Presenter

Contact: k.liu.10@fsw.leidenuniv.nl

Keywords:

Principal covariate regression, Nuclear norm penalty, Multivariate dimension selection

Abstract:

Principal Covariate Regression (PcovR) provides a unified theoretical framework for simultaneously performing regression and dimensionality reduction, making it particularly attractive for behavioral researchers working with noisy high-dimensional datasets, such as brain-wide association (BWAS) studies. Ideally, a PcovR algorithm should both optimize the number of components and estimate coefficients under a certain penalty in one procedure. However, current PCovR methodologies cannot achieve the aforementioned objectives simultaneously. Current PCovR methodologies contain two distinct approaches, each with its own advantages and limitations. Originally, the PCovR method determines the optimal number of components by evaluating the relative gain in R2 . However, it is prone to overfitting due to the lack of penalty parameters. In contrast, penalized PCovR methods address the overfitting issue by introducing ridge and lasso penalties on coefficients. However, this approach requires an a priori determined number of components (e.g., based on scree plots), which introduces potential bias and inconsistency in model selection. In this study, we propose a novel extension of the PcovR framework by introducing PcovR with a nuclear norm penalty (PcovRnnp). PcovRnnp supports simultaneously optimizing the number of components while estimating penalized coefficients. By incorporating a nuclear norm penalty parameter (which penalizes the singular values of the coefficient matrix) into the loss function, PcovRnnp selects the number of components like a lasso penalty (some singular values are shrunk to zero) while estimating the coefficients like the ridge penalty (individual coefficients are shrunk but not to zero). PcovRnnp is particularly advantageous when researchers aim to predict outcomes from large and noisy predictor sets (e.g., BWAS studies). We will present the preliminary results of our pilot simulation study focusing on PcovRnnp's predictive performance and dimensionality selection. We will also include one PcovRnnp application on BWAS data.

Session - Methodological advances in applied statistical modeling

Linking Brain Function And Structure To Phenotypes: a Preliminary Work On The Assessment Of Replicability In The Human Connectome Project.

Authors:

Lisa Verbeij[1*†], Mark de Rooij[1], Wouter Weeda[1]

[1] Leiden University

[*] Corresponding author [†] Presenter

Contact: l.g.verbeij@fsw.leidenuniv.nl

Keywords:

BWAS, Replicability, Multivariate Regression, PCovR

Abstract:

Brain-Wide Association Studies (BWAS) are studies examining the associations between inter-individual variability in the human brain and phenotypic, often behavioural, traits. Initial BWAS studies investigated the association between one brain feature (e.g. voxel or region) and one phenotype (e.g. cognitive ability). It turned out that these studies were very difficult to replicate. Namely, research has shown that true brain-behaviour effect sizes are small and that many studies are underpowered. Hence, a small sample size (a median across studies of n = 25) would lead to inflated results and replication failure. Researchers have turned to multivariate analysis to integrate brain features into a single model to improve power and reliability. Capturing the distributed patterns, instead of isolated features, results in greater power for detecting true associations. Some research however has suggested that multivariate BWAS associations are, like univariate BWAS, in-sample inflated and finding replicable brain-behaviour associations requires thousands of individuals. Hence, it suggests that small studies are most vulnerable to sampling variability and inflated brain-behaviour associations. In contrast, other research has shown that in-sample inflation is eliminated by appropriate cross-validation and recommends the implementation of both internal and external validation to improve replicability. Subsequently, they suggest that also smaller sample sizes (n = 75 - 500) can still lead to replicable results. In this preliminary study we will extend this line of research by adopting and reproducing the results that suggests that multivariate BWAS can be replicable with moderate sample sizes. We will perform several analyses on empirical data from the Human Connectome Project. The aim of this study is to expand on the notion of replicability by assessing the stability of results in terms of uni- and multivariate effect size and parameter stability across cross-validation folds and sample sizes. Bootstrapped subsamples for various sample sizes are generated to assess sampling variability, predictive probability and effect size measures for various phenotypes. While the univariate BWAS effect size (linear bivariate |r|) is the correlation between an individual brain feature and one phenotype, the multivariate BWAS effect size (Multiple correlation coefficient R), is the correlation between observed and fitted y value. Assessing R and |r| across folds helps to detect whether the phenotype encoded either by a distributed pattern across the brain or by local information is highly variable. Moreover, the stability of partial (multivariate) regression coefficients across cross-validation training folds is evaluated. A problem that often arises in multivariate, high-dimensional, analysis is collinearity. These strong correlations between brain features can lead to 'bouncing beta's', indicating that the size of regression coefficients is relative

and highly variable with the sample (size). Subsequently, this results in instable and hence non-replicable parameter estimates. We assess two models that can deal with collinearity: ridge regression and principal covariate regression (PCovR). While ridge regression shrinks coefficients toward zero, thereby reducing variance and improving stability of estimates, PCovR effectively balances the transformation of the (collinear) variables into orthogonal principal components, while also optimizing for the best prediction.

Evolution Of Students' Profiles Enrolled In Italian Distance Learning Universities Over The Last Decade

Authors:

Marta Campagnoli[1], Sophia Chiara Fiora[2], Rebecca Ghio[1]
Marta Magnani[1*†], Silvia Salini[1], Stefano Trancossi[1]

[1] University of Milan
[2] University of MIlan

[*] Corresponding author [†] Presenter

Contact: marta.magnani@unimi.it

Keywords:

clustering, data reduction, mixed-type data, administrative data

Abstract:

The expansion of distance learning represents a structural shift in higher education, reflecting both the growing demand for flexible educational formats and changing student expectations. In the Italian context, this phenomenon has accelerated in recent years, particularly following the COVID-19 pandemic. This study investigates how the demographic, educational, and geographical profiles of students enrolled in fully online degree programs have evolved over time, using individual-level data from the Italian National Registry of University Students (ANS - Anagrafe Nazionale degli Studenti e dei Laureati) spanning from 2010/11 to 2023/24. To address this question, we propose a flexible methodological framework for studying the temporal evolution of latent groups in high-dimensional data. The framework follows a two-step protocol: first, it reduces data redundancy and dimensional complexity while preserving the structure of categorical variables and their interdependencies; second, it identifies latent student profiles at different time points using clustering techniques. The resulting clusters are meant to allow interpretability and comparability over time, enabling consistent tracking of group compositions and transitions. The analysis begins with a descriptive overview of enrollment trends in traditional and distance learning universities, highlighting the discontinuity introduced by the pandemic. To go beyond static comparisons, we apply clustering methods to characterise the structure of the student population in the pre- and post-COVID periods, with particular attention to the cohorts from 2021 onward, when distance learning had fully stabilised following the initial emergency response of 2020. Special attention is given to the visualisation of groups' evolution. Our case study highlights how the composition of the distance learner population has changed, shedding light on shifts in students' profiles among different segments. The framework is designed to be adaptable and extensible, supporting both methodological innovation and empirical analysis of complex administrative datasets in higher education. Acknowledges financial support within the 'Fund for Departments of Excellence academic funding' provided by the Ministero dell'Università e della Ricerca (MUR), established by Stability Law, namely 'Legge di Stabilità n.232/2016, 2017' - Project of the Department of Economics, Management, and Quantitative Methods, University of Milan.

Some Statistical Aspects In The Development Of New Crash Frequency Models For Vulnerable Users

Authors:

Sirine Ben Assi[1*†], Wiem Neji[1], Giuseppe Cappelli[2]
Sofia Nardoianni[3], Simona Balzano[3], Mauro D'Apuzzo[3]
Giovanni Camillo Porzio[3]

[1] university of carthage
[2] university of Cassino
[3] university of cassino

[*] Corresponding author [†] Presenter

Contact: sirine.benassi@ihec.ucar.tn

Keywords:

Pedestrian crashes, Urban safety, Crash frequency modeling, Poisson regression, Negative Binomial regression, Zero-Inflated Negative Binomial (ZINB), Over-dispersion, Zero inflation, Road safety, Urban infrastructure, Traffic analysis zones, Model comparison, Rome, Policy implications, Georeferenced accident data

Abstract:

This study presents the development and comparative evaluation of statistical models for predicting pedestrian crash frequency in urban areas, using accident data provided by the Roma Mobility Agency. Given the discrete and over-dispersed nature of crash data, three count data regression techniques were employed: Poisson, Negative Binomial (NB), and Zero-Inflated Negative Binomial (ZINB) models. The Poisson model serves as a baseline, assuming equidispersion between mean and variance, while the NB model addresses over-dispersion by introducing a dispersion parameter. The ZINB model further accounts for the excess zeros often present in crash datasets, representing locations with no reported incidents. Data pre-processing involved the integration of georeferenced accident records with urban infrastructure and demographic variables across traffic analysis zones in Rome. Explanatory variables included road network characteristics, traffic flow, and vehicle speed. Model performance was evaluated using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and goodness-of-fit tests. Results indicate that the NB model significantly outperforms the Poisson model, suggesting considerable over-dispersion in the data. Furthermore, the ZINB model provides the best fit overall, effectively capturing the high proportion of zero-crash zones. These findings underscore the necessity of selecting appropriate statistical frameworks to accurately model pedestrian crash occurrences in urban contexts. The study offers actionable insights for urban safety planning and resource allocation, emphasizing the importance of considering over-dispersion and zero inflation in crash frequency modeling. The developed models can aid policymakers in identifying high-risk areas and implementing targeted interventions to enhance pedestrian safety in cities like Rome.

Young Generation And Sustainable Mobility: Findings From a Pls Structural Equation Model

Authors:

Edoardo Pascucci[1*†], Simona Balzano[1], Houyem Demni[1]
Luisa Natale[1], Giovanni C. Porzio[1]

[1] UNIVERSITA' DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE

[*] Corresponding author [†] Presenter

Contact: edoardo.pascucci@unicas.it

Keywords:

Latent Variables, Structural Equation Models, Technology Acceptance Model (TAM)

Abstract:

Understanding the factors influencing the adoption of sustainable mobility tools among younger generations is essential for comprehending future urban transport options. This study, based on a large survey, investigates the drivers behind young people's inclination to adopt sustainable mobility solutions (i.e., use of bicycles, e-bikes, and electric scooters) in cities. Specifically, it examines the roles of risk perception and perceived quality of residential areas in adopting sustainable mobility tools, considering the impact of personal beliefs about sustainable practices. An adapted version of the Technology Acceptance Model (TAM) is used for the study aims, where the model relationships are estimated using a Partial Least Squares Structural Equation Model approach. Data comes from a survey conducted from March to June 2024, targeting people aged 18 to 35 who commute for work or study in the metropolitan area of Rome. The dataset consists of responses from 1,003 participants. The data were collected using a CASI methodology. Key findings indicate that perceived risk negatively affects the intention to use bicycles, e-bikes, or electric scooters: those who perceive greater risks are less inclined to use them. Conversely, individuals with strong environmental awareness and positive beliefs about sustainability are more likely to adopt these options. Furthermore, the perceived quality of one's residential area also influences the intention to use these tools. Additional analysis comparing respondent groups revealed that intent varies based on living location (Rome versus suburbs) and commuting distance (living and working/studying in the same city versus different cities). These results can inform the development of policies and interventions by stakeholders to promote sustainable mobility solutions tailored to the needs of specific territories and the preferences and concerns of younger generations. Policymakers and urban planners can leverage these findings as a starting point to implement strategies that mitigate perceived risks, enhance the attractiveness of sustainable mobility, and foster a transition to more environmentally friendly urban transport networks.

Session - Modeling dependence and structure in graphs, space, and circular data

A Mardia-Sutton Distribution For Cylinders With Random Ray

Authors:

Yahia Hammami[1*†], Giovanni C. Porzio[2], Houyem Demni[2]
Amor Messaoud[1]

[1] Université de Carthage, Ecole Polytechnique de Tunisie Laboratoire d'Économie et de Gestion industrielle (LEGI-EPT)
[2] University of Cassino and Southern Lazio, European University of Technology (EUt+) Cassino, Italy

[*] Corresponding author [†] Presenter

Contact: yahia.hammami@ept.ucar.tn

Keywords:

Circular-linear modeling., Cylindrical distribution., Correlation on the cylinder.

Abstract:

Cylindrical data with one angle and two linear measurements occur in application areas such as meteorology and environmental science, where direction and size are relevant. In this work, we extend the classic Mardia-Sutton model to handle two linear and one circular variable, while capturing the correlation structure between the components. The model employs a von Mises marginal on the angle and a conditioned bivariate normal on the linear pair (conditioned on the angle). The maximum likelihood estimators of the model parameters are obtained. This extension increases flexibility by not assuming a fixed unit radius, allowing the distance from the origin to vary with direction, making the model more suitable for a wider range of real-world cylindrical data.

Robust Distance Correlation

Authors:

Sarah Leyder[1*†], Jakob Raymaekers[1], Peter Rousseeuw[2]

[1] University of Antwerp
[2] KULeuven

[*] Corresponding author [†] Presenter

Contact: sarah.leyder@uantwerpen.be

Keywords:

Breakdown value, Dependence measures, Independence testing, Influence Function, Robust statistics

Abstract:

Distance correlation is a widely used method for measuring dependence between random variables because it captures all forms of dependence - not just linear or monotone relationships. Its straightforward definition and applicability across diverse fields, ranging from genetics to machine learning, have contributed to its popularity. While it is often assumed that distance correlation exhibits a certain degree of robustness to outliers, this important aspect has not yet been thoroughly investigated. In this talk, we delve deeper into the robustness properties of distance correlation, as well as related quantities such as distance covariance and distance variance. By deriving influence functions and breakdown values, we uncover surprising results regarding its sensitivity to contamination. Specifically, we demonstrate that although the influence function of the classical distance covariance is bounded, the breakdown value is zero. This means that the measure can be completely disrupted by even a single outlier, contrary to common assumptions about its robustness. Additionally, we find that the sensitivity function is unbounded but converges to the bounded influence function as the sample size grows, highlighting a nuanced robustness behavior. To address these limitations, we introduce a novel, more robust version of distance correlation. Our approach is based on a new data transformation, which we call the biloop transformation. This transformation is designed to mitigate the influence of outliers and enhance the method's focus on the fundamental patterns present in the data. Through extensive simulations, we demonstrate that the resulting robust distance correlation performs well in the presence of outliers, maintaining strong power to detect dependencies. We further illustrate the practical advantages of our robust method using a real genetic dataset. We illustrate the new method using the biloop transformation on genetic data. Comparing the classical distance correlation with its more robust version provides additional insight in the dependencies between variables in a data set.

