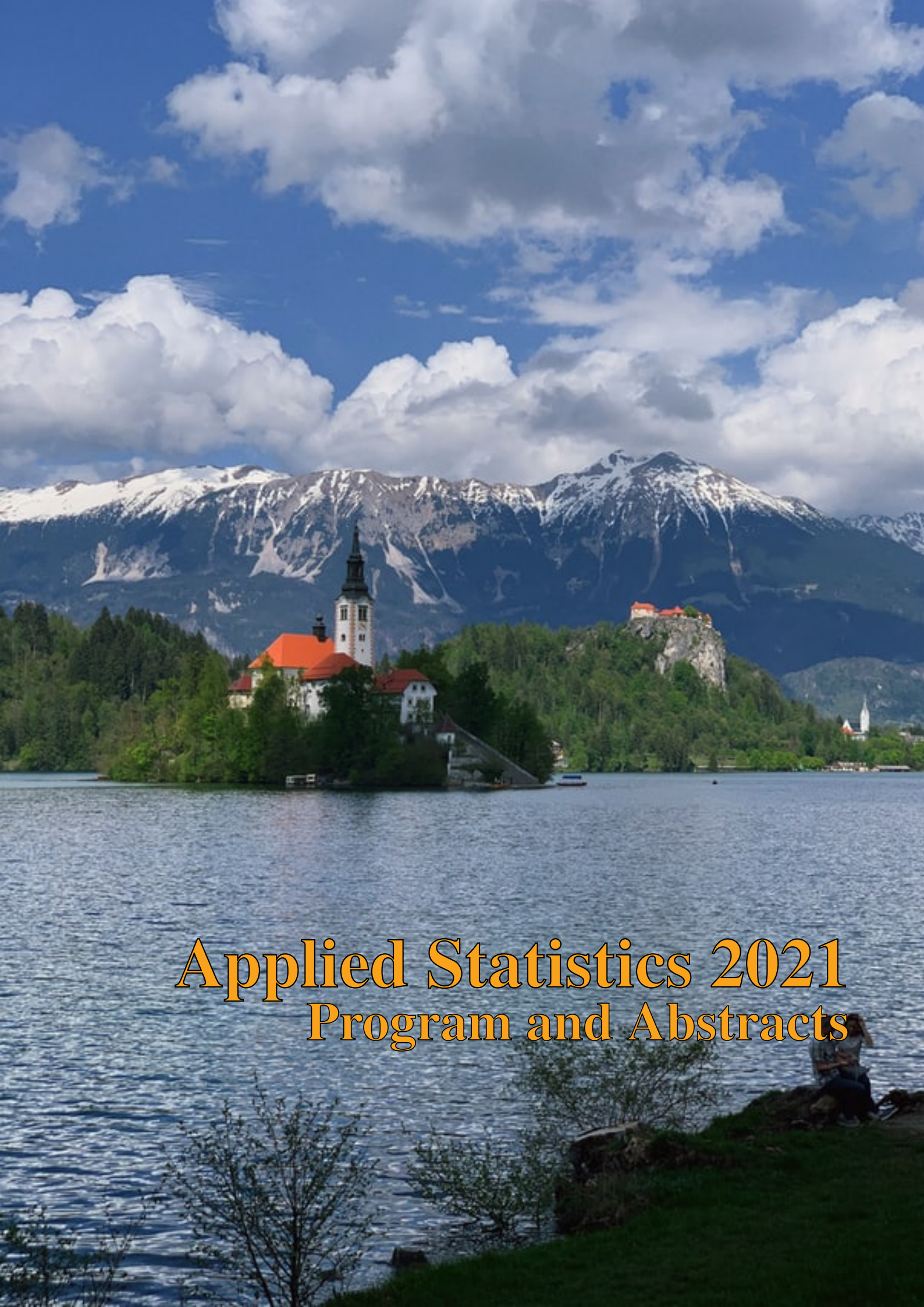

Applied Statistics 2021
Program and Abstracts

International Conference

APPLIED STATISTICS 2021

ABSTRACTS and PROGRAM

2021

Slovenia

https://stat-d.si/applied-statistics-conference/as2021

Organized by

Statistical Society of Slovenia

Scientific Program Committee

Lara Lusa (Chair), Slovenia
Vladimir Batagelj, Slovenia
Andrej Blejec, Slovenia
Maurizio Brizzi, Italy
Herwig Friedl, Austria
Katarina Košmelj, Slovenia
Stanislaw Mejza, Poland
Tamas Rudas, Hungary
Janez Stare (Scientific advisor), Slovenia
Mihael Perman (Scientific advisor), Slovenia
Matevž Bren, Slovenia
Anuška Ferligoj, Slovenia
Dario Gregori, Italy
Irena Križman, Slovenia
Jože Rovan, Slovenia
Vasja Vehovar, Slovenia

Organizing Committee

Irena Vipavc Brvar (Chair)
Andrej Blejec
Jerneja Čuk
Vanja Erčulj
Bogdan Grmek
Andrej Kastrin
Lara Lusa
Ana Zalokar

Edited by: Lara Lusa and Andrej Kastrin
Published by: Statistical Society of Slovenia
Litostrojska c. 54
1000 Ljubljana, Slovenia
Publication format: eBook
URL: `https://akastrin.si/as-book-2021.pdf`
Produced using: `generbook` R package
Front cover: Photo by `Kannidta Keawmontree` on `Unsplash`

ABSTRACTS and PROGRAM

PROGRAM

		Room 1	Room 2
Monday	9.10 – 10.00	Invited lecture	
	10.00 – 10.30	Break	
	10.30 – 12.00	Invited session	Actuarial & financial statistics
	12.00 – 14.00	Break	
	14.00 – 15.30	Young Statisticians Europe	Statistical applications I
	15.30 – 16.00	Break	
	16.00 – 17.00	Biostatistics & bioinformatics	
	17.00 – 17.30	Break	
