Data Analysis and Statistical Modeling

February 5-7, 2024

Venue: Department of Law, Economics, Management and Quantitative Methods (DEMM)
Piazza Arechi II – Palazzo De Simone
82100 Benevento, Italy

This workshop is supported by:
JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)
Institute for Mathematical Science, Tokyo – Japan
University of Sannio, Benevento – Italy
Waseda Research Institute for Science and Engineering, Tokyo – Japan

Data Analysis and Statistical Modeling

February 5-7, 2024

Program

Venue: Department of Law, Economics, Management and Quantitative Methods (DEMM)
Piazza Arechi II – Palazzo De Simone
82100 Benevento, Italy

Scientific Committee:

Anna Clara Monti (Chair), Pietro Amenta, Alessio Farcomeni, Luca Greco, Yan Liu, Antonio Lucadamo, Simona Pacillo, Stefano Pagnotta, Marco Riani, Masanobu Taniguchi.

This workshop is supported by:

- *JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290 (M. Taniguchi)*
- *Institute for Mathematical Science, Tokyo – Japan*
- *University of Sannio, Benevento – Italy*
- *Waseda Research Institute for Science and Engineering, Tokyo – Japan*

Monday February 5, 2024 – *Morning*

11:00 – 11:15: Opening

Anna Clara Monti
Masanobu Taniguchi

Session I (11:15 – 12:00)

Chaired by Ilia Negri

***Keynote speaker*:**

Philosophy of AIC
Masanobu Taniguchi (*Waseda University, Tokyo, Japan*)

Session II (12:00 – 13:00)

Chaired by Alessio Farcomeni

12:00 – 12:30

The Least Trimmed Squares for Time Series (LTSts): extensions and properties for policy support applications
Domenico Perrotta (*Joint Research Centre, European Commission, Ispra, Italy*)

12:30 – 13:00

Statistical models for electricity data forecasting
Luigi Grossi (*Univeristy of Parma, Italy*)
Joint work with F. Laurini

13:00 – 14:00 *Lunch*

Monday February 5, 2024 – *Afternoon*

Session III (14:00 – 15:30)

Chaired by Elena Stanghellini

14:00 – 14:30

Recent advances in theory and applications of mixture models with uncertainty for rating data
Rosaria Simone (*University of Naples, Italy*)

14:30 – 15:00

An approximate distribution of the dissimilarity index for CUB models
Domenico Piccolo (*University of Naples, Italy*)

15:00 – 15:30

Modelling ordinal data from repeated surveys
Marcella Corduas (*University of Naples, Italy*)

15:30 – 16:00 *Coffee break*

Session IV (16:00 – 17:30)

Chaired by Marco Riani

16:00 – 16:30

One-dimensional mixture-based clustering for ordinal responses
Marta Nai Ruscone (*University of Genova, Italy*)

16:30 – 17:00

High-dimensional data reduction in model based clustering: an application on a retailer's data
Gianluca Morelli (*University of Parma, Italy*)
Joint work with F. Laurini

17:00 – 17:30

Model based clustering for torus data
Antonio Lucadamo (*University of Sannio, Italy*)
Joint work with L. Greco and C. Agostinelli

Tuesday February 6, 2024 – *Morning*

Session V (9:30 – 11:00)

Chaired by Masanobu Taniguchi

9:30 – 10:00

Testing variable importance in high dimensions with deep neural network (NN)
Weining Wang (*University of Groningen, Netherlands*)
Joint work with Y. Zhao and J. Fan

10:00 – 10:30

On latent and selection nodes in graphical models for binary variables
Elena Stanghellini (*University of Perugia, Italy)*

10:30 – 11:00

Time-Interaction Point Processes with heterogeneity
Alessio Farcomeni (*University of Rome Tor Vergata, Italy)*
Joint work with Barone A., Mezzetti M.