Graph Embeddings Impact On Unsupervised Community Detection In Ring Of Cliques With Outlier Effects

Authors:

Ahmadali Jamali[1*†], Domenico De Stefano[1]

[1] University of Trieste, Piazzale Europa 1

[*] Corresponding author [†] Presenter

Contact: ahmadali.jamali@phd.units.it

Keywords:

ROC+, Community detection, Graph embedding, Node2vec

Abstract:

Unsupervised community detection in graph networks with outliers poses a significant challenge. Ring of Cliques with outliers (ROC+) is a synthetic network designed to incorporate topological anomalies, of- fering a novel benchmark simulation for network analysis. To deal with the community detection challenge in this synthetic network, this study explores combinations of graph embeddings, such as Node2vec, with dif- ferent clustering algorithm likewise Self-Organizing Map (SOM), DB- SCAN, and Affinity Propagation which techniques, we do not need the predefine number of groups. These combinations are evaluated on three variations of ROC+ networks, which vary in size, clique count, and out- lier configurations. By utilizing these synthetic networks, our research provides fresh insights into the efficacy of these combinations in man- aging topological anomalies as community detection in different varia- tions of ROC+ networks. This research demonstrates that Node2vec, as a shallow neural network without non-linear activation, combined with traditional clustering methods that do not require prior knowledge of the number of classes, effectively handles anomalies in different variations of ROC+ network configurations in both small- and large-scale networks. In addition, the results show that the community detection performance depends on the clustering algorithms combined with node embedding. Depending on the type of outlier network structure in ROC+, each clustering algorithm used for embedding clustering can have its own appli- cations and advantages.

Efficient Estimation Of Clustered Sdpd Models With Exogenous Components

Authors:

Sara Milito[1*†], Francesco Giordano[1], Maria Lucia Parrella[1]

[1] University of Salerno

[*] Corresponding author [†] Presenter

Contact: smilito@unisa.it

Keywords:

Spatial dynamic panel data models, Spatial clustering, Efficient estimation

Abstract:

The SDPD (Spatial Dynamic Panel Data) models have been proposed in the socio-econometric literature to analyze spatio-temporal data, with the aim of capturing multivariate dependence structures among spatial locations (i.e., spatial units) observed over time. Specifically, the SDPD model has several key components. First, it includes a "lag-0" spatial component that accounts for interactions among the contemporary values of neighboring spatial unit, allowing for the modelling of spatial relations. Second, a dynamic component captures serial dependence over time within each spatial unit. Third, the model incorporates a spatial-dynamic component, which reflects the interaction between space and time by considering the influence of neighboring unit values from the previous time period. Moreover, some exogenous components are introduced to model the effect of external covariates on the dependent variable. Finally, the model includes fixed effects to account for unobserved, time-invariant heterogeneity specific to each spatial unit. The error term of the model is assumed to be independently and identically distributed over time, with zero mean and constant variance, although it can be cross-correlated and heteroskedastic over space. Each of these elements may be influenced by location-specific parameters. In this work we consider a particular version of such models, called Clustered SDPD model, where the set of spatial units is assumed to be partitioned into clusters by relying on the parameters of the model. In particular, the spatial units are assumed to be homogeneous within clusters and heterogeneous across clusters, in the sense that they share the same parameters within clusters but have different parameters for different clusters. For this reason, the correct estimation (consistent but also efficient) of the parameters plays a central role in the correct identification of the unknown clusters. In this work we propose a new two-step procedure, which significantly improves the efficiency of the parameters estimators. In the first step, following the approach proposed in the literature for this class of models, individual parameters are estimated for each spatial unit by solving a system of Yule-Walker equations, obtaining consistent estimates of all the components. The second step, which represents the new contribution of this work, refines the estimation of the coefficients associated with the exogenous variables and the fixed effects. It follows by applying a new linear model between the error terms (defined by some parameters estimates from the first step) and the exogenous components. This new step allows us to achieve a more efficient estimation of the Clustered SDPD model parameters. The full set of estimated parameters, obtained with the proposed procedure, is then used as input for a Hierarchical Agglomerative Clustering algorithm, which identifies clusters of spatial units that share similar values in the parameters. A multiple testing procedure is employed to validate the resulting cluster configuration in order to select the optimal number of clusters. Simulation studies show the performance of the method in identifying both the number and composition of clusters under

various configurations of spatio-temporal dependence and heterogeneity.

Session - Latent variable models and dimensionality reduction methods for complex data II

Algebraic Laplacian Estimator To Select The Number Of Clusters In Spectral Clustering

Authors:

Cinzia Di Nuzzo[1*†], Matteo Farnè[2]

[1] University of Catania
[2] University of Bologna

[*] Corresponding author [†] Presenter

Contact: cinzia.dinuzzo@unict.it

Keywords:

Laplacian embedding, low-rank representation, nuclear norm, internal cluster validation, spectral clustering

Abstract:

Spectral clustering is a widely used technique for data partitioning, particularly effective in capturing non-linear and high-dimensional structures. However, one of its fundamental challenges remains the estimation of the number of clusters, which directly determines the dimensionality of the Laplacian embedding. This work introduces a novel and fully automatic method, called ALLE (ALgebraic Laplacian Estimator), to estimate the optimal number of clusters within the spectral clustering framework. The main idea of ALLE is to reformulate the cluster recovery task as a matrix decomposition problem, exploiting the fact that the Laplacian matrix of a well-separated clustering structure admits a low-rank plus sparse structure. Formally, ALLE estimates the Laplacian embedding by solving a penalised least squares problem, where the Laplacian matrix is decomposed into a low-rank component (representing the Laplacian embedding space) and a sparse residual (accounting for noise or perturbations). This is achieved by imposing a nuclear norm plus ℓ_1-norm regularisation, leveraging the convex relaxations of the non-convex rank and sparsity constraints. The low-rank component is recovered via Singular Value Thresholding, while the sparse part is identified through Soft Thresholding, in a proximal gradient algorithm that alternates the two steps. The regularisation parameters are adaptively selected using a minimisation criterion inspired by recent advances in high-dimensional covariance estimation. The novelty of the proposed approach lies in its algebraic formulation: by interpreting spectral clustering through the matrix decomposition, ALLE transforms the problem of selecting the number of clusters into the recovery of the latent rank of the Laplacian matrix. In contrast to existing methods that rely on heuristic eigengap analysis or internal validation indices, ALLE provides a data-driven solution, requiring only the similarity matrix as input. A simulation study confirms the effectiveness of ALLE in accurately recovering both the number of clusters and the embedding space under various scenarios, including different noise levels and overlapping clustering structures. These results support the potential of ALLE as a tool for unsupervised learning tasks where the number of clusters is unknown. Future developments will focus on theoretical guarantees for consistency and on extending the framework to dynamic and multiway data.

A Comparison Of Estimation Methods In Latent Variable Models For Binary Panel Data

Authors:

Lucia Guastadisegni[1*†], Silvia Bianconcini[1], Silvia Cagnone[1]

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna

[*] Corresponding author [†] Presenter

Contact: lucia.guastadisegni2@unibo.it

Keywords:

approximate likelihood methods, pairwise likelihood methods, high dimensional integrals

Abstract:

Longitudinal multivariate data are common in many research fields, such as biostatistics and psychology, and are often used to measure changes in outcomes over time or to identify the determinants of such changes in a set of individuals. To analyze such data, generalized linear latent variable models can be employed, as they can handle various types of responses and account for correlations both among different response variables at each time point and across multiple occasions, as well as the correlation of each response over time. In this work, we examine and compare various estimation methods for generalized linear latent variable models for multidimensional longitudinal binary data. In such cases, likelihood-based methods are problematic due to the high-dimensional integrals involved, which lack analytical solutions. Among the methods proposed in the literature to address this issue, we focus on approximate likelihood and composite likelihood methods. Specifically, within the first class, we examine the dimension-wise quadrature, which simplifies high-dimensional integrals by truncating the Taylor series expansion, providing precise approximations without the need for derivative computations. Within the second class, we consider the pairwise likelihood approach, which relies on marginal likelihoods related to pairs of observations, and its variant, the d-order pairwise likelihood, which relies on bivariate densities but selectively omits some of them from the likelihood. The pairwise likelihood methods are also implemented using the method of separate maximizations. In this approach, the composite likelihood is defined as the sum of all log pairwise likelihoods, with each pairwise likelihood maximized separately. The obtained parameter estimates are then combined to produce a single parameter estimate. The properties of the approximate likelihood and composite likelihood methods are evaluated through a comprehensive simulation study, along with a comparison of different weighting approaches used in the separate maximization methods.

Topic Homogeneity Test-Based Fuzzy Document Clustering

Authors:

Gian Mario Sangiovanni[1*†], Louisa Kontoghiorghes[2], Ana Colubi[3]
Maria Brigida Ferraro[1]

[1] Sapienza University of Rome
[2] King's Business School of London
[3] Faculty of Economics and Business Studies, Justus Liebig University of Giessen

[*] Corresponding author [†] Presenter

Contact: gianmario.sangiovanni@uniroma1.it

Keywords:

Bootstrap, Document Classification, Soft Clustering, Complex Data

Abstract:

A new fuzzy document clustering algorithm based on topic homogeneity is introduced. In detail, a novel dissimilarity measure is proposed, derived from the p-value of a hypothesis test that assesses the homogeneity of topic distributions between two documents. First, the topic distributions are derived through Latent Dirichlet Allocation, and then a bootstrap procedure is applied to obtain the p-value. Finally, the resulting dissimilarity matrix is integrated into the fuzzy relational clustering procedure. The performance of the proposal is evaluated using a benchmark dataset.

Mixture Of Experts Latent Trait Analyzers

Authors:

Dalila Failli[1], Maria Francesca Marino[2*†], Francesca Martella[3]

[1] University of Perugia
[2] University of Florence
[3] Sapienza, University of Rome

[*] Corresponding author [†] Presenter

Contact: mariafrancesca.marino@unifi.it

Keywords:

Model-based clustering, Finite mixtures, Concomitant variables, EM algorithm, Variational inference

Abstract:

Complex data are increasingly common across various fields due to advancements in technology that allow for the collection of large and diverse datasets. In this field, the use of advanced analytical methods allows to effectively extract insights from such data. A key aspect of complex data analysis often involves the identification of homogeneous groups (clusters) of units. The Mixture of Latent Trait Analyzers (MLTA) represents a model-based clustering approach specifically tailored to multivariate categorical (binary) data. It accommodates clustering of units through a finite mixture specification, as in a latent class framework. Further, the propensity of units in a given cluster toward specific values for the categorical outcomes is assumed to depend on a multi-dimensional continuous latent variable (trait), as in the latent trait framework. Therefore, compared to the latent class model, the MLTA also allows to capture the residual latent variability within each cluster, thus permitting to overcome problems related to the local independence assumption upon which the latent class model is based. In this paper, we extend the MLTA by allowing concomitant variables to influence cluster formation, the (conditional) outcome distribution, both models, or neither, as in standard mixtures of experts models. This approach offers several advantages. First, including covariates into the latent layer of the model allows to understand how the observed characteristics of units influence their clustering, better reflecting the underlying structures of the data. Second, allowing the distribution of response variables to also depend on covariates allows to capture the relationship between them. Overall, the inclusion of covariates in both the observed and latent layers of the model improves its flexibility and ability to reflect the complexity of the data.

Session - Advances in preference and perceptions statistical modeling

An Hybrid Preference Learning Framework To Refine The Consensus Ranking

Authors:

Maurizio Romano[1*†], Gianpaolo Zammarchi[1]

[1] University of Cagliari

[*] Corresponding author [†] Presenter

Contact: romano.maurizio@unica.it

Keywords:

Preference learning, Kemeny problem, tied rankings, heuristics, particle swarm optimization

Abstract:

The study of preference rankings, or preference learning, is becoming increasingly important in many scientific fields. Preferences are expressed when a group of judges (or raters) evaluate a collection of elements (or items), assigning an order to objects based on which ones are preferred over others. When there are many judges and also a large number of items to be evaluated, an aggregate measure is needed in order to solve the rank aggregation problem, providing an interpretable comparison of the ranked items and assessing the overall level of agreement among judges. The rank aggregation problem is an NP-hard problem because it becomes more difficult as the number of items increases significantly. Approaches such as the branch-and-bound can be applied to problems with a limited number of items (i.e. fewer than 200). When the number of items grows, heuristic techniques have been developed to provide approximate solutions. Many of these heuristics are based on Kemeny's axiomatic approach, which has shown to be valid with tied rankings. In this paper, we propose a framework that aims at providing more than just a choice between "this slow but extremely accurate algorithm" and "this just good and faster one". Thus, following a hybrid approach, the proposal permits a trade-off between the two options. A simulation study shows the performance of the proposed framework in a controlled environment. Furthermore, a real world data set with a large number of items is considered. As a result, the proposal provides significant improvements in the solution found with a reasonable additional amount of computational time. This improvement is mostly investigated while using the recently proposed PSOPR algorithm (as the faster one) and the state of the art QUICK (as the slowest one).

The Relevance Of Information In Changing The Structure Of Consumer Preferences: A Pre-Post Sensory Experiment On Seven Olive Oils

Authors:

Marco Cardillo[1*†], Alfonso Piscitelli[1], Raffaele Sacchi[1]

[1] Department of Agricultural Sciences, University of Naples Federico II, Naples, Italy

[*] Corresponding author [†] Presenter

Contact: marco.cardillo@unina.it

Keywords:

Consensus Ranking, Paired sample, consumer preferences, Sensory Experiment

Abstract:

This paper presents the design and methodological framework of a pre-post sensory evaluation experiment conducted to explore whether an educational intervention may affect the sensory perceptions and preference structure of consumers evaluating olive oil. The primary objective was to assess whether a short seminar, providing structured information about olive oil classification and quality, could influence the way individuals perceive and rank different olive oil samples. The experiment was designed to control for potential spontaneous maturation or test-retest effects that might arise from repeated exposure to the same oils in two separate tasting sessions. The intervention consisted of a one-hour seminar held between two sensory sessions. During the seminar, participants were introduced to key concepts related to the commercial classification of olive oils, olfactory and flavor attributes, common sensory defects, aftertaste profiles, and the markers of high-quality oils, with particular emphasis on extra virgin olive oil (EVOO). The first sensory session was conducted prior to the seminar and included a blind tasting of several olive oil samples. One week after the seminar, a second sensory session was conducted with the same panel of 98 participants who had completed the first session and attended the seminar. To prepare the data for analysis, the raw sensory scores assigned to each oil by participants were converted into ranks. These rank-transformed values were then used to construct representative rankings for both the pre- and post-seminar sessions. This approach allowed for the comparison of preference structures across the two sessions, supporting an evaluation of changes in the general perception of oil quality following the educational intervention.

Gender Stereotypes And Barriers In Stem: a Bayesian Statistical Analysis Of Perceptions And Challenges

Authors:

Rossella Duraccio[1], Maria Iannario[1], Claudia Tarantola[2*†]
Roberta Varriale[3]

[1] University of Naples Federico II
[2] University of Milan
[3] Sapienza University

[*] Corresponding author [†] Presenter

Contact: claudia.tarantola@unimi.it

Keywords:

Bayesian Analysis, Gender Parity, Marginal effect, Multilevel Logit Model, Social Challenges, STEM

Abstract:

This study investigates the persistent gender disparities in STEM (Science, Technology, Engineering, and Mathematics) fields through the lens of advanced statistical modeling within a Bayesian framework. Despite growing awareness and numerous initiatives to promote gender equity in STEM, women remain significantly underrepresented across most scientific and technological disciplines. This underrepresentation is not merely the outcome of individual preferences, but reflects the cumulative effect of structural barriers, social expectations, and internalized stereotypes. The aim of this work is to provide a comprehensive understanding of the perceptions, biases, and contextual factors that contribute to this imbalance by analyzing data collected through a national survey conducted in Italy between November and December 2021. The survey, administered to over 1,000 participants, explored individual perceptions of STEM subjects and careers, beliefs about gender roles, and attitudes toward people working in STEM. The questionnaire was disseminated through a snowball sampling strategy initiated via university student associations, ensuring broad demographic coverage across regions, generations, and educational backgrounds. Key items focused on internalized gender stereotypes, self-assessed scientific competence, the perceived role of teachers, and structural factors such as parental education and regional female employment rates. By employing supervised learning techniques for predictive analysis based on a Bayesian multilevel ordinal regression model, we aim to enhance the understanding of the barriers women face in STEM. Additionally, we estimate the marginal effects of key predictors to quantify the impact of gender, age, education, and workplace environment. This approach contributes to methodological innovation while offering data-driven insights into persistent social inequalities in STEM.