	17.30 – 19.00	Invited session	Modeling & simulation
	19.00 – 19.30	Break	
	19.30 – 21.30	Social event	
Tuesday	9.00 – 10.00	Invited lecture	
	10.00 – 10.30	Break	
	10.30 – 12.00	Invited session	Statistical applications II
	12.00 – 14.00	Break	
	14.00 – 15.30	Invited session	Network analysis
	15.30 – 16.00	Break	
	16.00 – 17.00	Invited lecture	
Wednesday	10.00 – 16.00	Workshop	

9.10–10.00 **Invited lecture**
Room 1 *Chair: Andrej Blejec*

1. **Clustering of attribute and/or relational data**
   Anuška Ferligoj

10.00–10.30 **Break**

10.30–12.00 **Invited session**
Room 1 *Chair: Anuška Ferligoj*

1. **Compositional data analysis of high-dimensional biological datasets: A revalidation of the additive logratio transformation**
   Michael Greenacre, Marina Martínez-Álvaro and Agustín Blasco
2. **Three approaches to supervised learning for compositional data with pairwise logratios**
   Germà Coenders and Michael Greenacre
3. **Combinatorial regression in abstract simplicial complexes**
   Andrej Srakar and Miroslav Verbič

10.30–12.00 **Actuarial & financial statistics**
Room 2 *Chair: Ana Zalokar*

1. **Consistently recovering the signal from noisy functional data**
   Siegfried Hörmann and Fatima Jammoul
2. **Modelling the polynomial time trend through spline function: A Bayesian procedure**
   Varun Agiwal
3. **Variable selection for mixtures of regression models with random effects**
   Luísa Novais and Susana Faria
4. **Estimation of multicomponent stress-strength reliability for unit Burr-XII distribution**
   Fatma Gul Akgul

