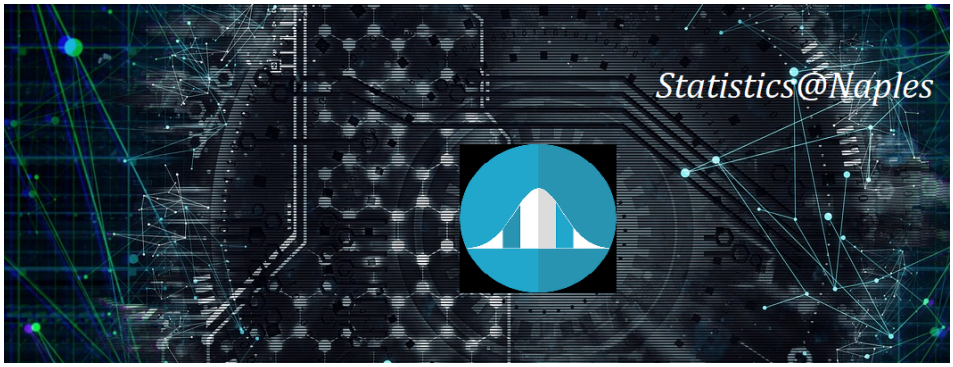
Statistics@Naples



Report of Contributions

https://indico.unina.it/e/statistics-at-naples

Contribution ID : **1** Type : **not specified**

Objective Shrinkage Priors Via Imaginary Data

mercoledì 28 giugno 2023 15:40 (20)

In this work, focus is given in the Bayesian variable selection problem for high-dimensional linear regression problems. The use of shrinkage priors, when the number n of available observations is less than the number p of explanatory variables, is a well-established method, which shares great theoretical and empirical properties. By using imaginary data and shrinkage priors as baseline priors, under the Power-Expected-Posterior (PEP) prior methodology, objective shrinkage priors are being created. In addition, we explore the idea of augmenting the imaginary design matrix in order to make it with orthogonal columns and thus to produce independent PEP-shrinkage priors, based on default baseline priors. Under this setup, properly chosen hyperpriors are placed on the power parameters of the PEP methodology, in order to produced mixtures of independent priors suitable for the variable selection problem when n « p. This second approach provides us with algorithmically flexibility and less time-consuming procedures. We check the theoretical properties of our proposed methods and we explore their behavior via simulated studies.

Keywords: Bayesian Variable Selection; Imaginary Data; Objective Priors; Shrinkage Priors

Primary author(s) : FOUSKAKIS, Dimitris (National Technical University of Athens); TZOUMERKAS, George (National Technical University of Athens)

Presenter(s) : FOUSKAKIS, Dimitris (National Technical University of Athens)

Session Classification : First Session

Contribution ID : **2** Type : **not specified**

MCMC or Reservoir computing? A direct sampling approach

venerdì 30 giugno 2023 11:00 (40)

Assume that we would like to estimate the expected value of a function f with respect to a density π by using an importance density function q. We prove that if π and q are close enough under KL divergence, an independent Metropolis sampler estimator that obtains samplers from π with proposal density q, enriched with a variance reduction computational strategy based on control variates, achieves smaller asymptotic variance than the one from importance sampling. We illustrate our results in challenging option pricing problems that require Monte Carlo estimation. Furthermore, we propose an automatic sampling methodology based on adaptive independent Metropolis that can successfully reduce the asymptotic variance of an importance sampling estimator and we demonstrate its applicability in a Bayesian inference problems.

Primary author(s) : LIU, Siran (UCL); TITSIAS, Michalis (Deep mind); DELLAPORTAS, Petros (AUEB and UCL)

Presenter(s) : DELLAPORTAS, Petros (AUEB and UCL)

Session Classification : Second Plenary Session

Contribution ID : **3** Type : **not specified**

Sparse Pairwise Likelihood Inference for Multivariate Time Series Models

venerdì 30 giugno 2023 09:30 (20)

Multivariate time series data is becoming an increasingly common research topic. Unlike univariate time series, the temporal dependence of a multivariate series includes both serial dependences and interdependences across different marginal series. Consequently, as the number of component series increases, multivariate time series models become overparameterized. In addition, there are many cases where the conditional distribution of the multivariate series given its past might have a complicated form. Given these challenges we develop methodology by replacing the full likelihood function by a pairwise likelihood that only requires the specification of bivariate marginals instead of the multivariate distribution. Clearly, the computational task of maximization of the pairwise likelihood is much simpler than maximization of the full likelihood function but still it poses the problem of combining all estimators. For this purpose, we rely on maximization of an approximate weighted least squares estimation criterion subject to a shrinkage penalty that allows for model selection. The suggested approach provides a general framework for multidimensional time series since it can be applied to both continuous and discrete time series but also to mixed mode time series data.

Primary author(s) : PEDELI, Xanthi (Athens University of Economics and Business, Department of Statistics)

Co-author(s) : FOKIANOS, Konstantinos (University of Cyprus, Department of Mathematics and Statistics); KARLIS, Dimitris (Athens University of Economics and Business, Department of Statistics)

Presenter(s) : PEDELI, Xanthi (Athens University of Economics and Business, Department of Statistics)

Session Classification : Sixth Session

Contribution ID : **4** Type : **not specified**

Subdata selection for big data regression based on leverage scores

mercoledì 28 giugno 2023 15:00 (20)

Data continues to become more abundant, and so the datasets that contain it. Even though big datasets can present insights and opportunities, they can pose significant challenges when it comes to statistical analysis. One of the biggest challenges, required to process and analyze large datasets, is the computational resources. Regression can be problematic in case of big datasets, due to the huge volumes of data. A standard approach is subsampling that aims at obtaining the most informative portion of the big data. We consider an approach based on leverages scores, already existing in the current literature for the selection of subdata for linear model discrimination. However, we highlight its importance on the selection of data points that are the most informative for estimating unknown parameters. We conclude that the approach based on leverage scores improves existing approaches, providing simulation experiments as well as a real data application.

Primary author(s) : CHASIOTIS, Vasilis (Athens University of Economics and Business, Department of Statistics); Prof. KARLIS, Dimitris (Athens University of Economics and Business, Department of Statistics)

Presenter(s) : CHASIOTIS, Vasilis (Athens University of Economics and Business, Department of Statistics)

Session Classification : First Session

Contribution ID : **5** Type : **not specified**

A model-robust subsampling approach in presence of outliers

mercoledì 28 giugno 2023 17:40 (20)

In the era of big data, several sampling approaches are proposed to reduce costs (and time) and to help in informed decision making. Some of these proposals (Drovandi et al., 2017; Wang et al., 2019; Deldossi and Tommasi (2022) among others) are inspired to Optimal Experimental Design and require the specification of a model for the big dataset. This model assumption, as well as the possible presence of outliers in the big dataset represent a limitation for the most commonly applied subsampling criterions. Deldossi et al. (2023) introduced non-informative and informative exchange algorithms to select "nearly" D-optimal subsets without outliers in a linear regression model.