A Joint Investigation Of Model-Based Classification Trees And Composite Indicator For Multivariate Ordinal Responses

Authors:

Rosaria Simone[1*†], Francesca Di Iorio[1], Carmela Cappelli[1]
Stefania Capecchi[1]

[1] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: rosaria.simone@unina.it

Keywords:

Model-based classification tree, Model-based synthetic indicators, Mixture models with uncertainty, Rating data

Abstract:

Social science surveys typically investigate multiple-dimensional concepts with ordinally scaled items, along with subject-specific features related to individual socio-economic and demographic characteristics and concomitant variables. Assuming the rationale of CUB models for the analysis of univariate rating distributions, we propose a joint investigation of recent developments in the setting of model-based classification trees for univariate response distributions and model-based composite indicators for the analysis of multivariate rating data. By model-based tree we refer to a binary partitioning algorithm assuming a maintained model to be fixed, and searching iteratively for the candidate variable splitting the observations classified in each node in two sub-samples on the basis of a given optimization criterion. This can be driven by maximization of the deviance in log-likelihood of the maintained model between father and descendant nodes, or by the maximization of the dissimilarity between children nodes. Model-based composite indicators refer, instead, to the procedure of identifying a unique (univariate) probability model (within a given class) that can be assumed as a fair representative of the rating distributions building up the data matrix. This can be derived either by resorting to a weighted average of the parameters characterizing the best fitting models estimated for each survey item (with weights possibly accounting for the association structure) or by estimating the statistical model that is the least dissimilar with all the items in the matrix. Given these premises, the proposal aims at discussing preliminary results on the joint implementation of a multivariate model-based classification tree for rating data, whose splitting criterion is built upon model-based synthetic indicators. In particular, at each step the procedure allows to identify the binary partitioning variable that allows to split the current data matrix into two sub-matrices for which the model-based composite indicators, assumed as representatives, are the most dissimilar among each other. The proposal derives groups of individuals whose responses to survey questions obey to a similar data generating process, as implied by the chosen modelling framework. Preliminary results are discussed on the wake of the fourth European Quality of Life Survey 2016.

Session - Innovative approaches in machine learning and clustering

Fast Weighted Linear Model Trees

Authors:

Flor Debois[1*†], Jakob Raymaekers[1], Thomas Servotte[1]
Tim Verdonck[1]

[1] University of Antwerp

[*] Corresponding author [†] Presenter

Contact: Flor.Debois@uantwerpen.be

Keywords:

Linear Model Trees, Weighted Regression, Imbalanced Regression, Boosting

Abstract:

Decision trees are powerful predictive learners widely used across various machine learning applications. However, simple decision trees like CART (Classification and Regression Trees) struggle to capture linear relationships effectively. Linear model trees address this limitation by fitting linear models in the leaf nodes, combining the structure of decision trees with the predictive power of linear regression. This results in highly interpretable models. Many linear model tree algorithms like MARS or M5 exist, each implementing this idea in a different way. Recently, a new approach called PILOT (Piecewise Linear Organic Tree) was introduced. PILOT trees offer the computational efficiency of traditional decision trees while incorporating the flexibility of linear model trees. This paper proposes a weighted extension of the PILOT tree algorithm, allowing its full potential to be exploited in a broader range of applications. During training, each observation can be assigned a weight that reflects its importance in the loss function. Crucially, incorporating these weights does not change the computational complexity nor the space complexity of the PILOT algorithm. As a result, the wide applicability of the original PILOT algorithm is preserved. Two applications illustrate the benefits of fast weighted PILOT trees. First, they are integrated into a one-step boosting model which aims to fit two complementary PILOT trees in an informed way. The approach strikes a balance between a single PILOT tree and an entire (random) forest of PILOT trees in terms of interpretability and performance. Secondly, weighted PILOT is applied to imbalanced regression datasets. By increasing the weights of underrepresented data points, often the points of interest, the predictions for these regions significantly improve. Experiments on several imbalanced datasets confirm the success of this strategy.

Sketchedrf: A Random Forest Framework For Classification Under Dataset Shift

Authors:

Laura Anderlucci[1*†], Angela Montanari[1]

[1] University of Bologna

[*] Corresponding author [†] Presenter

Contact: laura.anderlucci@unibo.it

Keywords:

sketching, random forests, dataset shift

Abstract:

In supervised classification, a fundamental assumption is that training and test data are drawn from the same underlying distribution. However, in many real-world applications, this assumption is often violated due to dataset shift-a discrepancy in the distribution of features, feature combinations, or class boundaries between the training and testing sets. These shifts can significantly degrade the performance of predictive models, especially when traditional learning algorithms are employed without adaptation. One strategy to address this issue is to perturb or reweight the training data to better reflect the characteristics of the test data. In this work, we propose a novel approach to handle dataset shift by leveraging matrix sketching within the framework of Random Forests. Unlike the classical Random Forest method, which generates training subsets via bootstrapping, our method constructs multiple synthetic versions of the training data using linear combinations of existing observations within each class. This "sketching" process is performed separately for each class and allows the generation of synthetic data that maintains class structure while introducing controlled variability. Moreover, the size of each synthetic class can be adjusted, enabling the balancing of class distributions and reducing class imbalance. The resulting Sketched Random Forests differ from traditional Random Forests in that no training sample is left out-of-bag, and each tree is built on a distinct, sketched dataset. However, we also show that independent validation data resembling the training distribution can be generated for evaluating model performance and variable importance. Empirical evaluations on real-world datasets demonstrate that the proposed approach significantly improves predictive accuracy, particularly in scenarios where the test set exhibits shifts in variance or structure compared to the training data. Here, the "sketchedRF" R package is presented. This research received funding by MUR-PNRR M4C2I1.3 PE6 project PE00000019 HEAL ITALIA, CUP J33C22002920006.

From Prediction To Explanation: Interpreting Risk Factors In Health Survey Analytics

Authors:

Agostino Gnasso[1], Massimo Aria[1*†], Roberta Siciliano[2]

[1] Department of Economics and Statistics, University of Naples Federico II, Italy
[2] Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: massimo.aria@unina.it

Keywords:

E2Tree, Machine Learning, Classification, Depression, EHIS

Abstract:

Machine Learning methods have gained significant attention for their ability to deliver high predictive accuracy and reveal complex, non-obvious patterns in data. Among these, Random Forest stands out as a popular ensemble technique, appreciated for its robustness and practical applicability, particularly in contexts where minimizing prediction errors is crucial. Nevertheless, the inherent complexity and lack of transparency in such models often limit their interpretability and can hinder user trust. In this study, we employ the Random Forest algorithm to analyze data from the European Health Interview Survey (EHIS), conducted by the Italian National Institute of Statistics (ISTAT), with the aim of identifying key risk factors associated with depression in Italy. To improve the interpretability of the model's output, we apply the Explainable Ensemble Trees approach, which allows for a deeper understanding of the internal decision-making mechanisms of Random Forest. The aim is to shed light on the determinants of mental health conditions, providing valuable insights to better understand the factors contributing to the risk of depression. Acknowledgments: this research has been financed by the following research projects: - PRIN 2022 SCIK- HEALTH (Project Code: 2022825Y5E - CUP: E53D23006110006); - PRIN 2022 PNRR The value of scientific production for patient care in Academic Health Science Centres (Project Code: P2022RF38Y - CUP: E53D23016650001).

Fuzzy Clustering Of Cylindrical Data: Some New Approaches

Authors:

Moatassam Bellah Alyani[1*†], Houyem Demni[2], Amor Messaoud[1]
Paolo Giordani[3], Giovanni C. Porzio[2]

[1] Université de Carthage, Ecole Polytechnique de Tunisie Laboratoire d'Économie et de Gestion industrielle (LEGI-EPT)
[2] University of Cassino and Southern Lazio, European University of Technology (EUt+), Cassino, Italy
[3] Sapienza University of Rome, Rome, Italy

[*] Corresponding author [†] Presenter

Contact: moatassambellahalyani@tbs.u-tunis.tn

Keywords:

Fuzzy Clustering Cylinders , Gower's Similarity , Cosine Dissimilarity , Tangent Space Projection

Abstract:

Due to its hybrid geometric structure, characterized by the integration of linear and angular components, cylindrical data pose significant challenges for their analysis. Particularly, conventional clustering methods for linear or spherical data may indeed fail to capture the intricate interplay between linear and circular variables. For this reason, some clustering techniques have been specifically designed to deal with this kind of complex data structure. Within this not-so-large literature, this work proposes and investigates some new approaches, all based on a fuzzy clustering perspective. The first is called Hybrid Fuzzy Partitioning Around Medoids (PAM) clustering. It encompasses a hybrid distance metric with customized weights, rescaled using distinct normalization techniques to align linear and circular components on comparable scales. The second uses a tangent space projection that linearizes angular data around the Fréchet mean, enabling compatibility with Euclidean-based clustering algorithms. Finally, a method based on the adaptation of Gower's similarity coefficient and the cosine dissimilarity is discussed. The proposed clustering algorithms are compared with some of the most popular traditional approaches (including k-means, PAM, and Fuzzy PAM), providing promising results.

Session - From stratified effects to latent trajectories: advances in statistical association

Testing For Constant Central Asymmetry Between Two Copulas

Authors:

Lorenzo Frattarolo[1*†]

[1] Department of Economics, University of Verona

[*] Corresponding author [†] Presenter

Contact: lorenzo.frattarolo@univr.it

Keywords:

Two-Sample Test, Dependence Asymmetry, Radial Symmetry, Reflection Symmetry

Abstract:

I develop a test of constant central asymmetry between two copulas estimated through their empirical counterpart. I provide inference under standard assumptions for stationary time series. The tie-break bootstrap is used for calculating p-values of the proposed Cramer-von Mises test statistic. Finite sample properties are assessed with Monte Carlo experiments. I apply the testing procedure to the US portfolio industry returns during the subprime crisis.

A Unified Approach To Inference On a Common Parameter Of Interest In Stratified 2×2 Tables

Authors:

Ruggero Bellio[1*†], Annamaria Guolo[2], Nicola Sartori[2]

[1] University of Udine
[2] University of Padova

[*] Corresponding author [†] Presenter

Contact: ruggero.bellio@uniud.it

Keywords:

Integrated Likelihood, Meta-Analysis, Risk Measure, Stratified binomial data

Abstract:

Stratified 2×2 tables are commonly encountered in several applied settings, including epidemiology, social sciences, and biostatistics. Notable instances within the latter field are meta-analysis studies and multicenter clinical trials. In broad generality, the data of interest consist in the outcome of K independent pairs of binomial data, given by $Y_{i1} \sim Bi(n_{i1}, p_{i1})$ and $Y_{i2} \sim Bi(n_{i2}, p_{i2})$, $i = 1, \ldots, K$, with Y_{i1} independent of Y_{i2}. Here, index 1 refers to the treatment group, and index 2 refers to the control group. In most applied settings, the starting point is the assumption of a constant measure of risk comparison between the two groups across the K strata. Three commonly adopted measures are the Odds Ratio (OR), $\psi = \{p_{i1}/(1-p_{i1})\}/\{p_{i2}/(1-p_{i2})\}$, the Risk Difference (RD), $\delta = p_{i1} - p_{i2}$, and the Risk Ratio (RR), $\gamma = p_{i1}/p_{i2}$. In settings with sparse data and potentially large number of strata, inference about the parameter of interest needs to account for the estimation of the strata-specific nuisance parameters. Under the assumption of a common OR, inference on ψ may be carried out using the conditional likelihood obtained by conditioning on $s_i = y_{i1} + y_{i2}$, which is a sufficient statistic for the nuisance parameters $\lambda_i = p_{i2}/(1-p_{i2})$. However, such a solution is unavailable for the RD and RR measures. Furthermore, boundary issues may occur when the individual estimated risks are 0 or 1, which is frequent in sparse data settings. This work proposes an integrated likelihood approach as a unified inferential approach covering all the three risk measures under the same methodological framework. In particular, the focus is on defining the integrated likelihood based on the zero-score-expectation parameterization for the strata-specific nuisance parameters, as proposed in the likelihood asymptotics literature. We describe the computation of the proposed solution for the various risk measures, showing that it provides accurate inferences, matching or even outperforming the best available methods existing in the literature. Moreover, the proposal is unaffected by boundary issues, without the need for continuity correction. Applications in meta-analysis studies illustrate the methodology.

Assessing Shape Heterogeneity In Regression And Smoothing Spline Models

Authors:

Marjolein Fokkema[1*†]

[1] Leiden University

[*] Corresponding author [†] Presenter

Contact: marjolein.fokkema@gmail.com

Keywords:

smoothing splines, regression splines, score based tests, decision trees

Abstract:

Non-linear associations are of key interest in many fields of study. Splines, or generalized additive models, provide state-of-the-art methods for modeling non-linear effects. For example, splines are widely used for modeling time series, dose-response curves, neuro-imaging data, and in geographical or climate studies. When spline models are fitted, an implicit assumption is made that the same shape of association fits all observations equally well. If there are subgroups with different shapes of association, these need to be known and specified a-priori. However, the one-shape-fits-all assumption may often be unrealistic. Moreover, researchers may often be specifically interested in discovering and explaining heterogeneity in the shapes of association. We propose the use of parameter stability tests to detect and explain shape heterogeneity in (unpenalized) regression splines as well as (penalized) smoothing splines. We discuss how the relevant (maximum likelihood) scores can be obtained from fitted regression and smoothing spline models. Specifically, we explore the use of fixed-effects model based scores as well as mixed-effects model based scores. We present results on the performance of different test statistics that can be computed from these scores, in testing parameter stability in simulated datasets. In an application to articulatory trajectories from experimental linguistics, we illustrate how the parameter stability tests can be used to discover subgroups that show different pronunciation patterns of English words. We present R package gamtree, that implements the proposed methodology in a user-friendly manner. Finally, we point out some of the computational challenges that arise, as well as avenues for future work.