12.00–14.00 **Break**

14.00–15.30 **Young Statisticians Europe**
Room 1 *Chair: Andrej Srakar*

14.00–15.30 **Statistical applications I**
Room 2 *Chair: Nataša Kejžar*

1. **Complex hypothesis testing on circular economy**
Stefano Bonnini, Michela Borghesi and Getnet Melak Assegie

2. **A detailed statistical analysis of COVID-19 worldwide effects on economic, social and health welfare**
Maurizio Brizzi and Diletta Cecilia Canini

3. **Spatial non-stationarity in the determinants of land use in Campania (southern Italy) based on the GWR model**
Gennaro Punzo, Rosalia Castellano and Emma Bruno

4. **Mixed models for anomaly detection in aggregate anti-money laundering reports**
Marianna Siino and Stefano Iezzi

5. **Applying multivariate statistical process control for mixed data to prosthetic rehabilitation after lower-limb amputation**
Gaj Vidmar, Neža Majdič and Helena Burger

15.30–16.00 **Break**

16.00–17.00 **Biostatistics & bioinformatics**
Room 1 *Chair: Herwig Friedl*

1. **The importance of imperfect detection in biological data: Large-scale climate effects meet an Amazonian butterfly**
Maja Kajin, Carla Maria Penz and Phil De Vries

2. **Years life difference compared to the general population**
Damjan Manevski, Maja Pohar Perme and Nina Ružić Gorenjec

3. **Comparison of clustering methods for diabetic kidney disease patients formalized through category theory**
Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli

17.00–17.30 **Break**

17.30–19.00 **Invited session**
Room 1 *Chair: Marianne Huebner*

1. **Managing research data for transparency and reusability**
Scout Calvert

2. **Data science research ethics and the challenges of inference, public data and consent**
   Jacob Metcalf

3. **Good data science practice: Moving towards a code of practice for drug development**
   Mark Baillie

17.30–19.00 **Modeling & simulation**
Room 2 *Chair: Mihael Perman*

1. **Estimating the conditional distribution in functional regression problems**
   Thomas Kuenzer, Siegfried Hörmann and Gregory Rice

2. **Univariate goodness-of-fit tests for randomly censored data: tests' adaptation versus data transformation**
   Marija Cuparić and Bojana Milošević

3. **Robust mixture regression modeling for heterogeneous data sets**
   Fatma Zehra Doğru and Olcay Arslan

4. **New class of goodness-of-fit tests based on independence-type characterizations**

   Katarina Halaj, Bojana Milošević, Marko Obradović and Maria Dolores Jiménez-Gamero

5. **Modeling complex histograms**
   Herwig Friedl

6. **Determining factor impacting electronic fitness tracker usage for health and wellness management via predictive analytics**
   Sinjini Mitra

19.00–19.30 **Break**

19.30–21.30 **Social event**
Chair: Andrej Srakar

9.00–10.00 **Invited lecture**
Room 1 *Chair: Janez Stare*

1. **On censoring (with a nod towards causality)**
   Jan Beyersmann

10.00–10.30 **Break**

10.30–12.00 **Invited session**
Room 1 *Chair: Aleš Žiberna*

1. **Blockmodeling dynamic networks: A Monte Carlo simulation study**
   Marjan Cugmas and Aleš Žiberna
2. **Disentangling homophily, community structure and triadic closure in networks**

   Tiago Peixoto
3. **Generalized direct blockmodeling of large valued networks**
   Carl Nordlund

10.30–12.00 **Statistical applications II**
Room 2 *Chair: Matevž Bren*

1. **Spatial statistical modeling of air pollution**
   Marek Brabec
2. **The impact of outliers on the IV and 2SLS estimators in the linear regression model with endogeneity**
   Aleš Toman
3. **The impact of missing data imputation procedures on the data topology**
   Blagoje Ivanović, Katarina Halaj, Bojana Milošević, Danijel Subotić and Mirjana Veljović
4. **Improving the representativeness of non-probability samples: A case study of two web surveys**
   Ana Slavec
5. **Toward unified criteria for assessing construct validity in quantitative, qualitative and mixed methods research**
   Joca Zurc and Anuška Ferligoj
6. **Statistical approximations to the Ising model on fractal lattices**
   Andrej Srakar