11:00 - 11:30 *Coffee break*

Session VI (11:30 – 13:00)

Chaired by Weining Wang

11:30 – 12:00

Ups and (Draw) downs
Tommaso Proietti (*University of Rome Tor Vergata, Italy)*

12:00 – 12:30

Detection of periodicity in multivariate functional time series
Yan Liu **(***Waseda University, Tokyo, Japan***)**
Joint work with R. Sagawa

12:30 – 13:00

Z-Process Method in Change Point Problems for Dependent Observations
Ilia Negri (*University of Calabria, Italy*)

13:00 – 14:00 *Lunch*

Tuesday February 6, 2024 – *Afternoon*

Session VII (14:00 – 15:00)
Chaired by Domenico Piccolo

14:00 – 14:30

On testing the equality between interquartile ranges
Luca Greco (*Giustino Fortunato University, Italy*)
Joint work with G. Luta and R. Wilcox

14:30 – 15:00

An empirical evaluation of gene-set enrichment test statistics if missing a reference ground truth.
Stefano Maria Pagnotta *(University of Sannio, Italy)*
Joint work with C. La Torella and D. Risso

15:00 – 15:30

Partial Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic.
Pietro Amenta (*University of Sannio, Italy*)
Joint work with A. Lucadamo

15:30 – 16:00 ***Coffee break***

16:00 – 17:30

Informal meeting

Wednesday February 7, 2024 – *Morning*

Session VIII (9:30 – 11:00)

Chaired by Fabrizio Laurini

9:30 – 10:00

Applied Robust Statistics through the Monitoring Approach
Marco Riani (*University of Parma, Italy*)

10:00 – 10:30

Robust diagnostics and monitoring for Linear Mixed Models
Aldo Corbellini (*University of Parma, Italy*)
Joint work with F. Laurini and L. Grossi

10:30 – 11:00

An impartial trimming algorithm for robust circle fitting
Simona Pacillo
Joint work with L. Greco and P. Maresca

11:00 - 11:30 *Coffee break*

Session IX (11:30 – 13:00)

Chaired by Luigi Grossi

11:30 – 12:00

Model-based clustering with cellwise outlier detection
Francesca Greselin (*University of Milano Bicocca, Italy*)
Joint with L.A. Garcia-Escudero, A. Mayo-Iscar, G. Zaccaria

12:00 – 12:30

A compared protocol to improve clustering procedures
Silvia Salini (*University of Milano Statale, Italy*)

12:30 – 13:00

Robust correspondence Analysis with applications in international trade
Francesca Torti (*Joint Research Centre, European Commission, Ispra, Italy*)

13:00 – 13:15: Closing

Anna Clara Monti
Masanobu Taniguchi

13:15 – 14:15 *Lunch*

ABSTRACTS

Pietro Amenta
University of Sannio, Italy

Partial Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic.

Partial correspondence analysis (Yanai, 1986, 1988) has been introduced in statistical literature to eliminate the effects of an ancillary criterion variable on the relationship between two categorical characters. It is well known that partial and classical correspondence analyses do not perform well if one (or both) of the variables forming the contingency table present an ordinal structure. Correspondence analysis of cumulative frequencies is a method that considers the information included in the ordinal variable(s). Nevertheless, in this case, a third categorical variable (ancillary) could also influence the existing relation. In this paper, we extend Yanai's partial approach to correspondence analysis of cumulative frequencies, and by using suitable orthogonal projectors, we obtain some properties. Finally, we present an actual case study.

Aldo Corbellini
University of Parma, Italy

Robust diagnostics and monitoring for Linear Mixed Models

Joint work with F. Laurini and L. Grossi

The resilience of linear mixed models (LMM) with random effects is investigated with the use of the forward search (FS). The constraints imposed by the model and its estimations provide new computing challenges when extending the FS to the LMM. The method is illustrated by looking at coffee shipments to the European Union using real data and looking for anomalies that could point to fraud.
Robust estimators may be utilized to monitor the detrimental effects of outliers and important data on the parameter estimations of these models. Certain current diagnostics suffer from "masking" when the real number of outliers is more than k. Most other diagnostic methods rely on the leave-k-out method. Despite its relatively high processing costs, the forward search (FS) strategy is an alternative "monitoring" technique that has shown to be particularly effective in dealing with the masking effect in numerous multivariate situations due to its high degree of flexibility, speed, and resilience.