In this study, we extend their proposal to account for model uncertainty. More precisely, we propose a model robust approach where a set of candidate models is considered; the optimal subset is obtained by merging the subsamples that would be selected by applying the approach of Deldossi et al. (2023) if each model was considered as the true generating process. The approach is applied in a simulation study and some comparisons with other subsampling procedures are provided.

Key-words: Active learning, D-optimality, Subsampling

References

Deldossi, L., Tommasi C. (2022) Optimal design subsampling from Big Datasets. Journal of Quality Technology 54(1): 93–101

Deldossi, L., Pesce, E., Tommasi, C. (2023) Accounting for outliers in optimal subsampling methods, Statistical Papers, https://doi.org/10.1007/s00362-023-01422-3.

Drovandi CC, Holmes CC, McGree JM, Mengersen K, Richardson S, Ryan EG (2017) Principles of experimental design for big data analysis. Statistical Sciences 32(3): 385–404

Wang H, Yang M, Stufken J (2019) Information-based optimal subdata selection for Big Data linear regression. Journal of American Statistical Association 114(525): 393–405

Primary author(s) : DELDOSSI, Laura; TOMMASI, Chiara

Presenter(s) : DELDOSSI, Laura

Session Classification : Second Session

Contribution ID : **6** Type : **not specified**

A structural equation model to integrate item responses, response times and item positioning in students' ability assessment

mercoledì 28 giugno 2023 17:20 (20)

In the context of students' ability assessment, considering collateral information in addition to item responses can be helpful in increasing the accuracy of the measurement. In this vein, the evaluation of students' abilities via computer based-devices has made response time data available at the item level (Wang et al., 2019). Besides, the literature (Becker et al., 2022) has highlighted an item position effect when the same items are presented in different positions within multiple test forms.

With the present contribution, we contribute to this research line by proposing a structural equation model (SEM) to jointly consider item responses, response times and item positioning in students' ability assessment. In particular, we assume that the response process is driven by two underlying latent variables: the first latent variable, denoted by Θ_i, represents the ability of individual i that is measured by the test items; the second latent variable, denoted by η_i, refers to the speediness of individual i to answer the test items.

We formulate the statistical model assuming that the item responses are directly affected by the ability Θ_i, whereas the response times depend both on the ability Θ_i and on the speediness η_i. Accordingly, response accuracy tends to increase with the ability level of individual i while response time tends to decrease with the speediness and ability levels. Moreover, we suppose that item positioning affects both item responses and response time. Under this setting, the correlation between Θ_i and η_i is modelled through the cross-relation function that models the relationships between Θ_i and the observed response times.

The empirical application of the proposed model was carried out on first-year Psychology students at the University of Naples Federico II, attending the introductory Statistics course. The test administered was composed of 30 multi-choice questions developed according to three of the five Dublin descriptors: Knowledge (10 items), Application (10 items) and Judgement (10 items). For each question, students' answers were coded as correct (2 credits), partially correct (1 credit) and wrong (0 credits). Data were collected through Moodle platform, which also provided the response time.

References

Becker, B., Van Rijn, P., Molenaar, D., and Debeer, D. (2022). Item order and speededness: Implications for test fairness in higher educational high-stakes testing. Assessment & Evaluation in Higher Education, 47(7):1030–1042.

Wang, C., Weiss, D. J., and Su, S. (2019). Modeling response time and responses in multidimensional health measurement. Frontiers in psychology, 10:51.

Primary author(s) : BACCI, Silvia (Department of Statistics, Computer Science, Applications "G. Parenti"); FABBRICATORE, Rosa (Department of Social Sciences); GALLUCCIO, Carla (Department of Statistics, Computer Science, Applications "G. Parenti"); IANNARIO, Maria (Department of Political Sciences)

Presenter(s) : GALLUCCIO, Carla (Department of Statistics, Computer Science, Applications "G.

Parenti”)

Session Classification : Second Session

Contribution ID : **7** Type : **not specified**

Variable selection via ranking in generalized linear models

mercoledì 28 giugno 2023 16:00 (20)

In many empirical domains, the availability of ultrahigh-dimensional data has led to the development of feature screening and variable selection procedures aiming to detect the informative variables of datasets and consequently remove unimportant features. In this context, we propose a ranking-based variable selection procedure that extends the Ranking Based Variable Selection technique (Baranowski et al., 2020) to general linear regression models. We explore the performance of our proposal using both simulated and empirical data. The algorithm is compared to two competitors: i) the Extended BIC (Chen and Chen, 2012); ii) the variable selection procedure based on the combination of the Sure Independence Screening (Fan and Song, 2010) and the Elastic Net (Zou and Hastie, 2005).

References

Baranowski R, Chen Y, Fryzlewicz P (2020), Ranking-based variable selection for high-dimensional data, Statistica Sinica, 30(3), 1485-1516.

Chen J, Chen Z (2012), Extended BIC for small-n-large-P sparse GLM. Statistica Sinica, 22(2), 555-574.

Fan J, Song R (2010), Sure independence screening in generalized linear models with NP-dimensionality. The Annals of Statistics, 38(6), 3567-3604.

Zou H; Hastie T (2005), Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B, 67(2), 301-320.

Primary author(s) : Prof. GIORDANO, Francesco (Università degli Studi di Salerno); NIGLIO, Marcella (Università degli Studi di Salerno); RESTAINO, Marialuisa

Presenter(s) : NIGLIO, Marcella (Università degli Studi di Salerno)

Session Classification : First Session

Contribution ID : **8** Type : **not specified**

Bayesian networks for complementing and building gender equality composite indicators

mercoledì 28 giugno 2023 17:00 (20)

Composite indicators are a common choice for synthesizing complex phenomena. Over the years, they have grown in popularity and are now applied in many social and environmental sciences. Among others, a subject of increasing interest is gender equality analysis. Gender composite indicators, even if easy to read, may provide a limited picture of the problem. Here we discuss the potentiality of Bayesian networks (BNs) to complement and build composite indicators. BNs are powerful tools for explaining the complex association structure in the dataset and developing scenarios to orient policy-making. Here we propose to use BNs to model the association structure among the gender equality index, its ingredient variables and other context socio-economic variables. In such a way the synergy between composite indicator and BN gives rise to both a monitoring tool for the gender equality gap status and a proactive inferential machine for proposing policies to reduce inequality. BNs can be also used to build the gender equality index, and, in general, any composite indicator. Specifically, we focus attention on an extension of BNs, namely Object-Oriented Bayesian networks (OOBNs). The modularity of the OOBN ensures a computational logic that is consistent with composite indicators, while also providing additional information about the relational structure of variables. An example is carried out on Italian province-level data.