12.00–14.00 **Break**

14.00–15.30 **Invited session**
Room 1 *Chair: Fabio Divino*

1. **COVID-19 in Slovenia, from a success story to disaster: What lessons can be learned?**
   Damjan Manevski

2. **From data to modelling: Why statistics is fundamental to manage the epidemic**
   Antonello Maruotti, Alessio Farcomeni, Fabio Divino, Giovanna Jona-Lasinio, Gianfranco Lovison, Pierfrancesco Alaimo Di Loro and Marco Mingione

3. **Reproducibility in COVID-19 experience: Pitfalls and challenges**
   Clelia Di Serio

14.00–15.30 **Network analysis**
Room 2 *Chair: Germà Coenders*

1. **Exploring the effect of extreme anchor labeling on research findings**
   Vanja Erčulj and Anže Mihelič

2. **Weighting in non-compensatory composite indices: The weighted Mazziotta-Pareto index**
   Matteo Mazziotta and Adriano Pareto

3. **Quality of mixed methods research in intervention studies: Preliminary results**
   Gizela Kopač and Valentina Hlebec

4. **Mixed field of mixed methods: Bibliographic analysis**
   Daria Maltseva, Stanislav Moiseev and Joca Zurc

5. **Using a predictive model to map the Russian information operation networks**
   Sachith Dassanayaka, Dimitri Volchenkov and Ori Swed

6. **Obtaining closed form Bayes factors from summary statistics in common experimental designs**
   Thomas Faulkenberry

15.30–16.00 **Break**

16.00–17.00 **Invited lecture**
Room 1 *Chair: Andrej Blejec*

1. **The seven deadly sins of big data – (and how to avoid them)**
   Richard De Veaux

10.00–16.00 **Workshop**
Room 1

1. **SHARE dataset and analysis of nonstandard data: Examples and applications**
   Andrej Srakar

ABSTRACTS

Invited lecture

Clustering of attribute and/or relational data

Anuška Ferligoj

University of Ljubljana, Ljubljana, Slovenia
anuska.ferligoj@fdv.uni-lj.si

A large class of clustering problems can be formulated as an optimizational problem in which the best clustering is searched for among all feasible clustering according to a selected criterion function. This clustering approach can be applied to a variety of very interesting clustering problems, as it is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and/or by the definition of the set of feasible clusterings. Both, the blockmodeling problem (clustering of the relational data) and the clustering with relational constraint problem (clustering of the attribute and relational data) can be very successfully treated by this approach. It also opens many new developments in these areas.

Invited session

Compositional data analysis

Michael Greenacre

Universitat Pompeu Fabra, Barcelona, Spain
michael.greenacre@upf.edu

Compositional data can be defined as vectors of positive components, where the relative importance of these components is what ultimately matters to the research questions. This relative importance is distilled in ratios, or, more precisely, the logarithms of these ratios, i.e. logratios, for their appropriate statistical treatment. The compositional data tradition started treating data with a fixed sum (compositions expressed as proportions, percentages, ppm, etc.) but has recently spanned all fields in which relative importance is of interest, be the data constrained to a fixed sum or not. Applications include chemical and geological compositions, time budgets, genomics, microbiomics, geology, financial ratios, pollution studies, dietary studies, text mining and content analysis, and so on. Standard statistical learning methods have to be adapted to this new scenario, as well as the interpretation of their results. The session presents three advancements in these domains.
The first contribution in this sessıon shows that for high-dimensional data found typically in genomic and microbiome studies the simplest logratio transformation, the additive logratio, can almost perfectly reproduce the exact logratio geometry and thus provide a practical solution that is easy to interpret. The second contribution involves mining the prediction of a criterion variable using logratios between pairs of components as predictors, where the issue of interpretation is crucial. The third contribution presents a new approach to compositional regression where the sampling units are not treated in a single simplex space but in a simplicial complex after combining the units into groups.