Marcella Corduas
University of Napoli Federico II, Italy

Modelling ordinal data from repeated surveys

Business and consumers survey data are the basis for several indicators describing the trend of macro-economic variables that are fundamental for monitoring the overall performance of the economic system. Qualitative surveys typically ask interviewees to express their perceptions or expectations about the current or future tendency of a reference economic variable (such as inflation or industrial output) using a trichotomous or a finer-tuned ordered scale. Surveys are carried out at regular interval by statistical offices, and collected data are traditionally published in aggregate form, reporting the proportions of positive, neutral or negative assessments. This contribution presents an innovative dynamic model that describes the probability distributions of ordered categorical variables observed

over time. For this aim, we extend the definition of the mixture distribution obtained from the Combination of a Uniform and a shifted Binomial distribution (CUB model), introducing time varying parameters. The model parameters identify the main components ruling the respondent evaluation process: the degree of attraction towards the object under assessment, the uncertainty related to the answer, and the weight of the refuge category that is selected when a respondent is unwilling to elaborate a thoughtful judgement. We suggest to use the model time-varying parameters as indicators of the diversity of respondents' opinions, shifting from an optimistic to a pessimistic state as the surrounding conditions evolve. For illustrative purpose, the dynamic CUB model is applied to the consumers' perception and expectations of inflation in Italy to investigate: a) the effect of external shocks on the respondents' perceptions; b) the impact of the respondents' income level on expectations.

Alessio Farcomeni
University of Roma Tor Vergata, Italy

Time-Interaction Point Processes with heterogeneity

Joint work with R. Barone and M. Mezzetti

We define time-interaction point (TIP) processes, that are general point processes where the occurrence of an event can increase or reduce the probability of future events.
We allow for parametric and non-parametric baseline risks. We further generalise TIPs by letting model parameters be modulated by an unobserved discrete-state continuous-time latent Markov process. To facilitate Bayesian inference for TIP processes, we propose a novel and efficient data augmentation method to approximate posteriors.
We illustrate with a simulation study; and an original application to terrorism in Europe in the period 2001-2017. Our analysis reveals the presence of two distinct latent clusters in the hazard of terrorist event occurrences, association with specific covariates, and self-exciting phenomena.

Luca Greco
University Giustino Fortunato, Italy

On testing the equality between interquartile ranges

Joint work with G. Luta and R. Wilcox

The interquartile range is a statistical measure of variability, that accounts for the variability of the central half of the data. It is particularly useful when the distribution of the data is asymmetric or irregularly shaped. The use of the interquartile range is investigated when the main aim is to compare the variability of two distributions using two independent random samples, without the need to make any distributional assumptions. Several techniques are compared through numerical studies and real data examples, with a particular attention given to the use of the Harrel-Davis quantiles estimator and the quantile regression.

Francesca Greselin
University of Milano Bicocca, Italy

Model-based clustering with cellwise outliere detection

In recent years, a great deal of attention has been paid to the accommodation and identification of unusual observations (outliers) in data. Outliers, disrupting the prevailing patterns within a dataset,

are a common occurrence in real-world data rather than an exception. Real-world databases, encompassing all fields, exhibit an estimated encoding error rate of approximately five percent.
Robust methods aim to mitigate the impact of outliers, allowing the remaining data point to shape the inferential results. Existing literature predominantly addresses outliers at the level of entire data rows. The present paper takes a different approach by focusing on the identification of cellwise outliers - individual cells within a data matrix affected by contamination. We presume that the remaining cells in these affected rows retain valuable information, a perspective particularly advantageous in a multivariate setting. We tackle the challenging scenario of unobserved heterogeneity, where distinct patterns exist in different parts of the data, and a global analysis benefits from mixture models. To address this, we introduce a model-based clustering methodology adept at managing missing data and outliers at the cell level. Parameter estimation utilizes an alternating expectation-conditional maximization algorithm, featuring a concentration step for pinpointing contaminated cells. The effectiveness of our proposed approach is demonstrated through applications to both synthetic and real datasets.