Diverging Career Trajectories Of Men And Women In Japan: A Comparison Through Hidden Markov Models

Authors:

Miki Nakai[1*†], Fulvia Pennoni[2]

[1] Ritsumeikan University
[2] University of Milano-Bicocca

[*] Corresponding author [†] Presenter

Contact: mnakai@ss.ritsumei.ac.jp

Keywords:

employment sequences, gender inequality, hidden Markov models, typologies of career trajectories

Abstract:

Male-centered, Japanese-style employment practices, whereby women undertake the majority of housework and childcare, have contributed to substantial gender inequality in career advancement in Japan. Although women are increasingly participating in the labor market and pursuing professional careers, the gender employment gap remains significant. This study investigates how men's and women's careers diverge over time. We also examine how individuals transition among various types of occupational states, focusing on professional qualification throughout the life course. By analysing employment sequences we seek to understand how gender disparities accumulate over time. To do this we apply both Markov and hidden Markov models tailored for sequence data that also account for individual-level covariates. This analytical framework of model based-clustering allows us identify distinct typologies of qualification trajectories and to assess their association with age cohort and educational background. The data come from Social Stratification and Social Mobility Survey conducted in 2015 in Japan, including 2,885 men and 3,447 women aged between 20 and 69. We construct sequences from yearly occupational statuses, categorized into eleven distinct states: professional, various types of white-collar and blue-collar employment (differentiated by firm size and self-employment), farming, education, unemployment/inactivity, and missing responses. To identify clusters of similar life-course career trajectories, we estimate separate models for men and women to capture the gendered dynamics of intragenerational career mobility. The analysis identifies eight distinct types of career trajectories among women. The results highlight a variety of work-centered pathways, as well as a trajectory dominated by homemaking and caregiving roles. Women in self-employed white-collar occupations are more likely to experience stable career trajectories over the life course. On the other hand, the life events such as marriage and childbirth may trigger withdrawal from the labor market for women who are working as white-collar employees in both large and smaller firms. Among men, we identify ten distinct groups that signify typical career trajectories.

Session - Advanced statistical modeling in financial markets

Isp Index: A Parsimonious Method To Predict Defaults

Authors:

Fausto Corradin[1*†], Antonio Peruzzi[1], Roberto Casarin[1]

[1] Ca' Foscari University of Venice

[*] Corresponding author [†] Presenter

Contact: Fausto.Corradin@unive.it

Keywords:

Default, Probability of Default, Logit, Machine Learning, Overfitting

Abstract:

The necessity of determining the probability of default often conflicts with the requirement for employing parsimonious methodologies. The inclusion of a large number of regressors in Logit models or Machine Learning approaches can lead to overfitting, thereby introducing biases that distort the results. This study aims to examine the extent to which an indicator that synthesizes information related to balance sheet metrics can achieve a performance comparable to that obtained through a comprehensive set of indicators. In this paper, we introduce the Synthetic Performance Indicator (ISP), which is derived from specific balance sheet indicators. We demonstrate its effectiveness as a synthetic measure of financial stability. Furthermore, we assess its potential to serve as a viable alternative to the broader panel of indicators from which it is constructed. Finally, we provide further evidence of how Machine Learning approaches, despite being effective in-sample, perform poorly out-of-sample.

Random Dynamic Systems As a Modeling Tool In Statistical Arbitrage In The Stock Market

Authors:

Przemyslaw Jasko[1*†]

[1] Krakow University of Economics

[*] Corresponding author [†] Presenter

Contact: jaskop@uek.krakow.pl

Keywords:

statistical arbitrage, financial time series, financial econometrics, JTTW statistical arbitrage test, random dynamical systems, Breitung linear cointegration test, Aparicio-Escribano cointegration-in-information test, Escribano RCC nonlinear cointegration tests, Breitung rank test for monotonic cointegration, Bayesian TVP-VECM-SV with shrinkage priors and SAVS, postsparsification, Warsaw Stock Exchange, WIG20

Abstract:

A statistical arbitrage is a long-short, market neutral, quantitative trading strategy. We present its definition and the verification procedure employing formal JTTW (Jarrow, Teo, Tse, Warachka) test of statistical arbitrage, based on stochastic process of strategy value. The first aim is to mathematically establish structures of random dynamical systems and their generators (in a form of stochastic difference equations) representing movement of (log) prices of assets, which enable us to pursue statistical arbitrage strategy based on modeled dynamics of prices of the related stocks. The second aim is to empirically find multivariate stochastic processes of related stock (log) prices, that will form a random dynamical system, whose properties will allow us to pursue a statistical arbitrage strategy based on it. The first aim is realized on a ground of random dynamical models theory. From the normal form of a random dynamical system, we can extract the formulas for random invariant manifolds and random foliations, which can be used in forming of a statistical arbitrage strategy portfolios. As tools to find related processes (in the form of time constant cointegration) of asset (log) prices we present following statistical tests: Johansen's and Breitung's tests for linear cointegration; cointegration-in-information and Record Counting Cointegration (RCC) tests for nonlinear cointegration, and Breitung's rank test for monotonic cointegration. Noting that parameters of the models could change in time, we also consider time varying cointegration Bayesian models, namely TVP-VECM-SV with shrinkage priors and SAVS (Signal Adaptive Variable Selector) postsparsification (which enables to simultaneously establish which stock prices processes are cointegrated and is this cointegration constant or time varying). During empirical study, first we use exploratory analysis to state statistical hypotheses, tested in the subsequent confirmatory analysis, within a procedure of specification and verification of statistical models. The dataset for the empirical research encompasses 21 time series of closing prices of WIG20 and its 20 constituent stocks, of length T=643. Cointegration tests we use point out that the (log) prices process of the following assets could be related: ALIOR-SANPL, CCC-JSW, and DINOPL-PGE-PZU (among others). For (log) prices of the three stated subsets of assets, we build separate time varying parameter Bayesian dynamical models, namely TVP-VECM-SV with shrinkage priors (in two variants: normal gamma priors and ridge regression priors) and SAVS postsparsfication of cointegration matrix. Such structure of a model enables us to simultaneously test if cointegration is present,

and when this is true, is it time varying (with possible time subperiods in which cointegration disappears, which can be easily established using postsparsification procedure for cointegration matrix) or time constant. Conclusions from TVP-VECM-SV models we construct, is that for the pairs ALIOR-SANPL, CCC-JSW there is no cointegration during all the analyzed time period. This situation excludes for these two pairs, construction of statistical arbitrage strategy, based on their models of price dynamics. For the triplet DINOPL-PGE-PZU time varying cointegration was present for short subperiods of time: 4% of analyzed time period, for model with normal-gamma priors, and 18% of considered time period, for model with ridge regression priors.

Machine Learning For Credit Risk Modelling

Authors:

Steven Mphaya[1*†], Marialuisa Restaino[2]

[1] University School of Advanced Studies IUSS Pavia and University of Salerno
[2] University of Salerno

[*] Corresponding author [†] Presenter

Contact: smphaya@unisa.it

Keywords:

Credit Risk, Machine Learning, Deep Learning, Imbalanced Samples

Abstract:

Granting credit is the crucial activity of lending institutions, yet it faces an inherent risk in the form of customer default. Default risk requires close attention as it threatens firms' financial management activities. Thus, modelling credit risk is a pivotal component of the lending system, guiding lenders in evaluating borrowers' default probability and mitigating financial losses. For a long time, parametric models such as logistic and probit regressions, as well as discriminant analysis, have been the standard choice for estimating default likelihood. The increasing complexity and volume of financial data necessitate more proactive, scalable, and robust machine learning (ML) and deep learning (DL) models, which demonstrate high performance in various classification tasks due to their flexibility and ability to learn data with granularity. This study explores whether and which ML and DL algorithms outperform the most commonly used regression models in determining suitable models for credit risk assessment, based on performance and computational efficiency. Based on firm-level financial data, the study looks to uncover the trade-offs among these models, providing data-driven guidance to lenders and financial analysts on deploying methods that balance accuracy, efficiency, and interpretability in their default prediction workflows. Special attention is given to the challenge of class imbalance, a common issue in credit risk datasets, by examining the sensitivity of these models to this data characteristic using appropriate evaluation metrics. To address the "black-box" challenge of ML and DL models, we implement Shapley Additive Explanations (SHAP) to decompose model outputs and identify consistent predictors of default. By integrating appropriate data pre-processing techniques and interpretation tools, the study will provide insights into which financial features are more influential in credit-granting decisions, helping to bridge the gap between model predictive performance and practical deployment. The findings will contribute to the growing body of literature advocating for data-driven approaches in financial services, potentially making lending processes more accurate, fair, and efficient.

Weighted Estimation Of Hidden Markov Stochastic Volatility Models

Authors:

Michael Trequattrini[1*†], Silvia Pandolfi[1], Francesco Bartolucci[1]

[1] University of Perugia

[*] Corresponding author [†] Presenter

Contact: michael.trequattrini@unipg.it

Keywords:

Stochastic volatility model, Hidden Markov model, Time series forecasting

Abstract:

Forecasting realized volatility is a cornerstone of modern financial econometrics, un- derpinning a wide range of applications including risk management, asset allocation, and derivative pricing. Accurate short-term volatility forecasts are particularly valuable to practitioners seeking to dynamically adjust portfolio exposures or compute value-at-risk metrics. In this paper, we investigate the modeling and prediction of realized volatility using stochastic volatility (SV) models, with a specific focus on extending the classical framework to account for temporal heterogeneity in the information content of past ob- servations. We begin by adopting a class of stochastic volatility models introduced in the previous literature, in which log-volatility evolves according to an autoregressive latent process. In particular, we consider models where the conditional variance of returns is driven by an AR(1) latent component, discretized via a finite-state hidden Markov model (HMM) approximation. This approximation allows us to circumvent the computational complexity inherent in direct likelihood evaluation for SV models, while enabling effi- cient likelihood-based estimation and forecasting. As a novel contribution, we propose a weighted variant of the standard estimation approach, in which observation-specific weights are incorporated into the likelihood function to account for the diminishing rel- evance of older data. Specifically, we modify the log-likelihood by introducing exponen- tially decaying weights, placing greater emphasis on recent observations. This approach is motivated by the stylized fact that financial time series often exhibit regime changes and evolving dynamics, for which equal-weighted estimation may be suboptimal. The weighted log-likelihood is computed by applying exponentially decaying weights to the sequence of one-step-ahead predictive log-probabilities, allowing for efficient integration into the forward algorithm of the HMM. The decay parameter governing the weighting scheme is selected via cross-validation. Inference is performed via Newton-Raphson op- timization, leveraging a likelihood surface constructed from Gaussian quadrature-based approximations to the latent volatility process. To evaluate the predictive performance of the proposed models, we undertake an analy- sis comprising a simulation study based on synthetic data and an application to real-world financial datasets. Initially, synthetic datasets are generated under varying levels of la- tent volatility to rigorously assess the robustness of the estimation procedure and isolate the impact of temporal weighting. Subsequently, the models are applied to real financial data, beginning with the weekly forecasting of realized variance for the SandP 500 index. To ensure the generalizability and robustness of our findings, we extend the analysis to include additional market indices. Realized variance is computed as the sum of squared log-returns at 5-minute intervals, aggregated over each week. Forecasts are generated using a rolling window procedure, with re-estimation of the model at each step to reflect information available at the time of prediction. The empirical results demonstrate a con-

sistent and meaningful improvement of approximately 20% in median absolute error (MAD) for the weighted model compared to its unweighted counterpart on real financial data, underscoring the effectiveness of adaptive weighting in volatility forecasting.

Session - Statistical modeling and machine learning in genomic and population health research

Prioritization Of Differential Methylation Regions For The Prediction Of Coeliac Disease: a Machine Learning Approach

Authors:

Paolo Dalena[1*†], Luigina De Leo[2], Elena Spinelli[3]
Giulia Barbati[3], Adamo Pio D'Adamo[1]

[1] IRCCS Burlo Garofolo and University of Trieste, Trieste
[2] IRCCS Burlo Garofolo, Trieste
[3] University of Trieste, Trieste

[*] Corresponding author [†] Presenter

Contact: paolo.dalena@phd.units.it

Keywords:

Coeliac disease prediction, Random Forest for classification, Group Lasso, DNA methylation

Abstract:

In biochemistry, methylation is an epigenetic modification of DNA and is associated with repression of transcription. Different methylation patterns regulate the switching on and off of certain genes. When two groups of patients are analysed, it is possible to identify Differential Methylation Probes (DMP), i.e. points in the genome where there is a recurring significant difference in methylation between the two groups. From the DMPs, it is possible to identify Differential Methylation Regions (DMRs), i.e. entire DNA segments where significant differences are recorded. The objectives of this analysis are to identify DMPs and DMRs for coeliac disease, to introduce a prioritization of DMRs, to observe associations between DMRs and outcome in order to identify DMRs to be considered for the diagnosis of coeliac disease from the subjects' DNA. The initial dataset included 731239 probes plus 8 variables with clinical and demographic information on 148 patients. The values for the methylation of the probes were normalized and, through the application of the ChAMP pipeline, the relevant DMPs were identified and categorized into DMRs. The DMRs were ordered based on the average distances between the DMPs in each group. First, we fitted a Random Forest (RF) model to classify coeliacs and controls including all the DMPs and the clinical and demographic variables, which was considered to extract the variables importance. Secondly, we performed a grouped variable selection using the Group Lasso method (selecting the shrinkage coefficients considering a 10-folds cross-validation), forcing the selection of groups of DMPs corresponding to the identified DMRs and considering the clinical and demographic variables as single variable groups. We fitted a second RF model, including only the groups that resulted as non-null in the variable selection stage. Both the models were optimized considering the number of variables randomly sampled as candidates at each split, the minimum size of terminal nodes, and the number of trees to grow. They were evaluated through the accuracy, the Out-Of-Bag estimation for the error rate (OOBe) and the test set error rate (TSe). From the initial dataset, 1284 DMPs and 148 DMRs were identified. The model with all DMPs presented a OOBe of 17.86% and a TSe of 22.22%. Considering the resulting variable importance, a greater presence of the DMPs included in the first 24 DMRs (ordered according to the prioritization identified) was clear: 36.73%, 41.41% and 47.88% of

the first 50, 100 and 500 most important variables were included in the first 24 DMRs. The Group Lasso identified only 11 groups of variables (including 9 DMRs) and 70 coefficients as non-zero and the reduced model considering only these groups had better performances: 16.96% for the OOBe and 11.11% for TSe. DMRs are possible functional regions involved in gene transcriptional regulation, so a study of DMRs would allow a condition to be identified from the subjects' DNA. It was possible to obtain a prioritization for DMRs that resulted associated with the coeliac outcome. Furthermore, by considering DMRs in a group variable selection, a reduced model with very good prediction performance was identified.

Enhancing Statistical Inference In Mixed-Effect Three-Tree Model: A Data-Carving Estimation Strategy With An Application On Amyotrophic Lateral Sclerosis Data

Authors:

Giulia Vannucci[1*†], Roberta Siciliano[1], Valentina Iuzzolino[2]
Gianmaria Senerchia[2], Raffaele Dubbioso[2]

[1] Department of Electrical Engineering and Information Technology, University of Naples Federico II
[2] Department of Neuroscience, Reproductive Sciences and Odontostomatology, University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: giulia.vannucci@unina.it

Keywords:

mixed three-tree model, post-selection inference, data carving

Abstract:

In the framework of mixed-effects models, this paper explores the Three-Tree Mixed-Effects Model for longitudinal data. This model is a semi-parametric extension of the linear mixed-effects model, comprising a linear component and three tree-based components. This approach results in a model capable of handling interactions and nonlinearities while ensuring interpretability. Moreover, we propose an algorithm for estimating model parameters based on the data-carving post-selection inference procedure. The performance of the proposed algorithm is evaluated through a Monte Carlo study. The proposed methodology is applied to a real case study on Amyotrophic Lateral Sclerosis (ALS).