Compositional data analysis of high-dimensional biological datasets: A revalidation of the additive logratio transformation

Michael Greenacre[1], *Marina Martínez-Álvaro*[2] *and Agustín Blasco*[3]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Scottish Rural College, Edinburgh, UK
[3]Universitat Politècnica de València, València, Spain
michael.greenacre@gmail.com, marina.alvaro@sruc.ac.uk, ablasco@dca.upv.es

Microbiome and omics datasets are, by their intrinsic biological nature, of high dimensionality, characterized by counts of large numbers of components (microbial genes, operational taxonomic units, etc.), and regarded as compositional since the total number of counts identified within a sample are irrelevant. The central concept in compositional data analysis is the logratio transformation, the simplest being the additive logratios with respect to a fixed reference component. A full set of additive logratios is not isometric in the sense of reproducing the geometry of all pairwise logratios exactly, but their lack of isometry can be measured by the Procrustes correlation. The reference component can be chosen to maximize the Procrustes correlation between the additive logratio geometry and the exact logratio geometry, and for high-dimensional data there are many potential references. As a secondary criterion, minimizing the variance of the reference component's log-transformed relative abundance values makes the subsequent interpretation of the logratios even easier. On each of three high-dimensional datasets the additive logratio transformation was performed, using references that were identified according to the abovementioned criteria. For each dataset the compositional data structure was successfully reproduced, that is the additive logratios were very close to being isometric. The Procrustes correlations achieved for these datasets were 0.9991, 0.9974 and 0.9902, respectively. It is thus demonstrated that, for high-dimensional compositional data, additive logratios can provide a valid choice as transformed variables, which (a) are subcompositionally coherent, (b) explain 100% of the total logratio variance and (c) come measurably very close to being isometric, that is approximating almost perfectly the exact logratio geometry. The interpretation of additive logratios is simple and, when the variance of the log-transformed reference is very low, it is made even simpler since each additive logratio can be identified with a corresponding compositional component.

Three approaches to supervised learning for compositional data with pairwise logratios

Germà Coenders[1] *and Michael Greenacre*[2]

[1]Universitat de Girona, Girona, Spain
[2]Universitat Pompeu Fabra, Barcelona, Spain
germa.coenders@udg.edu, michael.greenacre@upf.edu

The common approach to compositional data analysis is to transform the data by means of logratios. Logratios between pairs of compositional parts (pairwise logratios) are the easiest to interpret in many research problems, and include the well-known additive logratios as particular cases. When the number of parts is large (sometimes even larger than the number of cases), some form of logratio selection is a must, for instance by means of an unsupervised learning method based on a stepwise selection of the pairwise logratios that explain the largest percentage of the logratio variance in the compositional dataset. In this article we present three alternative stepwise supervised learning methods to select the pairwise logratios that best explain a dependent variable in a generalized linear model. The first method features unrestricted search, where any pairwise logratio can be selected. This method has a complex interpretation if some pairs of parts in the logratios overlap, but it leads to the most accurate predictions. The second method restricts parts to occur only once, which makes the corresponding logratios intuitively interpretable. This method can be related to the discriminative balance approach. The third method uses additive logratios, so that $k-1$ selected logratios involve exactly k parts. This method in fact searches for the subcomposition with the highest explanatory power and its objectives are thus connected to the regularized regression and selbal approaches. Once the subcomposition is identified, the researcher's favourite logratio representation may be used in subsequent analyses, not only pairwise logratios. We present an illustration of the three approaches on a dataset from a study predicting Crohn's disease, already used in the selbal approach. The first method excels in terms of predictive power, and the other two in interpretability.