Keywords: composite indicator, gender equality, multivariate dependencies, Object oriented Bayesian networks

References:

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems. Springer Verlag, New York

Musella, F., Vicard, P. (2015). Object-oriented Bayesian networks for complex quality management problems. Quality & Quantity, 49, 115–133

Primary author(s) : VICARD, Paola (University Roma Tre); GIAMMEI, Lorenzo (University of Milan-Bicocca); MUSELLA, Flaminia (Link Campus University); MECATTI, Fulvia (University of Milan-Bicocca)

Presenter(s) : VICARD, Paola (University Roma Tre)

Session Classification : Second Session

Contribution ID : **9** Type : **not specified**

On the predictability of a class of ordinal data models

giovedì 29 giugno 2023 16:00 (20)

The contribution aims at discussing some preliminary results on the evaluation of prediction performance for the class of mixture models with uncertainty (Piccolo and Simone, 2019). The ultimate goal of the analysis is the evaluation of the extent by which the uncertainty specification constitutes an added value for prediction of ordinal scores. A small simulation study is presented to assess prediction performance of competing models under miss-specification. The Ranked Probability Score is chosen as scoring rule since it is the most suited to deal with ordinal data, without the assignment of numerical scores to category. Finally, a variable selection procedure based on prediction performance can be outlined on a case study for the prediction of subjective probability to survive. Comparisons with cumulative link models are illustrated for the sake of completeness. Preliminary findings discussed in Simone and Piccolo (2022) indicate that uncertainty modelling improves prediction performance substantially. Hence, it is important to assess the information quality of the baseline preference model (the Binomial, for instance). To this aim, we introduce a new utility measure for preference models when contaminated with alternative uncertainty specifications in the sense proper to the framework of Information Quality. As a result, the mixing weight of the chosen feeling component within the mixture can be explicitly interpreted in terms of model predictive ability.

Keywords: CUB models; Predictability; Ranked Probability Score; Ordinal Data

References: D. Piccolo, R. Simone (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. STATISTICAL METHOD AND APPLICATIONS, Volume 28, pages 389-435. R. Simone and D. Piccolo (2022). On the predictability of a class of ordinal data models. In A. Balzanella, M. Bini, C. Cavicchia, and R. Verde, editors, Book of short papers SIS 2022, 51st Scientific Meeting of the Italian Statistical Society, pages 1053–1058. Pearson.

Primary author(s) : SIMONE, Rosaria (University of Naples Federico II); DOMENICO , Piccolo

Presenter(s) : SIMONE, Rosaria (University of Naples Federico II)

Session Classification : Fifth Session

Contribution ID : **10** Type : **not specified**

Fast Bayesian Variable Screening Using Correlation Thresholds

giovedì 29 giugno 2023 10:10 (20)

We propose a fast Bayesian variable selection method for Normal regression models, using Zellner's g-prior specification. The approach is based on using thresholds on Pearson and partial correlation coefficients. Nevertheless, the proposed methodology is derived using purely Bayesian arguments derived from thresholds on Bayes factors and posterior model odds. The proposed method can be used to screen out the non-important covariates and reduce the model space size. Then, traditional, computer-intensive, Bayesian variable selection methods can be implemented without any problem with the derived reduced model space. We focus on the g-prior where the Bayes factor computations and the corresponding correlation thresholds are exact. Nevertheless, the approach is general and can be easily extended to any prior setup. The proposed method is illustrated using simulated examples.

Primary author(s) : NTZOUFRAS, Ioannis (AUEB); PAROLI, Roberta (Universita Cattolica del Sacro Cuore)

Presenter(s) : NTZOUFRAS, Ioannis (AUEB)

Session Classification : Third Session

Contribution ID : **11** Type : **not specified**

Latent Feeling and Uncertainty of Perception and Expectations of Price levels over time: A Change Point Analysis

giovedì 29 giugno 2023 13:10 (20)

For the analysis of ordered categorical data, CUB modelling approach entails the estimation of two main structural latent components of the rating process: feeling and uncertainty, parameterized within a two-component mixture of Binomial and uniform distributions: see Piccolo and Simone 2019 for an overview. Featuring parameters can be possibly linked to subject covariates to determine twofold response patterns and they can be promptly estimated using the EM algorithm (as implemented in the R package 'CUB' available on CRAN). The contribution aims at presenting how change point detection of temporal series of estimated feeling and uncertainty can be pursued to identify if and to what extent Italian people modified their perception and judgments of price levels from 1994 to 2019. To this goal, we resort to the framework of Atheoretical Regression Trees (ART, Cappelli et al. 2008) considering the series of monthly response distributions to questions: 1-(Judgments): How do you think the price level changed over the previous 12 months? 2-(Expectations): How do you think the price level will change over the next 12 months? issued by the Italian National Statistical Institute (ISTAT) within the consumers' confidence survey. Responses are collected over a scale with m=5 categories (1 =fall ', 2 =stay about the same', 3 = rise slightly', 4 =rise moderately', 5 = 'rise a lot'). Preliminary results indicate that ART is effective in partitioning the series into sub-intervals characterized by different levels of the estimated model parameters, allowing to study and compare over time, the change points of both feeling and uncertainty. It's worth noticing that the model parameters refer to two different aspects of the respondents' perception and judgment of price level, thus the study of their change points may reveal that they show different number and location of break dates providing a further and valuable insight into the two components of respondents' answers. Performances of ART are also discussed comparatively with those of other techniques for structural change point detection, in particular with respect to Bai and Perron's procedure as ART mimics this procedure. Keywords: price expectation; price judgment; Atheoretical Regression Trees; CUB model; change point detection

References: C. Cappelli, R. N. Penny, W. S. Rea, M. Reale (2008). Detecting multiple mean breaks at unknown points in official time series, MATHEMATICS AND COMPUTERS IN SIMULATION, Volume 78, Issues 2–3, Pages 351-356, ISSN 0378-4754. D. Piccolo, R. Simone (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. STATISTICAL METHOD AND APPLICATIONS, Volume 28, pages 389-435.