Innovative Applications Of Supervised Learning In Addressing Missing Data: A Case Study On Social Surveys

Authors:

Simona Cafieri[1*†], Francesco Pugliese[1], Mauro Sodani[1]

[1] ISTAT

[*] Corresponding author [†] Presenter

Contact: cafieri@istat.it

Keywords:

missing data, supervised learning, imputation

Abstract:

In today's interconnected and dynamic global landscape, addressing key challenges such as public health, environmental sustainability, and social inequality demands coordinated, data-informed policy responses. Addressing these issues requires coordinated and evidence-based policy interventions, for which national statistical offices (NSOs) play a crucial role by providing high-quality and timely data. However, the presence of missing or incomplete information-whether arising from surveys or administrative sources-represents a persistent threat to the validity and reliability of statistical outputs. This study investigates the potential of advanced imputation techniques based on supervised machine learning (ML) and deep learning (DL) approaches, which fall within the broader domain of Artificial Intelligence (AI) , to improve the handling of missing data in official statistics. A comparative analysis is conducted using real-world microdata from the Italian National Institute of Statistics (Istat), specifically from the Multipurpose Survey on Households, which is characterized by a non-negligible incidence of item nonresponse. Preliminary findings indicate that AI-driven imputation strategies outperform traditional statistical methods in terms of accuracy and robustness, particularly in complex social datasets. The results contribute to the growing body of literature advocating for the integration of modern computational tools within the framework of official statistics, with the aim of enhancing data quality on critical societal issues.

Off-Target Analysis Through Latent Class Models And Machine Learning In Crispr Cas9: Tumor Protein 53 Sequence Application

Authors:

Ali Mertcan Köse[1*†]

[1] Istanbul Ticaret University

[*] Corresponding author [†] Presenter

Contact: alimertcankose@gmail.com

Keywords:

Off-target prediction, CRISPR, latent class analysis, machine learning

Abstract:

Introduction: Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are an adaptive immune system in archaea and bacteria that target alien DNA, like plasmids and viruses (Kalamakis and Platt, 2023). Although the Cas9 system is widely used, off-target effects remain challenging (Listgarten et al., 2018). This study uses Latent Class Analysis (LCA) to classify off-target levels and applies Machine Learning (ML) to predict mismatch positions, offering an alternative to traditional methods like CFD and MIT. Methods: LCA was used to analyze the relationships between categorical variables measured in nominal and ordinal variables in a dataset of 947 matching the target sequence and gRNAs based on 23 bases. Firstly, the matching DNA and gRNA were coded as multiple (4×4). Afterwards, off-target regions were predicted using ML methods such as XGboost, Support Vector Machines (SVM), Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and decision trees with 10-fold cross-validation procedures on both training and test datasets. Results: In the LCA applied to the benchmark dataset, three significant classes were produced regarding the off-target scores. These classes were determined as high (26.7%), middle (30.9%), and low (42.3%). According to the ML results, which are based on the proposed approach, CNN achieved the highest accuracy for both training and test datasets in the multiclass classification test. Additionally, CNNs and XGboost yielded the highest Area Under the Curve (AUC). Conclusions: Consequently, the off-target effects were minimized, and the regions effective in gRNA design were identified. These findings provide valuable insights into the complex interactions between gRNA and genetic sequences. This work will benefit future studies aimed at optimizing genome editing in CRISPR Cas9.

Session - Advances in statistical optimization

A Note On Gradient-Based Parameter Estimation For Energy-Based Models

Authors:

Luca Martino[1], Salvatore Ingrassia[1], Sofia Mangano[1*†]
Luca Scaffidi Domianello[1]

[1] University of Catania

[*] Corresponding author [†] Presenter

Contact: sofia.mangano@phd.unict.it

Keywords:

Energy-based models, Gradient descent, Maximum likelihood

Abstract:

Energy-based models (EBMs) are an important family of models where a piece of the likelihood is intractable, and hence unknown. For this reason, the parameter estimation in EBMs is a challenge for the standard estimation methods. In this paper, we present a critical discussion of gradient-based approaches for inference in energy-based models. We provide many details of different derivations, clarify connections and differences. We give practical suggestions for the application of the different schemes. Specifically, we focus on a suitable choice of the proposal/reference density that is crucial for the performance of the gradient-based procedures.

Local Optimization-Based Clustering

Authors:

Luca Frigau[1*†]

[1] University of Cagliari

[*] Corresponding author [†] Presenter

Contact: frigau@unica.it

Keywords:

Local Optimization, Non-Convex Clusters, Iterative Reassignment, Deviation Minimization

Abstract:

We propose a novel clustering algorithm that leverages local information through a process of localized optimization to partition data into coherent and meaningful groups. In contrast to traditional clustering techniques, which typically depend on global criteria, such as distances to centroids (as in k-means) or global density estimates (as in DBSCAN), our method emphasizes the internal coherence of small subsets of observations. It guides the clustering process by minimizing localized variability measures, thus enabling a more flexible and adaptive approach to group discovery. The core idea of the algorithm is to begin with an initial random configuration of candidate clusters, typically obtained by assigning data points to groups of similar size without imposing any strong structural assumptions. Each cluster is then evaluated using a local deviation function, which quantifies intra-cluster variability based on the specific configuration of points within it. The algorithm proceeds through an iterative process in which individual observations are reassigned between groups based on the potential gain in local cohesion. Crucially, the algorithm does not rely on global centroids, nor does it impose fixed distance thresholds. As a result, it can naturally adapt to the local geometry, density, and distribution of the data in ways that conventional clustering methods cannot. By focusing on local consistency rather than global partitioning, the method proves particularly effective in handling non-convex, overlapping, or heterogeneously shaped clusters, which frequently occur in real-world applications such as spatial data analysis, image segmentation, or socio-economic profiling. The number of clusters is specified a priori, but the algorithm is robust to small mis-specifications, as it self-corrects by optimizing the internal configuration of each group. To assess the effectiveness of the proposed approach, we conduct an extensive empirical evaluation on both synthetic and real-world datasets, comparing its performance to established methods such as k-means, hierarchical clustering, and density-based techniques. The results consistently show that our algorithm offers good performance, particularly in the presence of irregular cluster boundaries, varying densities, and heterogeneous distributions. In addition, it demonstrates robustness to initialization, reduced sensitivity to outliers, and improved adaptability in high-dimensional settings, where traditional clustering models often underperform. Overall, this research contributes to the growing field of locality-aware clustering, highlighting how localized optimization can uncover complex structures that global approaches may overlook. Future developments will explore extensions to fuzzy clustering, incorporation of domain knowledge via constraints, and online learning scenarios, where clusters must be updated dynamically as new data arrives.

Selection Accuracy And Errors In Sparse Models With The Horseshoe Prior

Authors:

Andrea Mascaretti[1*†], Anna Vesely[2], Aldo Gardini[2]

[1] Scuola Internazionale Superiore di Studi Avanzati
[2] Università degli Studi di Bologna

[*] Corresponding author [†] Presenter

Contact: amascare@sissa.it

Keywords:

High-dimensional data, Bayesian variable selection, error rate control

Abstract:

Variable selection in high-dimensional settings is a fundamental goal in many fields, including genomics, neuroscience, psychology, and economics. Researchers often aim to identify relevant features, such as genes differentially expressed between biological conditions or predictive biomarkers that accurately forecast patient outcomes. Selection inherently involves making decisions, each carrying the risk of (i) falsely including non-informative variables and (ii) missing the relevant signal. Global-local shrinkage priors, and notably the Horseshoe prior, demonstrate strong empirical performance in signal detection and provide posterior uncertainty quantification by inducing sparsity on coefficients via a scale-mixture of Gaussians. However, their behavior under formal decision-theoretic frameworks is only partially understood, limiting their adoption when formal error guarantees are required. We perform a systematic analysis of frequentist selection errors for the Horseshoe prior in the high-dimensional normal means problem: a benchmark problem in many applied fields. Specifically, we characterize the family-wise error rate (FWER), the probability of making at least one false discovery, and the false discovery rate (FDR), the expected proportion of false discoveries among all selections, under different decision rules: posterior mean thresholding, credible interval-based selection, and median probability model selection. We examine how the global shrinkage parameter, which reflects the overall expected proportion of active variables, affects these error rates across different estimation approaches, including empirical Bayes methods and plug-in estimators based on sparsity-level estimation. We evaluate these properties across diverse settings, including sparse and dense signals, non-Gaussian noise, and correlated error structures. We assess how different choices for the global parameter and the selection rules impact errors and statistical power. Our results provide practical guidance for applying shrinkage priors in high-dimensional settings, enabling practitioners to anticipate the expected magnitude of selection errors under various configurations and select procedures that align with their tolerance for false discoveries. Furthermore, they suggest promising directions for extending these ideas to more complex models and broader classes of global-local shrinkage priors.

Optimizing Predictive Ability Of Binary Regression For Imbalanced Data: A Simulation Study On Asymmetric Link Functions And Feature Selection

Authors:

Marcella Niglio[1*†], Marialuisa Restaino[1], Rosaria Simone[2]

[1] Università degli Studi di Salerno
[2] Università degli Studi di Napoli Federico II

[*] Corresponding author [†] Presenter

Contact: mniglio@unisa.it

Keywords:

Imbalanced binary data, Binary Regression, Asymmetric link function, Prediction, Feature selection

Abstract:

Imbalanced binary data occurs frequently in medical and financial applications, and in all circumstances in which the outcome of interest constitutes a rare event (bank/client default, stroke occurrence, to quote a few). Their analysis requires suitable classification methods or prior data resampling techniques. In the setting of regression, for instance, asymmetric link functions should be preferred over the classical logistic or probit links to derive unbiased estimates of model parameters and reliable fitting and predictive performance. An instance in this respect is provided by Generalized Extreme Values and Generalized Logistic link functions. The proposal put forth by our contribution is a joint investigation of model selection and variable selection for high-dimensional regression of imbalanced binary data, challenging together recent results on regression with asymmetric link function and on feature selection for binary classification. To this aim, we discuss the implementation of a forward search algorithm for binary regression with asymmetric link functions that iteratively optimize a pre-specified indicator of classification performance, to yield a parsimonious model. Beyond the classical AUC, further performance measures specifically defined to account for the imbalance in the binary outcome can be considered, either in the optimization process or for an overall assessment of the procedure. As a result, our study aims at the identification of the link function which enables to build the best-performing model with respect to classification of new cases for given data, with the assessement of the contribution of each predictor to the fitting and prediction performance afforded by the model. Our preliminary investigation is presented and discussed on the basis of an extensive simulation study, with varying rates of imbalance in the response data, and assuming a comparative performance with classical logistic regression and variable selection strategies.

Session - From structural models to machine learning: predictive approaches across domains

Structural Equation Modeling For Out-Of-Sample Prediction: A Comparative Study Of Methods

Authors:

Hsin Kao[1*†], Li Zeng[1], Katrijn Van Deun[1]

[1] Tilburg University

[*] Corresponding author [†] Presenter

Contact: h.kao@tilburguniversity.edu

Keywords:

structural equation modeling, out-of-sample prediction, high-dimension low-sample size, machine learning

Abstract:

Structural Equation Modeling (SEM) is widely utilized in social science research for examining complex relationships among observed and unobserved (latent) variables. Traditionally, its application has centered on explanatory modeling, this is, testing theory-driven hypotheses, prioritizing model fit and parameter significance to explain underlying causal mechanisms. In social and behavioral science practice, however, prediction on new cases (e.g., whether a patient is at risk of committing a suicide attempt) is of utmost importance. Despite SEM's strengths in modeling latent constructs and accounting for measurement error, its predictive capabilities have received limited attention; few studies have systematically evaluated how well different SEM-based approaches predict outcomes when applied to new data, especially under high-dimensional, low-sample-size (HDLSS) conditions. Since explanatory power does not guarantee predictive power, this lack of predictive validation leaves a critical gap in our understanding of when and how SEM-based approaches can be effectively applied for prediction. This study addresses this gap by comparing seven approaches for out-of-sample prediction. We focus on this setting because it is common in psychological and social science research. Our comparison includes two-stage estimation methods, including Structural After Measurement (SAM), Sparse Generalized Canonical Correlation Analysis (SGCCA), and Regularized Exploratory Structural Equation Modeling (Regularized ESEM), as well as a one-stage estimation method, the SEM-Based Prediction Rule. Traditional approaches, such as sum scoring and covariance-based SEM (CB-SEM) are also included as these are used by many practitioners. A machine learning approach is also incorporated for comparison. We assess both parameter recovery (for measurement and structural models) and predictive performance across these seven approaches, using simulated data.

Relaxed Sem-Based Out-Of-Sample Predictions

Authors:

Aditi M. Bhangale[1*†], Mark de Rooij[1], Julian D. Karch[1]

[1] Leiden University

[*] Corresponding author [†] Presenter

Contact: a.m.bhangale@fsw.leidenuniv.nl

Keywords:

prediction, structural equation modelling, cross-validation, machine learning, regularisation

Abstract:

Predictive modelling-which applies model parameters from one data sample to generate predicted values for new observations beyond that sample-can play a critical role in psychological research. Until recently, prediction mechanisms were limited to the traditional linear regression and machine learning frameworks. However, these approaches assume that psychological variables are measured without error, which is often not the case. Structural equation models (SEMs) do consider measurement error, and a prediction rule for SEMs with normally distributed, continuous data was recently proposed. Although the SEM-based prediction rule outperforms predictions based on (regularised) linear regression models in most cases, it is sensitive to model misspecification-specifically when additional direct effects between indicators on the predictor side and the latent response variable are included in the data-generating SEM. Regularising the SEM-based prediction rule-using methods like ridge regression or regularised discriminant analysis-can help address this issue. In this study, we propose using regularisation to achieve a data-driven compromise between a restricted SEM and a linear regression model fit to the same data, thereby producing regularised SEM-based out-of-sample predictions. The combination of regularisation and SEM indirectly accounts for the degree of model misspecification to produce more accurate and precise predictions by weighting the influence of the linear regression and the SEM. We hypothesise that the regularised SEM-based prediction rule will perform at least as well as the SEM-based prediction rule when the model is misspecified.

Evaluation Of Supervised Machine Learning Methods In Predicting Biogeographical Ancestry Through An Innovative Snp Panel.