Luigi Grossi
University of Parma, Italy

Statistical models for electricity data forecasting

Joint work with F. Laurini

The Forecasting electricity prices is a topic that has been extensively explored over the last two decades, driven by the deregulation of energy markets that began at the end of the previous century. There is a plethora of articles addressing the estimation of models for interpreting time dynamics and predicting electricity prices. The set of regressors used in these models is highly variable, necessitating criteria for the selection and ranking of the most predictive variables.
Machine learning models have recently garnered significant attention in the literature. Several reasons support this choice, with high flexibility and the ability to include a large number of regressors being among the key factors.
While some papers claim the superiority of machine learning methods over simpler linear models in predicting electricity prices, it remains unclear whether the application of very sophisticated black-box models is genuinely motivated by non- linearity tests, and whether their forecasting performances are truly better than those of simpler and more interpretable linear models. This paper seeks to contribute to this literature by comparing the performance of a set of machine learning models based on Neural Networks architecture with that of linear AutoRegressive (AR) models, both with and without LASSO penalization. The dataset used covers many of the most important and frequently studied electricity markets.
The impact of renewable sources in the generation mix of electricity is becoming overwhelming. For this reason, in the final part of the analysis, we have estimated the same set of models presented in the first part of the paper to forecast the electricity produced by agrovoltaic plants. This demonstrates how machine learning models could enhance forecasting performance when the underlying process is non- linear. The forecasting ability of these models could provide significant support in the development of a predictive maintenance framework, which is essential to make this technology economically sustainable.

Yan Liu
Waseda University, Japan

Detection of periodicity in multivariate functional time series

Joint work with R. Sagawa

We propose an information criterion for determining an unknown number of periodic components in functional time series. It has been a focal topic to find the number of frequencies when we obtain a large-scale time series. To achieve this goal, we suggest an iterative procedure based on the periodogram comprised of the residual process obtained by the least squares fitting. This iterative procedure has high applicability. We establish the consistency of the estimated number of periodic components which minimizes the information criterion. In addition, we apply this new procedure to multivariate functional time series. The efficacy of the procedure is illustrated by the numerical simulations. Some real data applications will be also provided in this talk.

Antonio Lucadamo
University of Sannio, Italy

Model based clustering for torus data

Joint work with L. Greco and C. Agostinelli

Torus data are multivariate circular observations that arise as measurements on a periodic scale and are often recorded as angles. In this paper, we focus on parsimonious model-based clustering for torus data by building on the *mclust* methodology. Therefore, covariance constraints are imposed on the completely general heterogeneous clustering model allowing a flexible and general framework to clustering torus data.

Marta Nai Ruscone
University of Genova, Italy

One-dimensional mixture-based clustering for ordinal responses

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows and columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. Mixture-based models are extended to include covariates and test what effect this has on the resulting clustering structures. The focus is on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. The models are fit using the Expectation-Maximization (EM) algorithm and assess their performance. Finally, an application of the models is also illustrated in the well-known arthritis clinical trial data set.

Gianluca Morelli
University of Parma, Italy

High-dimensional data reduction in model based clustering: an application on a retailer's data

Joint work with F. Laurini

Customer segmentation is a well-known strategic asset in the retail domain, but applied clustering methods are often selected for their computational efficiency rather than their accuracy. In this work we present a new approach of model based clustering applied on a large dataset characterized by strong sparsity and cases of strong correlations between variables. These two starting conditions, very typical in real data sets, have a double implication. In fact, the presence of variables containing a strong prevalence of zeros increases calculation times without probably adding useful information to clustering. The presence of correlated variables, however, if not managed with correct methods, inevitably leads to misleading classifications. The goal of this work is to create a ranking of the variables that most actively contribute to the formation of clusters and then remove those that do not have impact on classification. In a second stage, once the most important variables have been detected, we propose a new method to identify the optimal number of groups into which to classify the customers for the final output. The proposed heuristic dimensions reduction method attempts to reconcile the benefits of model based clustering applied to big data keeping down the often prohibitive calculation times.