Primary author(s) : CAPPELLI, Carmela (University of Naples Federico II); SIMONE, ROSARIA (University of Naples Federico II)

Presenter(s) : CAPPELLI, Carmela (University of Naples Federico II)

Session Classification : Fourth Session

Contribution ID : **12** Type : **not specified**

Modelling ordinal data from repeated surveys

giovedì 29 giugno 2023 09:30 (20)

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 the COVID pandemic on the respondents' perceptions; b) the impact of the respondents' income level on expectations.

Keywords: ordinal data; CUB model; consumers' perceptions; consumers' expectations

References

Corduas, M.: A dynamic model for ordinal time series: An application to consumers' perceptions of inflation. In: *Statistical Learning and Modeling in Data Analysis*, Balzano, S., Porzio, G.C., Salvatore, R., Vistocco, D., Vichi, M. (Eds.), Cham: Springer 2019, (pp. 37-45).

Piccolo, D., Simone, R. The class of cub models: statistical foundations, inferential issues and empirical evidence, (with discussion and rejoinder). *Stat. Meth.\& Appl.* 2019, 28, 389-435.

Primary author(s) : CORDUAS, Marcella (Departmento of Political Sciences, University of Naples Federico II)

Presenter(s) : CORDUAS, Marcella (Departmento of Political Sciences, University of Naples Federico II)

Session Classification : Third Session

Contribution ID : **13** Type : **not specified**

Assessing replication success via skeptical mixture priors

giovedì 29 giugno 2023 10:30 (20)

There is a growing interest in the analysis of replication studies of original findings across many disciplines. When testing a hypothesis for an effect size, two Bayesian approaches stand out for their principled use of the Bayes factor (BF), namely the replication BF (Verhagen and Wakenmalers, 2014) and the skeptical BF (Pawel and Held, 2022). In both cases replication data are used to compare an "advocacy" prior against a benchmark. For the replication BF, the latter is the standard point null hypothesis of no effect while for the skeptical BF it represents the prior of somebody who is unconvinced by the original findings. We propose a novel skeptical mixture prior which incorporates skepticism and limits prior-data conflict. We support our proposal with theoretical results on consistency of the resulting BF, we illustrate its features on an extended example, and we apply it to case studies from the Social Sciences Replication Project.

Keywords: Bayes factor, consistency, prior-data conflict, replication studies.

References:

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3), 879-911.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4), 1457.

Primary author(s) : CONSONNI, Guido (Università Cattolica del Sacro Cuore); EGIDI, Leonardo (Università degli Studi di Trieste)

Presenter(s) : EGIDI, Leonardo (Università degli Studi di Trieste)

Session Classification : Third Session

Contribution ID : **14** Type : **not specified**

Integrating model-based clustering and graphical models to explore the relationship with the digital self-image in (pre)adolescents

venerdì 30 giugno 2023 10:10 (20)

Digital revolution has dramatically changed not only the way people interact but also the relationship with the self-image. Increased data availability and computational power have significantly improved algorithms for facial feature detection which have been also successfully applied to develop face filter apps enhancing and "beautifying" self-portraits. Potential of these filters in altering facial appearance has raised concerns in parents, educators and health professionals as they promote unrealistic beauty standards increasing discrepancy between real and digital self. Actually taking, sharing and viewing edited selfies may have detrimental effects especially on younger users in a developmental phase where they are already facing significant identity construction processes, possibly giving rise to appearance-related cyberbullying. To investigate selfie-sharing/editing behaviour in (pre)adolescents, their relationship with digital self-image, problematic use of social network and possible internalizing symptoms an online questionnaire, including both validated and ad-hoc realized scales, has been developed. When examining the digital-self image, here the attention is narrowed to the face only, the protagonist of real and virtual interactions, and not to the whole body. In this setting, graphical models represent an appealing tool to model dependence structure between collected variables. To properly analyze collected data, the procedure should account for the fact that (i) data from psychological questionnaires are usually measured on discrete/ordinal levels thus violating the normality assumption and (ii) measured behaviors are rarely homogeneous and this heterogeneity should be properly modelled to obtain unbiased results. To tackle these issues, an approach integrating model-based clustering and graphical models (Fop et al., 2019), has been applied to copula transformed data collected on a sample of 229 middle school (pre)adolescents which took part to the online survey. A two-clusters solution was selected as best based on BIC criterion: the two clusters actually showed different covariance network and different management of online self-image and psychological status. Participants in the cluster displaying a worse management of online self-image and psychological status were mainly female reporting higher use of social networks. To better examine the relationships among variables within each cluster, partial correlation networks were estimated separately for the two clusters and compared using both global and local network statistics and inferential procedure for network comparison. Although graphical models have been widely used to model psychological phenomena as complex networks, the application to selfie behavior is original. Moreover, identifying clusters within a graphical model framework has important practical implications such as (i) aiding in the development of tailored training programs suited for improving digital wellbeing in younger users and (ii) uncovering new data-driven relationships among constructs thus generating new hypothesis to test in successive studies.

Reference Fop, M., Murphy, T.B. and Scrucca, L., 2019. Model-based clustering with sparse covariance matrices. Statistics and Computing, 29(4), pp.791-819. Kashihara, J., Takebayashi, Y., Kunisato, Y. and Ito, M. (2021). Classifying patients with depressive and anxiety disorders according to symptom network structures: A Gaussian graphical mixture model-based clustering. Plos one, 16(9), p.e0256902.

Primary author(s) : CUGNATA, Federica (CUSSB, Faculty of Psychology, Vita-Salute San Raffaele University); DI SERIO, Clelia (CUSSB, Faculty of Psychology, Vita-Salute San Raffaele University); BROMBIN, Chiara (CUSSB, Faculty of Psychology, Vita-Salute San Raffaele University)

Presenter(s) : BROMBIN, Chiara (CUSSB, Faculty of Psychology, Vita-Salute San Raffaele University)

Session Classification : Sixth Session

Contribution ID : **15** Type : **not specified**

Sparse Hierarchical Vector Autoregression for Psychopathological Network Estimation from Intensive Longitudinal Data

venerdì 30 giugno 2023 10:30 (20)

The use of networks as a tool for studying complex systems gained popularity in various scientific disciplines. In the past decade, the "network takeover" reached psychology, and networks were utilized to abstract complex psychological phenomena. In psychopathology, a network-based framework known as the *network theory of mental illness*, posits that mental disorders emerge as systems of causally interacting psychopathological symptoms. According to this framework, symptoms and other psychological or sociological factors are nodes in a psychopathological network, and the absence of an edge between two nodes corresponds to a conditional independence relationship. In contrast to other types of networks (e.g., social networks) where the structure is observed, in networks from psychopathology the dependence structure between the nodes is not known a priori and needs to be estimated from data.