Authors:

Cosimo Grazzini[1*†], Giorgia Spera[1], Stefania Morelli[2]
Daniele Castellana[1], Giulia Cosenza[2], Michela Baccini[1]
Giulia Cereda[1], Elena Pilli[2]

[1] Department of Statistics, Computer Science, Applications "Giuseppe Parenti" (DiSIA)
[2] IRIS (Infrastruttura per la Ricerca e l'identificazione degli Scheletri senza nome), Department of Biology

[*] Corresponding author [†] Presenter

Contact: cosimo.grazzini@unifi.it

Keywords:

BioGeographical Ancestry, Single Nucleotide Polymorphism panel, Supervised Machine Learning, Forensic samples

Abstract:

In recent years, there has been an increasing interest in BioGeographical Ancestry (BGA) -the biological component of ethnicity- due to its significant ability to provide greater informativeness across several fields. Indeed, BGA plays a crucial role in population studies, medicine, epidemiology, anthropology, and forensic science, particularly regarding its strategic help in identifying remains from unknown individuals. An individual's BGA can be inferred from DNA, notably through panels comprising numerous Single-Nucleotide Polymorphism (SNP) markers, which poses challenges for traditional statistical approaches. Most studies using Machine Learning (ML) techniques have focused on BGA inference at the continental level. Still, only a few attempts have been made to go below this macro level, while maintaining a global scale perspective. Starting from individuals with known BGA, we aim to evaluate selected ML methods, coupled with an innovative SNP panel, in inferring BGA at the inter-continental and a more detailed level on a global scale. The applied ML methods include Categorical Naive Bayes, Penalized Multinomial Logistic Regression, Linear Support Vector Machines, Random Forests, and tree-based Gradient Boosting. Model selection and assessment were performed adopting a stratified nested cross-validation strategy. At the inter-continental level, the ML methods achieved high performance, suggesting the potential for broader adoption. However, performance decreased at the more detailed level, which is consistent with the literature. Additionally, a descriptive analysis of the inaccuracies highlights plausible misclassification patterns at both BGA levels, suggesting coherent geographical and historical patterns among populations. These findings lay the groundwork for further research aimed at selecting a highly informative SNP panel for finer BGA classification, ultimately defining a kit of genetic markers to use in real casework.

Clustering Using Multidimensional Indicators: An Approach Without Feature Reduction

Authors:

Alessia D'Ambrosio[1*†], Giuseppe Gismondi[1], Alfonso Piscitelli[1]
Leonardo Salvatore Alaimo[2]

[1] Università degli Studi Di Napoli Federico II
[2] Università di Roma La Sapienza

[*] Corresponding author [†] Presenter

Contact: alessia.dambrosio@unina.it

Keywords:

Latent Structure Contribution, Distance matrices, Clustering, Social Phenomena

Abstract:

The complexity of social phenomena can be decomposed into different dimensions, each measured by a set of basic indicators. These elementary measures capture distinct aspects of each dimension, summarizing them into a composite indicator. In this framework, one of the main goals of statistical analysis is to partition individuals by using these dimensional indicators directly rather than all available sets of basic indicators, possibly leading to a loss of information and to the introduction of biases into the data. In this work, we propose a new methodological approach that allows the clustering of statistical units using all the recorded elementary measures, while preserving information about the dimensional structure of the phenomenon under study when characterizing the clusters obtained from the partitioning algorithm. Our method relies on the use of distance matrices between individuals belonging to different groups, with elements that possess the desirable property of multivariate additivity, ensuring that the total distance between individuals is the sum of block-specific distances, where each block corresponds to a latent dimension. This property allows quantifying the contribution of each block to cluster formation, maintaining the original variability and ordinal nature of the data. The framework involves two steps: (i) clustering statistical units using all variables via distance-based algorithms, and (ii) quantifying block contributions through intra- and inter-cluster distances. The results of applying this method to an example dataset are compared with those obtained using traditional methods, demonstrating that our method not only achieves comparable clustering accuracy but also offers better interpretability of latent dimensions.

Session - Statistical methods and optimization for complex decision environments

Decoding Locomotor Intentions: Neural Networks And Probabilistic Machine Learning For Customizable Exoskeletons

Authors:

Lorenza Cotugno[1*†], Stefano Pellegrino[1], Roberta Siciliano[1]

[1] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: lorenza.cotugno@unina.it

Keywords:

Recurrent Neural Network, Long Short-Term Memory network, Gated Recurrent Unit, Latent Budget Model

Abstract:

Lower limb assistive technologies can help transfemoral amputees restore locomotion and perform diverse daily activities. The ENcyclopedia of Able-bodied Bilateral Lower Limb Locomotor Signals (ENABL3S) serves as a benchmark dataset, capturing neuro-mechanical signals via wearable sensors during locomotor tasks such as sitting, standing, walking and ascending or descending stairs and ramps, where these tasks constitute an imbalanced target variable. This paper aims to offer a unified perspective on both predictive accuracy and interpretability by applying two complementary methodologies to the same complex dataset. Real time locomotor prediction is accomplished using Recurrent Neural Networks (RNN), including Long Short-Term Memory Networks(LSTM) and Gated Recurrent Units (GRU). In parallel, the Latent Budget Tree, a probabilistic machine learning model, is employed to optimize feature settings for all tasks, including underrepresented classes, providing valuable insights into locomotion dynamics. This dual perspective highlights the need for personalized exoskeletons and effective management of complex motor transitions.

Computing An Agreement Measure For Crowdsourcing In Information Retrieval By Accurate Estimation Of a Two-Way Beta Regression Model

Authors:

Giuseppe Alfonzetti[1*†], Ruggero Bellio[1], Paolo Vidoni[1]

[1] University of Udine

[*] Corresponding author [†] Presenter

Contact: giuseppe.alfonzetti@uniud.it

Keywords:

Beta Regression, Crowdsourcing, Measure of Agreement, Modified Profile Likelihood, Two-way Model

Abstract:

In micro-task crowdsourcing, several workers carry out some tasks. Here, we focus on applications in the field of information retrieval, such as the crowdsourcing of the relevance of a document to a query. In broad generality, each worker assesses a set of items by rating each of them, and every item is rated by several workers: measuring the agreement among workers on the same task is essential. In the relevant literature of the field, a specialized measure has been proposed, denoted by Φ. Such a measure satisfies several desirable properties, and to a large extent, it is preferable to alternative existing measures. A key feature of the original formulation of the Φ measure is the capability of adjusting for the individual relevance level of different items; a simple extension, considered here, allows to adjust for the individual relevance level of different workers as well. Such flexibility derives from the fact that Φ is a one-to-one transformation of the precision parameter ϕ of a beta regression model, with the response given by the task rates and including individual dummies for different items and workers in the model for the mean. In inferential terms, evaluating the agreement measure Φ thus corresponds to estimating the precision parameter of interest, reducing or even removing the bias caused by the estimation of many incidental nuisance parameters. Here, we follow the econometric literature on two-way models with fixed effects and estimate the precision parameter of interest by modifying the profile likelihood for it, including some terms designed to reduce the incidental nuisance parameter bias. We illustrate how to streamline the model estimation, obtaining a real-time calculation of the agreement measure even in datasets with several hundred or thousands of different items and workers, with a noteworthy improvement in computation time and accuracy over the existing computational algorithms. Real data analyses and Monte Carlo studies illustrate and validate the methodology. Finally, we hint at extending the computation to cover the case of ordinal rates, achieved by the recourse to a beta regression model with a discretized response.

Enhancing Small-Sample Inference For Health Indicators: A Machine Learning Framework Applied To Eu Countries

Authors:

Cafieri Simona[1*†], Petrucci Francesco[2]

[1] Istat
[2] Sapienza University

[*] Corresponding author [†] Presenter

Contact: cafieri@istat.it

Keywords:

Health indicators, Small Sample Prediction, Machine Learning

Abstract:

Health is a complex and multidimensional concept that plays a central role in both individual well-being and societal development. This study investigates the use of machine learning techniques for predicting health-related indicators in the 27 European Union (EU) countries, with a particular focus on overcoming the limitations posed by small sample sizes. Two distinct but complementary indicators are examined: the share of individuals reporting "bad or very bad" perceived health, and life expectancy at birth. These are modelled using a common set of socioeconomic, environmental, and lifestyle-related covariates. The analysis relies on data from the 2014 and 2019 waves of the European Health Interview Survey (EHIS) and additional Eurostat sources. The proposed methodology includes the application of regression trees, random forests, and neural networks. To mitigate the effects of data sparsity (only 27 observations per year), a tailored data augmentation strategy is introduced. This method generates synthetic data by adding Gaussian noise to the original observations, thereby expanding the training set while preserving the statistical structure of the variables. The use of principal component analysis (PCA) further improves model performance by reducing dimensionality and controlling noise. Results show that while traditional models perform poorly due to the small sample, the combined use of data augmentation and PCA enables the training of more robust machine learning models. Neural networks in particular exhibit superior predictive accuracy, especially for the more volatile and subjective indicator of perceived health. The best-performing model achieves a low root mean squared error (RMSE) and high explained variance when predicting 2019 outcomes using only 2014 data. Interestingly, the architecture of the optimal neural network differs depending on the outcome: a simpler structure suffices for perceived health, while life expectancy requires deeper networks, suggesting structural differences in the two phenomena. A scenario analysis is also presented, simulating the impact of potential policy changes-such as reduced smoking rates or increased physical activity-on the predicted health indicators. While perceived health responds sensitively to changes in the covariates, life expectancy appears more inert, likely reflecting its cumulative and long-term nature. This asymmetry supports the hypothesis that subjective health measures, despite their limitations, may offer a more immediate picture of population well-being and can be useful for short-term policy planning. Finally, the study compares the performance of machine learning models with traditional regression approaches. While data augmentation improves all models to some extent, machine learning methods consistently outperform classical techniques in predictive tasks. This underscores the potential of AI-based tools to support evidence-based decision-making in official statistics, even in contexts

with limited data availability. The approach is generalizable and may serve as a blueprint for health monitoring and forecasting across diverse national or regional contexts.

Bayesian Inference With Besov-Laplace Priors For Spatially Inhomogeneous Binary Classification Surfaces

Authors:

Matteo Giordano[1*†]

[1] Università degli Studi di Torino

[*] Corresponding author [†] Presenter

Contact: matteo.giordano@unito.it

Keywords:

Bayesian nonparametric inference, Posterior contraction rates, Spatially inhomogeneous functions

Abstract:

In this article, we study the binary classification problem with supervised data, in the case where the covariate-to-probability-of-success map is possibly spatially inhomogeneous. We devise nonparametric Bayesian procedures with Besov-Laplace priors, which are prior distributions on function spaces routinely used in imaging and inverse problems in view of their useful edge-preserving and sparsity-promoting properties. Building on a recent line of work in the literature, we investigate the theoretical asymptotic recovery properties of the associated posterior distributions, and show that suitably tuned Besov-Laplace priors lead to minimax-optimal posterior contraction rates as the sample size increases, under the frequentist assumption that the data have been generated by a spatially inhomogeneous ground truth belonging to a Besov space.

Session - Latent structures and regularization in graphical, causal and deep clustering models

Generalized Estimating Equation Methods With a New Interpretation Of Goodness Of Fit Measures. The Impact Of Digitalization Level On Gdp Of The European Countries.

Authors:

Anna Crisci[1*†], Pasquale Sarnacchiaro[1]

[1] University of Naples Federico II

[*] Corresponding author [†] Presenter

Contact: anna.crisci@unina.it

Keywords:

Generalized Estimating Equations, Goodness of fit measure, Composite indicator, DESI, GDP

Abstract:

In this research, we consider Generalized Estimating Equations (GEE) for longitudinal data, offering a new interpretation of goodness-of-fit measures. GEE is a population-based approach built on a quasi-likelihood function, providing marginal or "population-averaged" estimates of the parameters. The central idea behind GEE is to generalize and extend the usual likelihood equations used in Generalized Linear Models by incorporating the covariance matrix of the responses. A key advantage of GEE is that it does not require full specification of the response distribution; only the mean structure, the mean-variance relationship, and the covariance structure need to be defined. Specifically, we present coefficients of determination based on three statistics-Wald, Likelihood ratio test, and Lagrange Multiplier-and illustrate their inequality relationships. The case study examine, within the European context, the relationship between a country's level of wealth and its level of digitalization. To this end, we collected GDP data for the 27 European countries from 2017 to 2022, along with several indicators related to digitalization, used in the calculation of the Digital Economy and Society Index (DESI). The DESI is a composite index that summarizes key indicators of Europe's digital performance and monitors the progress of EU Member States across four main dimensions: Human Capital, Connectivity, Integration of Digital Technology, and Digital Public Services. The GEE method is applied to assess how various indicators of digitalization in European countries-such as Advanced Skills and Development, Broadband Price Index, Digital Intensity, e-Commerce, e-Government, Fixed Broadband Coverage, Fixed Broadband Take-up, Internet User Skills, and Mobile Broadband-affect the gross domestic product (GDP) of each country.

Reframing Fair Pca: A Multiset Tensor Decomposition Perspective

Authors:

Violetta Simonacci[1*†], Michele Gallo[2], Lucio Palazzo[2]

[1] University of Naples Federico II - Dept. of Political Sciences
[2] University of Naples - L'Orientale

[*] Corresponding author [†] Presenter

Contact: violetta.simonacci@gmail.com

Keywords:

Sensitive Attributes, Dimensionality Reduction, Unfair Subspace, Algorithmic Fairness, PARAFAC2

Abstract:

Fairness in statistical modeling has become increasingly important as data-driven systems influence decisions in multiple domains. Acknowledging and mitigating bias is essential to prevent models from perpetuating or amplifying existing social inequalities. In multivariate analysis, one prominent approach to fairness is Fair PCA, which seeks to remove information related to sensitive attributes by projecting data away from directions deemed unfair. These directions are typically defined using differences in group means and second moments, forming what is known as the unfair subspace. The "null-it-out" method then ensures that the resulting low-dimensional representation is statistically independent of the sensitive attribute. In this work, we propose a novel reinterpretation of this framework through a tensor-based lens. Rather than aggregating group-level statistics, we treat the data associated with each group of the sensitive attribute as a separate matrix, forming a multi-subject three-way array. This tensor structure naturally captures variation across groups and allows us to model the data using a PARAFAC2 decomposition, a flexible multilinear model designed for multiset data. By enforcing orthogonality constraints, we extract a shared loading matrix that captures dominant directions of variation across groups, which can be interpreted as the unfair subspace. Projecting the original data away from this space yields a fair representation. Unlike existing methods, our approach models all groups jointly, capturing shared and structured unfairness without relying on pairwise comparisons. It is particularly well-suited for settings with multiple sensitive groups of large size or heterogeneous structure. Preliminary results on synthetic and real-world datasets show that our method effectively removes sensitive information while preserving meaningful variance.

Revealing The Nature Of Italian Life Expectancy: A Comparative Study Of Arima Models Using Covid-19 Shock

Authors:

Girolamo Franchetti[1], Massimiliano Politano[2*†]

[1] University of Naples Federico II
[2] University "Federico II" of Naples

[*] Corresponding author [†] Presenter

Contact: politano@unina.it

Keywords:

Life expectancy, ARIMA models, COVID 19

Abstract:

This study investigates how alternative ARIMA model specifications can be used to infer the underlying trend structure-deterministic or stochastic-of the life expectancy at birth time series for the Italian population over the period 1974-2024. By comparing two ARIMA(1,d,0) models and two ARIMA(1,d,1) models, each estimated with and without a deterministic trend component, we aim to assess not only forecast accuracy but also the capacity of each model to capture the structural dynamics of the series, particularly in the presence of exogenous shocks. The COVID-19 outbreak in 2020 is treated as a structural shock and serves as a testing ground for evaluating model adaptability and long-run behavior. Our analysis employs stationarity tests, residual diagnostics, impulse response functions (IRFs), model fit statistics, and forecast error measures. Results indicate that while trend-based ARIMA models tend to provide better in-sample statistical fit, they often fail to capture the persistent deviations induced by structural breaks. In contrast, the ARIMA(1,d,1) model without a deterministic trend offers greater flexibility and superior post-shock forecasting performance. The paper concludes by proposing a structured approach to model selection under structural uncertainty, highlighting how comparative model analysis can inform our understanding of time series behavior over a given historical period.