Ilia Negri
University of Calabria, Italy

Z-Process Method in Change Point Problems for Dependent Observations

The Z-process method was introduced as a general unified approach based on partial estimation functions to construct test statistics for a wide range of statistical change-point problems. This method can simultaneously test for changes in any of the model's parameters. We investigate the asymptotic distribution of the test statistics under the null hypothesis and under a very general alternative hypothesis. Applications of the method to change-point problems in a diverse set of models with dependent observations are also discussed. Specifically, it is applied to diffusion processes, both for ergodic models and for some models with random Fisher information matrices. Furthermore, we address the issue of testing for parameter changes in linear time series models. Finally, some simulated studies are presented.

Simona Pacillo
University of Sannio, Italy

An impartial trimming algorithm for robust circle fitting

Joint work with L. Greco and P. Maresca

Accurate circle fitting can be seriously compromised by the occurrence of even few anomalous points. In the work it is proposed a robust fitting strategy based on the idea of impartial trimming. An iterative algorithm detects the trimmed points simultaneously to parameters estimation. The global robustness properties of the method are established and the finite sample behavior of the proposed estimator has been investigated according to some numerical studies and real data examples.

Stefano Maria Pagnotta
University of Sannio, Italy

An empirical evaluation of gene-set enrichment test statistics if missing a reference ground truth.

Joint work with C. La Torella and D. Risso

Gene set enrichment analysis is a methodology that joins statistical results from genomic data with the biological event associated with the samples. For example, two groups of patients administered with different therapies provide tissues for transcriptome sequencing. After properly statistically comparing the groups with a gene level resolution, the question is uncovering which biological event characterizes each. The solution are enrichment analysis methods. To simplify, these methods assign a p-value to a group of consistent genes.
The first methodology appeared in 1999, as microarray technology allowed the measurement of the mRNA abundance associated with each known gene. Since then, a plethora of procedures have been proposed. In the last 20 years, with the growth of enrichment analysis methodologies, proposals for comparisons of methods appeared in the literature. A general misunderstanding still goes on because the performance of the test statistics is not separate from a) the gene-level summarization of the groups and 2) the source of variability needed to compute the p-value.
This study isolates the test statistics for enrichment analysis from confounding elements and evaluates its performance. The aim is to understand which method captures, at best, the biology in the groups. To reach this primary result, we had to define a new paradigm for comparing the methods if a theoretical ground truth is unavailable.

Domenico Perrotta
Joint Research Centre, European Commission, Ispra, Italy

The Least Trimmed Squares for time series (LTSts)

International organizations need to monitor large amounts of economic and financial data to prevent or uncover potential problems in policies implementation. The analysis of such time series cannot ignore the potential presence of anomalies and structural changes. In this paper, we elaborate on a robust framework based on least trimmed squares analysis, which we extend to treat outliers and points where a change in level takes place in operationally intensive environments requiring accurate and stable outcomes. We study its properties and introduce instruments for its use in concrete cases related to trade policies of major relevance for the European Union.

Domenico Piccolo
University of Naploli Federico II, Italy

An approximate distribution of the dissimilarity index for CUB models

A useful approach to fit and interpret ratings and preferences has been proposed by introducing *CUB* models, a mixture of two discrete distributions aimed to capture the subjective generating process of eliciting a definite choice among an ordered sequence of modalities. Differently from the standard paradigm of cumulative models, these structures do not necessarily require subjects' covariates. Thus, the range of their applicability is wider: from visualization to clustering, from classification to prediction, from composite indicators to time stability, etc.
In this regard, it is important to check the adequacy of the estimated models to empirical data by using the normalized dissimilarity index which compares the relative frequencies of ratings and the

corresponding fitted probabilities fitted by the model: indeed, it expresses the quota of the observed responses to change in order to obtain a perfect fit. Unfortunately, it is difficult to assess an exact distribution for this index and thus inferential issues cannot be advocated.
This work aims to study the behaviour of the sample distribution of an appropriate transformation of the dissimilarity index when data are generated by CUB models. Based on statistical considerations and extensive simulations over the parameter space, approximate quantiles are derived; then, the effect of uncertainty and feeling parameters on the quantiles for varying sample size are examined. In addition, power considerations for testing the adequacy of the fitted model are presented to confirm the substantial goodness of the approach. To make these results operational, some final suggestions conclude the work.