Typically, after estimating a psychological network, summary statistics are used to describe its structural properties both at the global and local levels. In the psychological literature, clinical outcomes such as illness severity have been associated with network summary statistics such as network density. The aim of this study is to test i) whether the network density differs across populations of increasing illness severity, and ii) whether local network statistics can be used to identify symptoms that are associated with illness severity. For this purpose, we use intensive longitudinal data from a 90-day diary study called Mapping Individual Routes of Risk and Resilience (MIRORR). The data consists of 8640 observations within $N = 96$ individuals, divided over four subgroups representing different early clinical stages ($n_1 = 25$, $n_2 = 27$, $n_3 = 24$, $n_4 = 20$). Participants in the lowest risk group were randomly selected from the general population in the north of the Netherlands based on their score on the Community Assessment of Psychic Experiences (CAPE) test. Inclusion criteria for the study were: aged between 18 and 35 years, reading and speaking Dutch fluently, being capable of following the research procedures, provide informed consent. Exclusion criteria for participating in the study were: psychotic episodes (current or in the past) according to the Diagnostic and Statistical Manual of Mental Disorders 4 (DSM-4), hearing or visual problems, and pregnancy. Participants were excluded from the study when they missed more than 22 measurements in total or missed five or more measurements in a row. Items in the diary assessment covered a wide range of feelings and (subclinical) psychotic experiences, depression, anxiety, mania, obsessive-compulsive behavior, and anger. Participants were required to complete a digital questionnaire on psychopathological symptoms, emotions, functioning, and stress once a day for 90 consecutive days.

For estimating the network structure for each group of participants, we propose a hierarchical extension of the graphical vector autoregressive (GVAR) model that aims to model the heterogeneity in intensive longitudinal data. The parameters of the proposed hierarchical GVAR model are estimated within a two-step procedure that combines penalized linear mixed models with graphical LASSO (gLASSO). The estimated networks are then used to calculate global and local network statistics, which are compared across groups using statistical tests.

Our results showed that global network statistics such as network density and connectivity do not significantly differ as mental illness becomes more severe. However, we propose the use of local network characteristics such as centrality indices to identify emotions that correlate significantly with increasing illness severity.

Primary author(s) : BALAFAS, Spyros (Vita-Salute San Raffaele University); Prof. MARCO, Grzegorczyk (Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen,); Dr. CUGNATA, Federica (CUSSB, Faculty of Psychology, Vita-Salute San Raffaele University); Dr. BOOIJ, Sanne (Department of Developmental Psychology, University of Groningen); Prof. WIGMAN, Johanna T. W. (Interdisciplinary Center Psychopathology and Emotion regulation (ICPE), University Medical Center Groningen); WIT, Ernst C. (Universita della Svizzera italiana)

Presenter(s) : BALAFAS, Spyros (Vita-Salute San Raffaele University)

Session Classification : Sixth Session

Contribution ID : **16** Type : **not specified**

Power Prior's Weight Parameter elicitation via Bayes Factor-Calibrated p-values

giovedì 29 giugno 2023 12:30 (20)

In recent times, the integration of historical data in the design and analysis of new clinical trials has gained considerable attention, owing to ethical reasons and difficulties encountered in recruiting patients. In the Bayesian framework, the process of informative prior elicitation is widely recognized as a complex and multifaceted undertaking, requiring the careful quantification and synthesis of prior information into an appropriate prior distribution. Hence, there is a pressing need for developing techniques and methods that can facilitate synthesizing and quantifying prior information more effectively and efficiently. Within this context, the concept of *power priors* (Chen and Ibrahim, 2000) has emerged as a popular approach for incorporating historical data into the prior distribution of a treatment effect, in a flexible and controlled manner. The power prior methodology heavily relies on the *weight parameter* δ, ranging between 0 and 1, that is a crucial factor in determining the degree to which the historical data influences the prior distribution, and for which multiple elicitation strategies are available. A modification of the power prior allows a hierarchical prior specification by taking δ as a random quantity

$$\pi\left(\theta, \delta \mid D_0\right) \propto L\left(\theta \mid D_0\right)^{\delta} \pi_0(\theta) \pi_0(\delta),$$

where D_0 is an historical dataset with corresponding likelihood $L(\theta \mid D_0)$, $\pi_0(\theta)$ and $\pi_0(\delta)$ are the initial priors for θ and δ, respectively. Furthermore, a significant benefit of incorporating a normalizing factor in the power prior methodology is its adherence to the likelihood principle, as demonstrated by the joint normalized power prior

$$\pi\left(\theta, \delta \mid D_0\right) = \frac{L(\theta|D_0)^{\delta} \pi_0(\theta) \pi_0(\delta)}{\int_{\Theta} L(\theta|D_0)^{\delta} \pi_0(\theta) d\theta}.$$

Consequently, in a fully Bayesian approach, the ability to effectively elicit an appropriate initial prior distribution for the weight parameter δ is a crucial step. As far as we know from reviewing the existing literature, a comprehensive justification underlying the choice of a Beta distribution with fixed hyper-parameters, that is an usual choice for this framework, is pretty vague.

The Bayes factor (BF) constitutes a valuable statistical tool for model comparison; however, we explore the use of the Bayes Factor to discriminate between competing models that incorporate distinct initial Beta prior distributions for the weight parameter by exploiting some BF p-value calibration techniques (Garcia-Donato and Chen, 2005). This would enable the selection of candidate models based on a more accurate and reliable assessment of the available evidence, thereby enhancing the validity and robustness of statistical inference.

Keywords: Beta distribution, Clinical trial, Historical information, Robust selection.

References

1. Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. Statistical Science, 15(1):46 – 60.
2. Garcia-Donato, G. and Chen, M.-H. (2005). Calibrating Bayes factor under prior predictive distributions. Statistica Sinica, 15(2):359–380.

Primary author(s) : MACRÌ DEMARTINO, Roberto (University of Padova)

Co-author(s) : Prof. EGIDI, Leonardo (University of Trieste); Prof. TORELLI, Nicola (University of Trieste)

Presenter(s) : MACRÌ DEMARTINO, Roberto (University of Padova)

Session Classification : Fourth Session

Contribution ID : **17** Type : **not specified**

Bayesian MANOVA for the combined evaluation of handwriting evidence

mercoledì 28 giugno 2023 15:20 (20)

Forensic science is a broad field that uses scientific principles and technical methods to help with the evaluation of evidence in legal proceedings of criminal, civil, or administrative nature. Forensic scientists examine recovered traces that can be given by glass fragments, fingerprints, body fluids, textile fibers, digital device data and handwriting. The handwriting examination is a well-known field of analysis. Consider a case involving a handwritten document of unknown origin. Handwritten features extracted from this questioned document will be compared to those extracted from a document written by a person that is suspected of being at the origin of the anonymous document. The propositions of interest are the following:

H_p: the suspect is the author of the manuscript;

H_d: the suspect is not the author of the manuscript.