From Vectors To Networks: Comparing Conventional And Graph-Based Approaches To Unsupervised Text Categorisation

Authors:

Maria Spano[1], Michelangelo Misuraca[2], Luigi Celardo[3*†]

[1] University of Naples Federico II
[2] University of Salerno
[3] National Research Council - Institute of Polymers, Composites and Biomaterials

[*] Corresponding author [†] Presenter

Contact: luigicelardo@cnr.it

Keywords:

unsupervised classification, community detection, comparative study

Abstract:

One of the primary tasks of text mining is to organise a large number of unlabeled documents into a smaller set of meaningful and coherent clusters that are similar in content. Clustering algorithms typically operate on document × term matrices, where each document is represented as a vector in an algebraic format. Alternatively, a collection of documents can be represented using a documents × documents structure, which can be viewed as an adjacency matrix and graphically depicted as a graph. In network analysis, community detection is used on these graphs to identify groups of nodes that share common characteristics and perform similar functions. This paper aims to evaluate different data structures and grouping criteria, showing the effectiveness of various alternatives in a text categorisation strategy. We conduct a comparative study involving classical text clustering methods and community detection approaches, examining and discussing their performances.

Session - Structured and multi-way data modeling

Nomclust 3.0: An R Package For Agglomerative Hierarchical Clustering Of Categorical And Mixed-Type Data

Authors:

Zdenek Sulc[1*†], Jaroslav Horníček[1]

[1] Prague University of Economics and Business

[*] Corresponding author [†] Presenter

Contact: zdenek.sulc@vse.cz

Keywords:

R package, hierarchical clustering, mixed-type data, dissimilarity measures, clustering with missing data

Abstract:

The nomclust R package, which is available on CRAN, was developed to cover agglomerative hierarchical clustering of objects characterized by the categorical variables, from a dissimilarity matrix computation over the choice of a clustering method to the final cluster evaluation. The large set of dissimilarity measures and many evaluation criteria developed explicitly for categorical data make this package unique. This contribution presents the third major release of this package that brings four substantial enhancements. First, it adds several recently proposed dissimilarity measures for mixed-type data that can serve as alternatives for the Gower distance measure commonly used in this field. Some of them outperform the Gower measure substantially in relation to the previous research based on generated datasets. Second, two modifications of the original BIC and AIC evaluation criteria adjusted for mixed-type data are included in the package. These criteria can help a researcher assess the produced clusters' quality and determine their optimal number. Third, the package enables a user to cluster incomplete data using one of several recently proposed methods, such as a modification of the average method for the quantitative data, which have in common the ability to replace the missing values temporarily. These imputation techniques are implemented within the hierarchical clustering process and impute a possibly different value in each clustering step for a given missing. Thus, no imputation method during data preprocessing is necessary. The last main enhancement is a new optimization method for dissimilarity matrix calculation that can substantially reduce the computation time of clustering objects in large and purely categorical datasets. It is based on the principle that many objects in categorical datasets have the same profile, so their mutual dissimilarities can be calculated only once. The presented enhancements further improve the functionality of the third release of the package, which is usable even for more possible scenarios.

Closed-Form Information Matrix Expressions For Matrix-Normal Mixtures

Authors:

Marco Berrettini[1*†], Giuliano Galimberti[1]

[1] University of Bologna

[*] Corresponding author [†] Presenter

Contact: marco.berrettini2@unibo.it

Keywords:

asymptotic variance, score vector, Hessian matrix, three-way data

Abstract:

Three-way data, i.e., data organized along three distinct modes or dimensions, frequently occur in diverse scientific areas including psychometrics, bioinformatics, and neuroimaging. When dealing with these data, each observation can be represented as a matrix with variables measured across multiple occasions, calling for statistical models that preserve this inherent matrix structure. The matrix-variate normal distribution extends the multivariate normal framework to matrix-valued data by modeling dependence across both rows and columns via separate covariance matrices. When data originate from heterogeneous populations, mixture models provide a flexible framework to capture latent group structures. By assuming matrix-variate normal distributions as mixture components, it is possible to model complex three-way data while preserving their multidimensional nature. Parameter estimation in these models typically relies on maximum likelihood via the EM algorithm. However, compact closed-form expressions for the score vector and Hessian matrix had not been derived before. This work derives such expressions by exploiting properties of the trace operator and the Kronecker product, enabling fast and accurate estimation of standard errors without resorting to numerical differentiation. Unlike approaches that vectorize the data and ignore matrix structure, this method avoids overparameterization and preserves separable covariance forms. Results show that standard errors computed from the Hessian matrix closely match those obtained via numerical approximations when sample sizes are large, with the added benefit of greater numerical stability and reduced computational time. In smaller samples or partially overlapping clusters, score-based estimates may sometimes yield improved accuracy for certain parameters. Numerical methods occasionally fail due to instability in likelihood evaluation, while the closed-form approach remains robust. Deriving the exact score vector and Hessian matrix for mixtures of matrix normal distributions opens up possibilities for additional applications, such as testing model misspecification via the information matrix test. Further extensions consider parsimonious model by imposing constraints on covariance and mean parameters, helping to address high-dimensional settings and reduce model complexity.

On Relations Of Piecewise-Linear Approximations

Authors:

Daniyal Kazempour[1*†], Emil Lambert[1], Claudius Zelenka[2]
Peer Kröger[1]

[1] Christian-Albrechts-Universität zu Kiel
[2] Kiel University

[*] Corresponding author [†] Presenter

Contact: dka@informatik.uni-kiel.de

Keywords:

GMM, MASO, LLE, ORCLUS, VQPCA, Piecewise-Linear Approximations, Dimensionality Reduction

Abstract:

Gaussian Mixture Model (GMM), Max-Affine Spline Operator (MASO), Locally Linear Embedding (LLE), arbitrarily ORiented projected CLUSter generation (ORCLUS) and Vector-Quantised Principal Component Analysis (VQPCA) were introduced to solve five seemingly unrelated problems: (a) probabilistic density estimation, (b) spline-based neural-network approximation, (c) manifold learning, (d) subspace clustering and (e) non-linear dimensionality reduction. Despite these distinct motivations, we observe that each of these methods, whether implicitly or explicitly, appears to leverage a form of piecewise-linear approximation (PLA) to model the data. This suggest a shared principle: learning a latent data model through a collection of local, linear representations. This leads us to hypothesize that the core logic of each algorithm can be viewed through a unified lens. We postulate the learning process of these method can be conceptualized through the interplay of two sub-tasks. Given a dataset X and a set of parameters π: The first is a partition function, $\Psi(X, \pi_{clustering})$ on a, that groups data into distinct clusters. The second is a dimnsionality reduction function, $\Delta(X, \pi_{manifold})$, that defines a local, lower-dimensional subspace for each of those clusters. In essence, we postulate that all five methods involve some form of this join process, mapping data points to (cluster, subspace) pairs. This unifying perspective motivates following pivotal question: *To what extent can the unique behaviours, objectives, and potential for mutual emulation among these five algortihms be explained through the specific interplay of their core components: data partitioning (Ψ) and local dimensionality redution (Δ)?* Towards answering this questions we compare, for instance, the probabilistic, soft clustering of GMM with the hard, geometric clustering of a MASO, and contrast the manifold learning method LLE with the clustering method ORCLUS. We explore this question, offering a new perspective on the connections between these algortihms.

Measuring Dynamics In Spatio-Temporal Clusters

Authors:

Andrzej Sokołowski[1*†], Małgorzata Markowska[2]

[1] Andrzej Frycz Modrzewski Krakow University
[2] Wroclaw University of Economics and Business

[*] Corresponding author [†] Presenter

Contact: sokolows@uek.krakow.pl

Keywords:

Cluster analysis, Time series analysis, Regional data, Cluster stability

Abstract:

In dynamic cluster analysis data, we have a set of Y objects characterized by Z variables, and the data covers period of time T. The question is which objects and when are similar, and can form clusters. If there are m objects and n time points, we have m*n points in the multidimensional space of variables Z. Dynamic cluster can be described by the binary membership matrix B with rows representing objects, and columns representing time units. If i-th objects was present in the cluster, in j-th time unit, than bij=1, and otherwise it is 0. In the paper we discuss eight versions of cluster stability measure in seven artificial simple membership matrices. Examples under discussion consist of five objects in six consecutive time points. Seven test membership matrices show different dynamics patterns, from full stability (B matrix has only 1s) to the situation when each object in present only once within time span covered by the cluster. The analysis leads to the final recommendation. As the real data example, the dynamic clustering of 20 NUT2 regions of Italy from the point of view of economic development level has been performed. They are characterized by the following variables: Gross Domestic Product, Bank services, Gross fixed investment, National consumption (home plus collective) and Total units of labour. First four are originally in millions of Euro. All these variables are finally expressed per capita. Variables were downloaded from CRENoSTerritorio database, covering the period of 35 years. The number of dynamic clusters was identified by the analysis of agglomerative distances in Ward's method, and final partition was obtained by k-means method.

Session - Validation and innovation in clustering: from hierarchical stability to spatio-temporal pattern discovery

Instability Measures For Hierarchical Clustering In Categorical Data

Authors:

Jaroslav Horníček[1*†]

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business

[*] Corresponding author [†] Presenter

Contact: horj31@vse.cz

Keywords:

bootstrapping, categorical data, clustering instability, evaluation criteria, hierarchical clustering

Abstract:

The contribution introduces two novel evaluation criteria for determining the optimal number of clusters based on the bootstrapping estimation of clustering instability. Various measures can express this quantity, but the shared key idea is to assess and utilize a disagreement between partitions obtained from repeated iterations of a clustering process. The approach can help to determine the optimal number of clusters regardless of the clustering algorithm and data type. However, the assumption about data type is very hypothetical. Whereas the clustering instability measure is well studied for quantitative data, the situation is different when data are categorical. The proposed modifications of the clustering instability measure address the specific and often challenging properties of categorical data and remove the disadvantages of the original approach, which tends to prioritize a solution with a lower number of clusters. The idea behind the novel criteria is to compare the obtained partitions against the random ones and utilize the concept of Shannon entropy. The contribution compares the original and proposed adjusted versions of the clustering instability measure with three other well-studied criteria on completely simulated datasets covering multiple parameters. The same approach is replicated on a sample of three real datasets. The experiment results show that the proposed modifications can better find an optimal number of clusters than the original ones and underestimate or overestimate this value less than other compared criteria. An additional advantage of the clustering instability function is its potential application in determining the optimal number of clusters, even in data with missing values.

Comonotonic-Based Time Series Clustering With Soft Spatial Constraints

Authors:

Alessia Benevento[1*†], Fabrizio Durante[1], Roberta Pappadà[2]

[1] University of Salento
[2] University of Trieste

[*] Corresponding author [†] Presenter

Contact: alessia.benevento@unisalento.it

Keywords:

Time Series, Copulas, Geo-referenced Data

Abstract:

Clustering methods are widely adopted in the analysis of geo-referenced time series, which requires incorporating spatial information into the definition of a suitable dissimilarity. Among the dissimilarity-based clustering methods, the algorithms using soft constraints are those embedding spatial information on the time series into the clustering procedure without requiring that the resulting clusters strictly adhere to proximity constraints. To focus on the dependence among different time series, we explore copula-based dissimilarity functions that capture their comovements and/or tail dependence behavior, regardless of marginal modeling. These measures are of particular relevance in environmental sciences, e.g., for analyzing joint extremes such as maxima of precipitations, temperature, or in modeling flood risks. Specifically, we present a general approach to time series clustering based on three paradigms: (C1) the use of copulas to focus on the dependence among time series; (C2) the detection of comovements that are interpreted as proximity of the pairwise copula between two time series to the comonotonicity case; (C3) the presence of some constraints that may guide the clustering process from a semi-supervised viewpoint. In particular in the latter paradigm, we discuss how spatial constraints can be embedded into the clustering algorithm, providing two comprehensive model architectures to handle this problem. Such architectures exploit an aggregation step for merging spatial and temporal dependencies that is based either on dissimilarity matrices or on copula functions. The presented framework is illustrated via simulated scenarios in order to highlight the comparison with the pure temporal or pure spatial clustering and discuss the critical issue of selecting the number clusters and the hyperparameter tuning the related influence of the spatial component on the whole procedure.

Graph-Based Deep Learning Approach For The Classification Of Earthquake Magnitudes In Space And Time

Authors:

Orietta Nicolis[1*†], Billy Peralta[1], Alonso Rivera[1]

[1] Universidad Andres Bello

[*] Corresponding author [†] Presenter

Contact: nicolisorietta@gmail.com

Keywords:

Classification, Earthquake Magnitude, Deep Learning, LSTM, GCN-LSTM

Abstract:

Earthquake prediction remains one of the most complex and urgent challenges in natural hazard research, due to the highly nonlinear, dynamic, and heterogeneous nature of seismic processes across both space and time. This study addresses the problem as a spatio-temporal classification task, where the goal is to categorize the daily maximum seismic magnitude into discrete intensity classes (e.g., low, moderate, high) across different regions of Chile. The proposed framework integrates Graph Convolutional Networks (GCNs) with Long Short-Term Memory (LSTM) models to capture both spatial and temporal dependencies in earthquake behavior. The study utilizes an earthquake catalog from the Chilean Seismological Center, covering the years 2000 to 2023. During the data preprocessing stage, the raw seismic records are filtered to remove duplicates, correct inconsistencies, and eliminate events below a defined magnitude threshold. Following this cleaning process, the spatial distribution of epicenters is used to apply K-means clustering, extracting 20 distinct seismic zones across the country. Each cluster is represented as a node in a graph, and spatial proximity among clusters defines the edges. Two deep learning models are developed and compared: (1) a standard LSTM network, which independently learns temporal patterns in each seismic zone, and (2) a hybrid GCN-LSTM model, which enhances predictive accuracy by incorporating spatial correlations between regions through graph convolutions. Both models are trained to classify the magnitude class expected for each zone on the following day, resulting in a multi-region, multi-output classification problem. Experimental results show that the GCN-LSTM consistently outperforms the LSTM, particularly in geologically complex areas such as northern Chile. The hybrid model achieves an average F1-score of 0.72, compared to 0.58 for the LSTM, with precision and recall improvements of up to 20 % in the high-magnitude class. These gains demonstrate the added value of modeling spatial relationships in seismic forecasting tasks. Future developments will aim to enhance this framework in several directions. First, we plan to incorporate additional geophysical variables, such as ground deformation from GNSS networks. Second, we aim to explore attention-based graph architectures to dynamically learn the importance of connections between seismic regions. Finally, we will investigate transfer learning techniques to adapt the trained models to other tectonic regions (e.g., Turkey or Japan), enhancing generalizability and robustness. These developments will move us closer to a reliable and interpretable data-driven tool for seismic risk forecasting.