Tommaso Proietti
University of Roma Tor Vergata, Italy

Ups and (Draw) downs

The drawdown measures the potential loss of a financial asset associated with a deviation of the current value from its local historical maximum. It is used to provide measures of market risk, to construct portfolios as well as risk-adjusted measures of performance, and to define trading strategies. The paper aims to characterize the time series properties of the drawdown process and those of related processes, such as the drawup and their duration. They depend on the distribution of multiperiod returns and on the nature of the measurement process. The latter is such that the time lag from the current maxima and minima generates first-order Markov chains, which are homogeneous and ergodic under unrestrictive assumptions on the returns process. The paper also shows how to use the two Markov chains to date turning points and bear and bull market phases. We finally consider time series prediction of future drawdowns and robust estimation in the presence of noise.

Marco Riani
University of Parma, Italy

Applied Robust Statistics through the Monitoring Approach

The goal of the talk is to provide a summary of the forthcoming Springer Verlag book “Applied Robust Statistics through the Monitoring Approach” by Atkinson et al. (2024). This book is about statistics, computing and graphical methods in the service of robust data analysis. The recent development of the computational technique we call monitoring provides an envelopment of many robust methods, allowing the data to choose important parameters of the analysis. The procedures we describe yield highly informative graphical output that allow comparison of a variety of robust methods for each set of data. Robustness is required because data may be contaminated and models may be wrong. Successful data analysis requires methods which reveal any divergence between the data and the fitted model and provide indications of how both may be reconciled. When the data contain outliers, that is observations in disagreement with the model, the methods of traditional robust statistics are intended to give such observations reduced importance in the analysis, either by reducing the weight of the outliers, or by deleting them altogether. Traditional robust statistics was first formalized in the The Princeton Robustness Study of 1973. From that time the robust analysis of data has had something of a black box about it. The box needs some inputs apart from the data. The most important is the proportion of outliers that the data analyst believes is present in the data. Given the

choice of robust procedure, the box then delivers a robust estimate of the parameters, by downweighting or deleting observations during the fitting process. If too few observations are deleted, the parameter estimates will be biased by the outliers. If too many observations are deleted, information is lost and the parameter estimates will be unnecessarily imprecise. Over the last 25 years we have been developing an alternative approach to robust statistics, which we call the Forward Search. The search starts from a small, robustly chosen, subset of the data that excludes outliers. We then move forward through the data, adding observations to the subset used for parameter estimation and removing any that have become outlying due to the changing content of the subset. As we move forward, we monitor statistical quantities, such as parameter estimates, residuals and test statistics. One importance of this approach is that sequential testing for outliers can be used to determine the proportion of outliers in the data and so to provide an empirical estimate of how many observations should be deleted in the robust estimation of the parameters. The realisation exploited in the above mentioned book was that procedures similar to those that were used for the Forward Search could be used to provide non-asymptotic results for other estimators by enumerating their behaviour over a range of values of the expected proportion of outliers in the data. It is thus possible to compare the finite sample performance of many robust estimators.

Silvia Salini
University of Milano Statale, Italy

A compared protocol to improve clustering procedures

In this paper we study a widely used machine learning dimensionality reduction techniques, such as t-SNE (t-Distributed Stochastic Neighbor Embedding), in the presence of outliers and/or inliers, with the purpose to understand whether and how they can be used to improve well-known statistical clustering procedures, such as k-means or t-clust.
t-SNE was introduced by Van der Maaten and Hinton in 2008 (Journal of machine learning research) as an improvement of Stochastic Neighbor Embedding in the context of large datasets. It is a dimensionality reduction technique that emphasizes the preservation of local and global structure in high-dimensional data. It employs a probabilistic approach to model pairwise similarities between data points and minimizes the Kullback-Leibler (KL) divergence between high-dimensional and low-dimensional similarities.