Handwriting individualization is still largely dependent on the experience of the document examiner, though studies have been conducted with the aim of supporting handwriting examiners to reduce the degree of subjectivity of their expertise. Marquis et al. (2005) proposed to increase the degree of objectivity of handwriting analyses by implementing elements of Fourier analysis in order to describe the contour shape of loops of characters. Specifically, the characters that contain loops can be described by means of Fourier descriptors, which can be used to characterize the shape complexity and other geometric attributes. The analyses conducted showed that these features have a good discriminating power. With the aim of implementing the use of these handwriting features for handwriting identification, Bozza et al. (2008) proposed a Bayesian probabilistic approach by modeling the data with multivariate Normalinverse-Wishart distribution (NIW). The value of the evidence is subsequently assessed by means of the Bayes factor, which can be interpreted as a measure of the strength of support provided by the evidence in favor of the hypothesis H_p against the hypothesis H_d. This approach was accomplished to take into account the correlation between variables, the variability between-writers and within-writer variability. However, the above model is implemented separately for each different type of handwritten character. This can be problematic because it can lead to a different conclusion depending on the type of character that is retained.

In this research, it is proposed the implementation of a Bayesian Multivariate Analysis of Variance (MANOVA) via using the loop characters as predictors. The indicator of the loop character is transformed into a dummy variable (corner-point representation), so that it is possible to model variables describing the handwriting characters jointly taking into account the variability between characters, the variability between-writers for every character and within-writer variability. The Bayesian MANOVA is compared with the two-level random effect model (NIW) proposed by Bozza et al. (2008), that is implemented by modelling all characters jointly or separately. Three different methods for estimating the marginal likelihoods are used; the Generalized Harmonic Mean, the Laplace-Metropolis and the Bridge Sampling. Finally, the performances of the NIW and MANOVA models are compared with those of an alternative one, where a conjugate approach is chosen. This does not allow to model the within and between variation separately, but the marginals can be obtained analytically.

Firstly, we estimate the Bayes factor of the two data models for each writer to determine which model is more compatible with the data. Secondly, there have been selected handwriting features originating from the same writer or from different writers to evaluate the rate of false negatives (that is cases where the BF is smaller than one for characters originating from the same source) and false positives (that is cases where the BF is greater than one for characters originating from differ-

ent sources). Finally, the sensitivity of models is examined in two critical aspects: the misleading background information and the choice of degrees of freedom for the Wishart-inverse distribution that is used to model the handwriting variability. With reference to the misleading background information, the prior distributions were elicited by selecting writers characterized by either small or marked differences.

Keywords : 1. Handwriting Evidence 2. Fourier Analysis 3. Multivariate Bayesian Modelling 4. Bayes Factor 5. Sensitivity

References

Bozza, S., Taroni, F., Marquis, R. & Schmittbuhl, M. (2008), 'Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship', Journal of the Royal Statistical Society: Series C (Applied Statistics) 57(3), 329–341.

Marquis, R., Schmittbuhl, M., Mazzella, W. D. & Taroni, F. (2005), 'Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of the capital character o', Forensic Science International 150(1), 23–32.

Primary author(s) : TZAI, Lampis

Co-author(s) : Prof. TARONI, Franco (University of Lausanne); Mr. NTZOUFRAS, Ioannis (Athens University of Economics and Business); Prof. BOZZA, Silvia (University of Lausanne and University "Ca' Foscari" of Venice)

Presenter(s) : TZAI, Lampis

Session Classification : First Session

Contribution ID : **18** Type : **not specified**

Bayesian effect measures for a location scale model

giovedì 29 giugno 2023 15:40 (20)

We consider a Bayesian approach for the analysis of rating data when a scaling component is taken into account, thus incorporating a specific form of heteroskedasticity. Our approach includes model-based probability effect measures that enable comparisons of distributions among multiple groups. These effect measures are adjusted for explanatory variables that have an impact on both the location and scale components. To estimate the parameters of our fitted model and derive the associated effect measures, we employ Markov Chain Monte Carlo techniques. Through an analysis of students' evaluations of a university curriculum counselor service, we assess the performance of our method and highlight its valuable support in the decision-making process. Our findings demonstrate the effectiveness of our approach and emphasize its ability to enhance decision-making processes by providing valuable insights and support to stakeholders involved.

Primary author(s) : TARANTOLA, Claudia (University of Pavia)

Co-author(s) : IANNARIO, Maria (University of Naples Federico II); KATERI, Maria (RWTH AACHEN)

Presenter(s) : TARANTOLA, Claudia (University of Pavia)

Session Classification : Fifth Session

Contribution ID : **19** Type : **not specified**

Capturing Correlated Clusters Using Mixtures of Latent Class Models

giovedì 29 giugno 2023 13:30 (20)

Latent class models rely on the conditional independence assumption, i.e., it is assumed that the categorical variables are independent given the cluster memberships. Within the Bayesian framework, we propose a suitable specification of priors for the latent class model to identify the clusters in multivariate categorical data where the independence assumption is not fulfilled. Each cluster distribution is approximated by a latent class model, leading overall to a mixture of latent class models. The Bayesian approach allows to identify the clusters and fit their cluster distributions using a one-step procedure. We provide suitable estimation and inference methods for the mixture of latent class models and illustrate the performance of this approach on artificial and real data.

Keywords: Bayesian inference, model-based clustering, prior on the number of components, telescoping sampler.

Fop, M., K. M. Smart, and T. B. Murphy (2017). Variable selection for latent class analysis with application to low back pain diagnosis. The Annals of Applied Statistics 11 (4), 2080-2110.

Fruehwirth-Schnatter, S., G. Malsiner-Walli, and B. Gruen (2021). Generalized mixtures of finite mixtures and telescoping sampling. Bayesian Analysis 16 (4), 1279–1307.

Malsiner-Walli, G., S. Fruehwirth-Schnatter, and B. Gruen (2017). Identifying mixtures of mixtures using Bayesian estimation. Journal of Computational and Graphical Statistics 26 (2), 285–295.