Combinatorial regression in abstract simplicial complexes

Andrej Srakar[1] *and Miroslav Verbič*[2]

[1]Institute for Economic Research and University of Ljubljana, Ljubljana, Slovenia
[2]School of Economics and Business, University of Ljubljana and Institute for Economic Research, Ljubljana, Slovenia
andrej.srakar@ier.si, miroslav.verbic@ef.uni-lj.si

Regression analysis with compositional data in mathematical statistics has so far been limited to regressions on a single simplex space. We extend this to regression in abstract simplicial complex (as a family set of simplicial objects), developing a novel regression perspective, labelled combinatorial regression, based on combining n-tuplets of sampling units into groups and treating them on a simplicial complex (Lee, 2011; Korte, Lovasz and Schrader, 1991) as the regression sample space. The novel perspective is estimated in two stages: in the first (estimating initial regression output), combining Multivariate Distance Matrix Regression (McArdle and Anderson, 2001) and Plackett-Luce approaches, and in the second extending random walk perspectives on simplicial complexes (Mukherjee and Steenbergen, 2016) with the recent regression simplicial complex (neural) network perspective (Firouzi et al., 2020). It allows extensive number of perspectives in the analysis of, for example, triplets, quadruplets or quintuplets (or any n-tuplet) and using as measure of disparity between the units (to construct regressors) different distance and/or divergence measures. It also allows applications to very small datasets as the number of units in the new model can be expressed in terms of generalized factorial products (Dedekind numbers) of units of original sample. Computational issues, prone to statistical and probabilistic work on simplicial complexes are solved using approaches of computational topology (e.g. van Ditmarsch et al., 2020). In this article, we provide the analysis of new approach for different n-tuple combinations using Jensen-Shannon and generalized Jensen-Shannon divergence measures and provide the asymptotic limits of the approach and exploring its properties also in a Monte Carlo simulation study. In a short application we present analysis of sessile hard-substrate marine organisms image data from Italian coast areas which allows to explore the new approach in relative abundance data setting.

Actuarial & financial statistics

Consistently recovering the signal from noisy functional data

Siegfried Hörmann and Fatima Jammoul

Graz University of Technology, Institute of Statistics, Graz, Austria
shoermann@tugraz.at, f.jammoul@tugraz.at

We consider noisy functional data $Y_t(s_i) = X_t(s_i) + u_{ti}$ that has been recorded at a discrete set of observation points. Naturally, the goal is to recover the underlying signal X_t. Commonly, this is done by non-parametric smoothing approaches, e.g. kernel smoothing or spline fitting. These methods act function by function and do not take the overall presented information into consideration. We argue that it is often more accurate to take the entire data set into account, which can help recover systematic properties of the underlying signal. Other approaches using functional principal components do just that, but require strong assumptions on the smoothness of the underlying signal. We show that under very mild assumptions, the signal may be viewed as the common components of a factor model. Using this discovery, we develop a PCA driven approach to recover the signal and show consistency. Our theoretical results hold under rather mild conditions, in particular we do not require specific smoothness assumptions for the underlying curves and allow for a certain degree of autocorrelation in the noise. We demonstrate the applicability of our approach with simulation experiments and real life data analysis. Our considerations show that even in settings that are advantageous for competing methods, the factor model approach provides competitive results. In particular we observe that for growing sample size, the factor model approach shows an improving fit, which is not the case for classic spline smoothers. The proposed method performs particularly well in cases of rough data and provides insight into the nature of underlying functional structure in real life data cases.