Rosaria Simone
University of Napoli Federico II, Italy

Recent advances in theory and applications of mixture models with uncertainty for rating data

The presentation will survey some of the most recent theoretical and applied advances related to the class of mixture models with uncertainty for ordinal rating data.
The original idea of specifying a discrete model to explain feeling and uncertainty of ordinal evaluations goes back to Piccolo (2003). Over the last 20 years, several methodological extensions and applications have enriched the literature on categorical data in this framework: see Piccolo and Simone (2019) for an updated overview, with discussions and rejoinder. The most recent extensions concern model-based trees for ordinal data to perform classification and to derive response profiles in terms of feeling and possibly different uncertainty components (Cappelli et al., 2019; Simone et al., 2019). In this setting, residuals' diagnostics for ordinal data models (Liu and Zhang, 2018) can be

exploited to implement flexible uncertainty trees with better explicative and predictive performance than classical model-based trees relying on a single maintained model (Simone, 2023).
To summarize, the presentation will highlight the extent by which uncertainty specification improves both explicative and predictive performance of a given preference model assumed for the underlying latent trait (Simone and Piccolo, 2022).

Elena Stanghellini
University of Perugia, Italy

On latent and selection nodes in graphical models for binary variables

I will discuss the distortions induced by marginalization and conditioning on some parameters of interest in systems of binary random variables .Marginalization accounts for unobserved confounders while conditioning accounts for some kind of nonrandom sampling, such as case-control or self-selection. I shall discuss point identification as well as sensitivity analysis. Links to nonparametric modelling will also be made. If time permits, an instance where the introduction of a latent variable is beneficial will also be presented.

Masanobu Taniguchi
Waseda University, Japan

Philosophy of AIC

Joint work with J. Hirukawa

In statistical model selection, we have to infer the order p of parametric models from data. The best known rule for determining the true value p0 of p is probably Akaike's Information Criterion (AIC). In this talk we propose a generalized Akaike's information criterion (GAIC), which includes the usual AIC as a special case, for general class of stochastic models (i.e., i.i.d., non-i.i.d., time series models etc.). We derive the asymptotic distribution of selected order $\hat{p}$ by GAIC, and show that $\hat{p}$ is not consistent (i.e., over estimated).
Next we suppose that the true model g would be incompletely specified, and be "contiguous" to a fundamental parametric model. Under this setting we derive the asymptotic distribution of $\hat{p}$. In comparison with the other criteria, e.g., BIC and HQ, we show that GAIC has admissible properties. That is, we elucidate what AIC aims.

Francesca Torti
Joint Research Centre, European Commission, Ispra, Italy

Robust correspondene Analysis with applications in international trade

Correspondence analysis is a method for the visual display of information from two-way contingency tables. Riani et al. (2022) introduce a robust form of correspondence analysis based on minimum covariance determinant estimation. This leads to the systematic deletion of outlying rows of the table and to plots of greatly increased informativeness. The robust method requires that a specified proportion of the data be used in fitting. To accommodate this requirement we provide an algorithm that uses a subset of complete rows and one row partially, both sets of rows being chosen robustly. The approach has been applied to the analysis of international trade flows.

Weining Wang
York University, United Kingdom

Testing variable importance in high dimensions with deep neural network (NN)