Primary author(s) : MALSINER-WALLI, Gertraud (Vienna University of Economics and Business); GRUEN, Bettina (Vienna University of Economics and Business); FRUEHWIRTH-SCHNATTER, Sylvia (Vienna University of Economics and Business)

Presenter(s) : MALSINER-WALLI, Gertraud (Vienna University of Economics and Business)

Session Classification : Fourth Session

Contribution ID : **20** Type : **not specified**

Regularized Joint Mixture Models

giovedì 29 giugno 2023 15:00 (20)

Regularized regression models are well studied and, under appropriate conditions, offer fast and statistically interpretable results. However, large data in many applications are heterogeneous in the sense of harboring distributional differences between latent groups. Then, the assumption that the conditional distribution of response Y given features X is the same for all samples may not hold. Furthermore, in scientific applications, the covariance structure of the features may contain important signals and its learning is also affected by latent group structure. We propose a class of mixture models for paired data (X, Y) that couples together the distribution of X (using sparse graphical models) and the conditional Y | X (using sparse regression models). The regression and graphical models are specific to the latent groups and model parameters are estimated jointly (hence the name "regularized joint mixtures"). This allows signals in either or both of the feature distribution and regression model to inform learning of latent structure and provides automatic control of confounding by such structure. Estimation is handled via an expectation-maximization algorithm, whose convergence is established theoretically. We illustrate the key ideas via empirical examples. An R package is available at https://github.com/k-perrakis/regjmix.

Keywords: distribution shifts, heterogeneous data, joint learning, latent groups, mixture models, sparse regression

Primary author(s) : PERRAKIS, Konstantinos (Department of Mathematical Sciences, Durham University); LARTIGUE, Thomas (Aramis Project Team, Inria & Center of Applied Mathematics, CNRS, Ecole Polytechnique, IP Paris); DONDELINGER, Frank (Lancaster Medical School); MUKHERJEE, Sach (German Center for Neurodegenerative Diseases, Bonn, Germany & MRC Biostatistics Unit, University of Cambridge)

Presenter(s) : PERRAKIS, Konstantinos (Department of Mathematical Sciences, Durham University)

Session Classification : Fifth Session

Contribution ID : **21** Type : **not specified**

Mixture Models for Repeatedly Measured Survey Data

giovedì 29 giugno 2023 11:00 (40)

Survey data items are commonly collected on a Likert scale and may have an additional "don't know" category. It is also typical to have questions that are not applicable to some individuals or to observe floor or ceiling effects on ordinal or interval responses. These situations necessitate the use of mixture models to properly account for the structure of the data. The model formulation also needs to account for correlations among repeated measures within individual. We present a couple of mixture models with random effects for such situations. In particular, we use logistic sub-models to handle "don't know", inapplicable or floor effects and appropriate generalized linear sub-models for the remaining data. Correlated random effects link the sub-models together. For illustration we use data from tobacco surveys. Maximum likelihood estimation methods are used for model fitting and inference. The software implementation is using PROC NLMIXED in SAS. Simulation studies evaluate bias and efficiency of the parameter estimates.

Primary author(s) : GUEORGUIEVA, Ralitza (Yale University); BUTA, Eugenia (Yale University)

Presenter(s) : GUEORGUIEVA, Ralitza (Yale University)

Session Classification : First Plenary Session

Contribution ID : **22** Type : **not specified**

φ-Divergence based Modelling of Categorical and Rank Data

giovedì 29 giugno 2023 09:50 (20)

Standard models for categorical and ordinal data, such as log-linear, association models and logistic regression models for binary or ordinal responses, as well as the Mallows model for rank data are revisited and defined through statistical information theoretic properties in terms of the Kullback–Leibler (KL) divergence. In the sequel, replacing the KL by the φ-divergence, which is a family of divergences including the KL as special case, these models are generalized to flexible families of models. The suggested models are discussed in terms of their properties, estimation and fit. Finally, their potential is illustrated by characteristic examples.

Key-words: Cressie–Read power divergence, distance-based probability models, maximum likelihood estimation

Primary author(s) : KATERI, Maria (RWTH Aachen University)

Presenter(s) : KATERI, Maria (RWTH Aachen University)

Session Classification : Third Session

Contribution ID : **23** Type : **not specified**

Bayesian learning of network structures from interventional experimental data

venerdì 30 giugno 2023 09:50 (20)

Directed Acyclic Graphs (DAGs) provide an effective framework for learning causal relationships among variables given multivariate observations. Under pure observational data, DAGs encoding the same conditional independencies cannot be distinguished and are collected into Markov equivalence classes. In many contexts however, observational measurements are supplemented by interventional data that improve DAG identifiability and enhance causal effect estimation. We propose a Bayesian framework for multivariate data partially generated after stochastic interventions. To this end, we introduce an effective prior elicitation procedure leading to a closed-form expression for the DAG marginal likelihood and guaranteeing score equivalence among DAGs that are Markov equivalent post intervention. Under the Gaussian setting we show, in terms of posterior ratio consistency, that the true network will be asymptotically recovered, regardless of the specific distribution of the intervened variables and of the relative asymptotic dominance between observational and interventional measurements. We validate our theoretical results in simulation and we implement on both synthetic and biological protein expression data a Markov chain Monte Carlo sampler for posterior inference on the space of DAGs.

Primary author(s) : PELUSO, Stefano

Co-author(s) : CASTELLETTI, Federico

Presenter(s) : PELUSO, Stefano

Session Classification : Sixth Session

Contribution ID : **24** Type : **not specified**

Maximum likelihood estimation of multivariate regime switching Student-t copula models

giovedì 29 giugno 2023 12:50 (20)

We propose a novel estimation method for multivariate regime switching models based on a Student-t copula function. These models account for the interdependencies between multiple variables by considering the correlation strength controlled by specific parameters. Moreover, they address fat-tailed distributions through the number of degrees of freedom. These parameters, in turn, are governed by a latent Markov process.

We consider a two-steps procedure carried out through the Expectation-Maximization algorithm to estimate model parameters by maximum likelihood. The primary computational challenge lies in estimating both the matrix of dependence parameters and determining the number of degrees of freedom for the Student t-copula. To address this, we introduce a new approach that leverages Lagrange multipliers, simplifying the estimation process.

Through a comprehensive simulation study, we demonstrate that our estimators possess desirable properties in finite samples. Additionally, the estimation procedure shows good computational efficiency.

We apply our model to analyze the log-returns of five different cryptocurrencies. The results enable us to identify distinct bull and bear market periods based on the intensity of correlations observed between the crypto assets. This finding highlights the model's efficacy in capturing and characterizing market dynamics within the cryptocurrency domain.