Session - Bayesian and nonparametric frontiers in network and spatial classification

Advances On Combined Permutation Tests In Multivariate Problems

Authors:

Stefano Bonnini[1*†], Michela Borghesi[1], Massimiliano Giacalone[2]

[1] University of Ferrara
[2] University of Campania "Luigi Vanvitelli"

[*] Corresponding author [†] Presenter

Contact: stefano.bonnini@unife.it

Keywords:

Multivariate Statistics, Nonparametric Inference, Combined Permutation Test

Abstract:

When addressing complex statistical problems, such as multivariate analyses or scenarios requiring multivariate tests, a powerful and versatile approach is offered by the family of Combined Permutation Tests (CPT). This methodology involves conducting a set of partial tests and combining their p-values using an appropriate function to produce an overall (univariate) test statistic valid for the multivariate problem. CPTs are widely applicable in categorical data analysis, big data problems, regression models, and beyond. One of their main advantages over parametric methods is that it is not necessary to assume a certain multivariate distribution of the test statistics and then a dependence structure among variables. In fact, there is no need to model or estimate such dependencies explicitly. As permutation tests are distribution-free, they offer robustness and flexibility, particularly in the presence of deviations from normality. Dependence among partial tests is implicitly handled through row permutations of the data matrix. An effective combining function for the partial p-values should satisfy a few reasonable conditions: (1) it should be a monotonic non-increasing function of the p-values; (2) it should approach the supremum when a p-value approaches zero; (3) it should have a bounded acceptance region. Traditional parametric approaches, both univariate and multivariate, often rely on strict assumptions that may be unrealistic or only justified asymptotically. CPTs provide a valid alternative, performing well even when parametric conditions are met and excelling when they are not. Moreover, in the context of multiple testing, controlling the family-wise error rate (FWE) is essential. This allows to identify the partial tests which contribute to the possible overall significance while avoiding inflation of the type I error rate. The work deals with advances of this approach, in order to identify a more performant and robust method and investigate its properties.

A Multiple Random Scan Strategy For Bayesian Latent Space Models

Authors:

Antonio Peruzzi[1*†], Roberto Casarin[2]

[1] Ca' Foscari University of Venice
[2] University Ca' Foscari of Venice

[*] Corresponding author [†] Presenter

Contact: antonio.peruzzi@unive.it

Keywords:

Latent Space Models, Bayesian Inference, Adaptive MCMC

Abstract:

Latent Space (LS) network models project the nodes of a network on a d-dimensional latent space to achieve dimensionality reduction of the network while preserving its relevant features. Inference is often carried out within a Markov Chain Monte Carlo (MCMC) framework. Nonetheless, it is well-known that the computational time for this set of models increases quadratically with the number of nodes. In this work, we build on the Random-Scan (RS) approach to propose an MCMC strategy that alleviates the computational burden for LS models while maintaining the benefits of a general-purpose technique. We call this novel strategy Multiple RS (MRS). This strategy is effective in reducing the computational cost by a factor without severe consequences on the MCMC draws. Moreover, we introduce a novel adaptation strategy that consists of a probabilistic update of the set of latent coordinates of each node. Our Adaptive MRS adapts the acceptance rate of the Metropolis step to adjust the probability of updating the latent coordinates. We show via simulation that the Adaptive MRS approach performs better than MRS in terms of mixing. Finally, we apply our algorithm to face-to-face interaction data and to climate-related Facebook data. We show how our adaptive strategy may be beneficial to empirical network applications.

Adaptive Multiscale Clustering Of Spatial Panels Via Bayesian Wavelet Decomposition

Authors:

Antonio Pacifico[1*†]

[1] University of L'Aquila

[*] Corresponding author [†] Presenter

Contact: antonio.pacifico@univaq.it

Keywords:

Fuzzy Clustering, Spatial Panel Data, Dynamic Bayesian Classification

Abstract:

This work proposes an advanced classification framework for the analysis of spatial panel data characterized by structural heterogeneity, temporal nonstationarity, and spatial dependence. The methodology, termed Wavelet Adaptive Spatial Partitioning-Clustering (WASP-C), combines multiscale feature extraction, dynamic Bayesian variable selection, and spatially informed clustering to address the complexities inherent in high-dimensional spatio-temporal datasets. Initially, a Discrete Wavelet Transform (DWT), which decomposes each temporal trajectory into a multiscale basis. This allows the simultaneous representation of both smooth long-term trends and localized high-frequency variations, capturing transitory anomalies that traditional time-averaging methods tend to obscure. The resulting wavelet coefficients provide a rich set of time-frequency features for each spatial unit. To perform effective dimensionality reduction and noise control, a dynamic Bayesian shrinkage procedure is applied, using the Horseshoe prior to estimate Posterior Inclusion Probabilities (PIPs) for each wavelet coefficient. PIPs are employed as adaptive feature weights, ensuring that the clustering process focuses on the most informative and persistent components, while attenuating the impact of noisy or irrelevant fluctuations. The classification phase utilizes a fuzzy c-medoids algorithm specifically adapted to incorporate spatial information. A composite dissimilarity measure is constructed, integrating feature-based distances with spatial adjacency constraints. This enables the framework to account for partial membership, essential for correctly classifying units located in transitional or frontier zones, and to preserve geographic coherence in the clustering results. The methodology operates in a dual-stage fashion. The "static" clustering is performed by applying the fuzzy spatial clustering algorithm to the multiscale decomposition of local coefficient estimates obtained from a spatial panel model, fitted separately for each cross-sectional unit across the complete time series. This step captures persistent spatial heterogeneity in the covariate effects. Subsequently, the "dynamic" clustering focuses on the multiscale structure of the residuals, identifying latent shocks and transient spatial patterns not explained by the observed covariates. This two-stage approach enables a more comprehensive understanding of both structural and evolving behaviors within the data. Validation is performed through silhouette scores, Bayesian Information Criterion (BIC), and stability analyses based on Jaccard similarity. The results confirm the ability of the WASP-C framework to generate coherent, interpretable, and robust clusters even in the presence of strong multiscale variability and spatial dependence. By integrating multiscale analysis, Bayesian adaptivity, and spatially constrained fuzzy clustering, the proposed framework advances current classification techniques for complex spatio-temporal data, with wide applicability across economics, environmental sciences, and social studies.

Understanding Esg Scores Through Network Analysis: A Study Using Graph Neural Networks

Authors:

Brian Daniel Bernhardt[1], Chiara Marciano[1*†], Sara Gigli[1]
Lucia Maddalena[2], Mario Rosario Guarracino[1]

[1] University of Cassino
[2] National Research Council, Naples, 80131 Italy

[*] Corresponding author [†] Presenter

Contact: chiara.marciano@unicas.it

Keywords:

ESG Perception Index, Decision-Makers, Graph Neural Network, Interpretability

Abstract:

Graph neural networks (GNNs) are powerful tools for analyzing graph-structured data and have numerous applications. In this study, GNNs are used to explore the connections between a set of Italian companies, focusing on how sharing decision-makers influence their Environmental, Social, and Governance (ESG) scores. The research compares various GNN models, including GraphSAGE, Graph Attention Networks, and Graph Neural Additive Networks, with traditional classification techniques such as kNN, Random Forest, Decision Trees, and Na¨ıve Bayes. The results demonstrate that GNNs provide a higher node classification accuracy and suggest that more central nodes tend to have higher ESG scores, indicating that companies with greater network connectivity may provide higher ESG performance. This suggests that sharing decision-makers could be a strategic tool to enhance efforts in sustainability and social responsibility.

Session - Automated financial analysis and funding matching: statistical validation of AI modeling

Can Genai Match Human Standards In Financial Reports?

Authors:

Fabio Leone[1*†], Antonella Meccariello[1], Daniele Rossi[1]
Luca Ruocco[1], Carmine Santone[1], Sergio Beraldo[1]
Antonio D'Ambrosio[1], Carmela Iorio[1], Giovanni Walter Puopolo[1]

[1] Department of Economics and Statistics, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: fabio.leone@unina.it

Keywords:

Generative AI, Automated Financial Reporting, Human Supervision, Semantic Similarity, AI Reliability

Abstract:

Recent developments in the use of Generative Artificial Intelligence (GenAI) promise to accelerate the composition of financial reports, yet concerns remain about their reliability, semantic accuracy, and ability to meet professional standards. This study presents a GenAI-driven tool, developed by the authors, designed to automatically generate financial reports. The tool was developed with the aim of providing support to analysts and investors, combining the efficiency of automated production with expert human supervision. Such an approach seeks to ensure interpretative consistency, numerical accuracy and alignment with professional reporting standards. First, the work outlines the methodological framework that constituted the functionalities of the GenAI-driven tool and the role of human supervision, which was a key feature of the entire process. Second, a simulation study is proposed in order to assess the reliability of the report generated. To this end, the system generated n reports of the same company. Then, a rigorous analytical comparison was conducted between each of the automatically generated reports and a human-validated benchmark. Therefore, the winning report of the 2024 Chartered Financial Analyst (CFA) Institute Research Challenge was selected, since its outstanding quality has already been validated in a competitive context by specialist evaluators. In order to execute the comparison, a semantic similarity approach has been followed, using Cosine Similarity. Furthermore, Jaccard similarity and Dice coefficient has been estimated as robustness checks. The results show interesting evidence of the proposed GenAI-driven tool reliability within a professional and competitive context. Furthermore, the achievements illustrated contribute to the ongoing debate about the integration of GenAI systems into financial reporting routines. Acknowledgements: Financial support from the National Centre for HPC, Big Data and Quantum Computing (000004_PNRR_CN_HPC_BIG_DATA_SPOKE_9 - PNRR CN00000013 - National Centre for HPC) is gratefully acknowledged. The present work reflects only the authors' view and the Funding Agency can not be held responsible for any use that may be made of the information it contains.

Enhancing Access To European Funding Through Ai-Powered Tool

Authors:

Fabio Leone[1], Antonella Meccariello[1], Daniele Rossi[1]
Luca Ruocco[1], Carmine Santone[1*†], Sergio Beraldo[1]
Antonio D'Ambrosio[1], Carmela Iorio[1], Giovanni Walter Puopolo[1]

[1] Department of Economics and Statistics, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: carmine.santone@unina.it

Keywords:

Small and Medium Enterprises (SMEs), European Fundings, Semantic Matching, Large Language Model (LLM), Embedding Model

Abstract:

Small and medium-sized enterprises (SMEs) are widely acknowledged as the backbone of the European economy, representing approximately 99% of all firms within the European Union. Access to external funding is essential to sustain growth, foster innovation, and support internationalization efforts. Despite the availability of multiple funding instruments-including Horizon Europe, COSME, and the European Structural and Investment Funds-many SMEs face persistent challenges in identifying and successfully applying to the most relevant calls for proposals. Institutional platforms such as the EU Funding and Tenders Portal often rely on rigid taxonomies and user-specific search histories, creating barriers for firms with limited prior experience or insufficient familiarity with EU programs. In the framework of an advanced corporate reporting tool, this work explores how artificial intelligence can enhance the discovery of relevant European funding opportunities for SMEs. The tool seeks to align the core objectives of a business project with thematically related calls for proposals, moving beyond conventional keyword-based search methods by capturing the conceptual meaning of project and funding call descriptions. The methodology is composed of three main stages. First, to mitigate irrelevant matches caused by generic terminology, active and forthcoming calls are classified into predefined thematic areas using curated keywords that reflect EU strategic priorities. Second, project description is refined through a Large Language Model (LLM) to improve clarity and ensure alignment with the language commonly used in funding programs. Finally, only the textual descriptions of the project and the calls consistent with its thematic area are embedded in a high-dimensional semantic space. Then, cosine similarity is computed to rank the most relevant opportunities. The proposed tool was empirically evaluated on a dataset of real-world projects that successfully obtained EU fundings. For each project, a set of temporally overlapping calls was reconstructed to rigorously assess the tool's capability in retrieving the exact awarded call and to provide a benchmark comparison against conventional keyword-based search methods. Results show that the proposed tool can substantially enhance thematic precision and reveal relevant funding opportunities that may remain undetected through conventional keyword-based searches. However, its effectiveness decreases in cases where funding eligibility is driven mainly by procedural requirements or specific applicant characteristics, rather than thematic alignment. Therefore, these findings highlight the value of complementing the tool's semantic matching capabilities with rule-based filters and eligibility checks. Such integration ultimately contributing to more reliable and inclusive decision-support systems for SMEs seeking EU funding opportunities. Ac-

knowledgements: Financial support from the National Centre for HPC, Big Data and Quantum Computing (000004_PNRR_CN_HPC_BIG_DATA_SPOKE_9 - PNRR CN00000013 - National Centre for HPC) is gratefully acknowledged. The present work reflects only the authors' view and the Funding Agency can not be held responsible for any use that may be made of the information it contains.

From Data To Decision: A Scalable Ai Approach To Public And Private Funding Discovery

Authors:

Fabio Leone[1], Antonella Meccariello[1], Daniele Rossi[1]
Luca Ruocco[1*†], Carmine Santone[1], Sergio Beraldo[1]
Antonio D'Ambrosio[1], Carmela Iorio[1], Giovanni Walter Puopolo[1]

[1] Department of Economics and Statistics, University of Naples Federico II, Italy

[*] Corresponding author [†] Presenter

Contact: luca.ruocco@unina.it

Keywords:

Artificial Intelligence, Semantic Matching, Public and Private Funding, Decision Support Systems, Small and Medium Enterprises (SMEs)

Abstract:

Access to public and private funding remains a critical yet underutilized opportunity for small and medium-sized enterprises (SMEs). This is often hindered by fragmented information, complex eligibility criteria, and the difficulty in identifying relevant opportunities. The present paper sets forth a modular AI-based framework that semantically matches enterprises with funding initiatives, including national, regional, and private grants. This framework leverages natural language processing, large language models (LLMs), and structured data. The system processes firm-level information extracted from institutional databases, legal documentation, and company websites. A concise, business-oriented profile is generated for each enterprise using LLMs, complemented by additional content automatically extracted via web scraping to capture sector-specific nuances and strategic positioning. Concurrently, official data concerning funding calls is standardized and transformed into structured textual records, with a focus on objectives, beneficiaries, sectors, and geographic scope. The textual content, originating from both the companies and the funding programmes, is embedded through the utilization of OpenAI's text-embedding-3-small model. This enables the execution of semantic comparisons via cosine similarity. An additional rule-based filtering layer further refines the results based on technical eligibility constraints, including sector classification, regional applicability, and legal status. The system generates a ranked list of suitable and actionable funding opportunities for each firm. A real-world case study is presented involving a high-tech training consortium operating in Southern Italy, which demonstrates the framework's effectiveness in surfacing both relevant and previously unexplored opportunities. The proposed pipeline is fully automated, interpretable, and scalable. It represents a concrete contribution toward improving access to funding resources through artificial intelligence, supporting a more strategic alignment between enterprise development and available financial instruments. Acknoweledgments: Financial support from the National Centre for HPC, Big Data and Quantum Computing (000004_PNRR_CN_HPC_BIG_DATA_SPOKE_9 - PNRR CN00000013 - National Centre for HPC) is gratefully acknowledged. The present work reflects only the authors' view and the Funding Agency can not be held responsible for any use that may be made of the information it contains.

Author Index