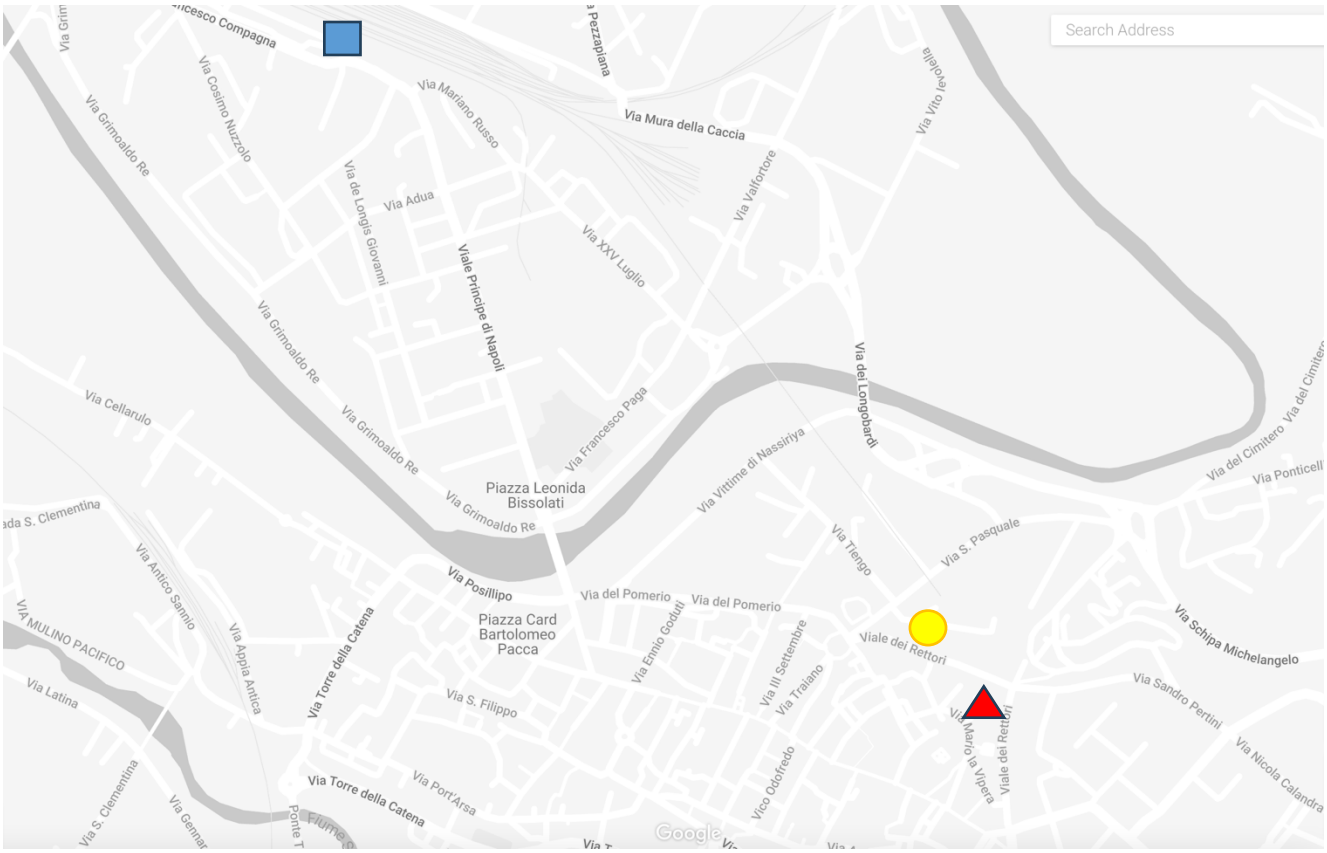
Joint work with Y. Zhao and J. Fan

We present a significance test for any given variable in nonparametric regression with many variables. The test is based on the moment generating function of the partial derivative of an estimator of the regression function, where the estimator is a deep neural network whose structure is allowed to become more complex as the sample size grows.

This test finds applications in model specification and variable screening for high-dimensional data. To render our test applicable to high-dimensional inputs, whose dimensions can also increase with sample size, we make the assumption that the observed high-dimensional predictors can effectively serve as proxies for certain latent, lower-dimensional predictors that are actually involved in the regression function. Additionally, we finely adjust the regression function estimator, enabling us to achieve the desired asymptotic normality under the null hypothesis, as well as consistency for any fixed certain scenarios and certain local alternatives.

PRACTICAL INFORMATION

Benevento map, with some interest places for the conference



TRAIN STATION, PIAZZA VITTORIA COLONNA

HOTEL VILLA TRAIANO, VIALE DEI RETTORI, 9

DEPARTMENT DEMM, PALAZZO DE SIMONE - PIAZZA ARECHI II