Keywords: statistical models for financial analysis, cryptocurrencies, time series, Expectation-Maximization algorithm, latent variable models

Primary author(s) : CORTESE, Federico (Università degli studi di Milano-Bicocca); PENNONI, Fulvia (Università degli studi di Milano-Bicocca); BARTOLUCCI, Francesco (Università di Perugia)

Presenter(s) : CORTESE, Federico (Università degli studi di Milano-Bicocca)

Session Classification : Fourth Session

Contribution ID : **25** Type : **not specified**

How to peel the network: an algorithm for weighted triad census

venerdì 30 giugno 2023 12:30 (20)

In Network Analysis, the interaction between three nodes is called a "triad" and represents the minimal group structure that can be observed. According to the presence and the type of the relations between three nodes, sixteen triadic configurations (called the isomorphism classes) are defined and their distribution is denoted as "triad census". This kind of analysis is used in different situations concerning relational data and the conventional approach is well-defined for unweighted networks. As a consequence, the information regarding the weights is not taken into account. To exploit this information in the triad analysis, we propose a new algorithm denoted as "network peeling" to count the different configurations of triads in weighted networks. The algorithm computes the triad census over the network layers generated at each step. The resulting matrix (with dimensions layers x isomorphism classes) can be summarized through a set of descriptive measures representing the weighted triad census. With the aim to highlight the appropriateness of our approach, we consider some real scenarios and a simulation study, comparing weighted and conventional triad censuses.

Primary author(s) : RONDINELLI, Roberto; IEVOLI, Riccardo (University of Ferrara); PALAZZO, Lucio (University of Naples Federico II)

Presenter(s) : RONDINELLI, Roberto

Session Classification : Seventh Session

Contribution ID : **26** Type : **not specified**

Semiparametric regression for competing risks data with missing not at random cause of failure

giovedì 29 giugno 2023 15:20 (20)

The cause of failure in cohort studies that involve competing risks is frequently incompletely observed. Failure to deal with this issue can lead to substantially biased estimates. To the best of our knowledge, all the methods that have addressed the issue in the context of semiparametric competing risks models rely on a missing at random (MAR) assumption. Nevertheless, the MAR assumption is not realistic in many real-world settings. In this work we relax the latter assumption by allowing for a class of missing not at random (MNAR) mechanisms, which contain the MAR mechanism as a special case. Due to the inherent non-identifiability issues under MNAR, we propose an approach for hypothesis testing that does not require the estimation of the non-estimable parameters. Using modern empirical process theory, we show that the proposed estimators are uniformly consistent under the assumed class of MNAR mechanisms. We also show that our estimators converge weakly to tight zero mean Gaussian processes and propose rigorous methodology for the computation of confidence intervals which achieve a coverage rate of at least 100*(1 - α)%, asymptotically, for the true unknown parameters of interest. The proposed methodology is applied to competing risks data from a large multicenter HIV study in sub-Saharan Africa where a substantial portion of causes of failure is missing not at random.

Primary author(s) : BAKOYANNIS, Giorgos (Athens University of Economics and Business); YIANNOUTSOS, Constantin T. (Indiana University)

Presenter(s) : BAKOYANNIS, Giorgos (Athens University of Economics and Business)

Session Classification : Fifth Session

Contribution ID : **27** Type : **not specified**

Network Integration with INet algorithm

venerdì 30 giugno 2023 12:50 (20)

Nowadays, network data integration is a demanding problem and still an open challenge, especially when dealing with large datasets. When collecting several data sets and heterogeneous data types on a given phenomenon of interest, the individual analysis of each data set will give only a particular view of such phenomenon. In contrast, integrating all the data will widen and deepen the results giving a more global view of the entire system.

We developed a novel statistical method named INet algorithm, for data integration based on weighted multilayer networks. Under the assumption that the structure underneath the different layers has some similarity that we want to emerge in the integrated network, we generate a "consensus network" through an iterative procedure based on structure comparison, capable of pulling out important information about the phenomenon under study. The procedure tries to preserve common higher-order structures of the original networks in the integrated one, i.e. neighborhood. Once obtained the consensus network, we compared it with the starting networks extracting "specific networks", one for each layer, containing peculiar information of the single data type not present in all the others.

We tested our method on simulated networks to analyze the performance of our algorithm and we analyzed virus and vaccine gene co-expression networks to better understand infectious diseases.

Primary author(s) : POLICASTRO, Valeria; Prof. MAGNANI, Matteo (Uppsala University); Dr. ANGELINI, Claudia (CNR); Dr. CARISSIMO, Annamaria (CNR)

Presenter(s) : POLICASTRO, Valeria

Session Classification : Seventh Session

Contribution ID : **28** Type : **not specified**

An approach to structural equation modeling in a multiblock framework

venerdì 30 giugno 2023 13:10 (20)

In many application fields, the variables used to measure a phenomenon are gathered into homogeneous blocks that measure partial aspects of the phenomenon. For example, in sensory analysis, the overall quality of products may depend on the taste and odor variables, etc. In consumer analysis, consumer preferences may depend on physical-chemical and sensory variables. In some contexts, a structure of relations between the different blocks may exist that gives rise to a chain of influences. Within each link, the blocks of predictor variables are called input blocks, while the block of dependent variables is called the output block. If the input blocks do not depend on any other block, then they are defined as exogenous blocks, while those that rely on other input blocks in the same relation are called intermediate blocks. If there is a chain of the relationship between the blocks, we are then dealing with what is often called a mediation model and must interpret both indirect and direct effects among blocks. Within the scope of multiblock data analysis with a directional path among the blocks, we will present a new approach named SO-PLS path modelling (SO-PLS-PM). The approach splits the estimation into separate sequential orthogonalized PLS regressions (SO-PLS) for each output block. The new method is flexible and graphically oriented and allows for handling multidimensional blocks and diagnosing missing paths. New definitions of total, direct, indirect, and additional effects in terms of explained variances will be proposed, along with new methods for graphical representation. In this research, some interesting properties of the method will be shown both on simulated and real data. The actual data concerns consumer, sensory and process modelling data. Results will also be compared to those of alternative path modelling methods.

Keywords: path analysis, graphical modelling, multiblock regression

References R. Romano, O. Tomic, K.H. Liland, A. Smilde, T. Næs (2019). A comparison of two PLS-based approaches to structural equation modeling. Journal of Chemometrics, 33 (3), e3105. T. Næs, R. Romano, O. Tomic, I. Måge, A. Smilde, K.H. Liland. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. Journal of Chemometrics, 35 (10), e3243.

Primary author(s) : ROMANO, Rosaria (University of Naples Federico II)

Presenter(s) : ROMANO, Rosaria (University of Naples Federico II)

Session Classification : Seventh Session