Modelling the polynomial time trend through spline function: A Bayesian procedure

Varun Agiwal

Indian Institute of Public Health, Hyderabad, India
varunagiwal.stats@gmail.com

In this paper, we develop an estimation procedure for an autoregressive model with polynomial time trend approximated by a spline function. Spline function has the advantage of approximating the non-linear time series in an appropriate degree of polynomial time trend model. For Bayesian parameter estimation, the conditional posterior distribution is obtained under two symmetric loss functions. Due to the complex form of the conditional posterior distribution, Markov Chain Monte Carlo (MCMC) approach is used to estimate the Bayes estimators. The performance of Bayes estimators is compared with that of the corresponding maximum likelihood estimators (MLEs) in terms of mean squared error (MSE) and average absolute bias (AB) via a simulation study. To illustrate the proposed study, import series of Brazil, Russia, India, China, and South Africa (BRICS) countries are analyzed.

Variable selection for mixtures of regression models with random effects

Luísa Novais and Susana Faria

University of Minho, Guimarães, Portugal
luisa_novais92@hotmail.com, sfaria@math.uminho.pt

In recent years, the technological advances have led to the existence of large and highly complex databases, which can lead to models that contain a large number of explanatory variables. As such, classical variable selection methods become unfeasible with the increasing size of the databases, being technologically too challenging to be used in practice. Thus, variable selection has become crucial in any modeling study, requiring the search for the simplest model that adequately describes the observed data. In the last few decades, there has been the need to develop new variable selection methods that allow us to overcome the issue of the computational complexity and that are capable of dealing with databases with a large number of explanatory variables. Among the new methods, methods based on penalizing functions have received great attention in the literature. These methods, unlike the classical methods, can be used in large database problems as they estimate the effect of the non-significant variables to be zero, removing them from the model, which, in consequence, drastically decreases the computational load. In this work, we investigate different variable selection methods based on penalty functions that act on the coefficients of the variables, which simultaneously allows variable selection and variable estimation, in particular the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO), the HARD and the Smoothly Clipped Absolute Deviation (SCAD), comparing their performance in identifying the most relevant subset of explanatory variables using the Expectation-Maximization (EM) and the Classification Expectation-Maximization (CEM) algorithms. In order to compare the performance of both algorithms in variable selection for mixtures of regression models with random effects for the different methods based on penalty functions, an extensive simulation study is developed and the developed methodologies are applied to a set of real data. The research of L. Novais was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference number SFRH/BD/139121/2018.

Estimation of multicomponent stress-strength reliability for unit Burr-XII distribution

Fatma Gul Akgul

Artvin Coruh University, Artvin, Turkey
ftm.gul.fuz@artvin.edu.tr

In this study, we consider the classical and Bayesian estimation of reliability in the multicomponent stress-strength model when both the stress and strengths are drawn from the unit-Burr XII distribution. The maximum likelihood (ML) and Bayesian methods are used in the estimation procedure. The Bayesian estimates of reliability are obtained by using Lindley's approximation and Markov Chain Monte Carlo (MCMC) methods due to the lack of explicit forms. The asymptotic confidence intervals are constructed based on the ML estimators. The MCMC method is used to construct the Bayesian credible intervals. A Monte-Carlo simulation study is conducted to investigate and compare the performance of the proposed methods. Finally, analysis of the real data set is presented for illustrative purposes.

Statistical applications I

Complex hypothesis testing on circular economy

Stefano Bonnini , Michela Borghesi and Getnet Melak Assegie

University of Ferrara, Ferrara, Italy
stefano.bonnini@unife.it, michela.borghesi@unife.it, getnet.melakassegie@unife.it

Circular Economy (CE) is nowadays a much-discussed topic because the idea that a linear production system is no longer sustainable from an environmental point of view is becoming more widespread. Some empirical studies have been published on the topic. However, there is a lack of literature about valid statistical approaches for testing complex hypotheses about CE. For example, an interesting hypothesis concerns the effect of companies' age in the propensity of SMEs to undertake CE activities. The main difficulty of such problem, ignored in the literature, is due to the presence of confounding factors such as company size and business sector. To verify the aforementioned hypothesis, it is suitable to stratify with respect to size and sector, or to consider homogeneous companies in terms of size and/or sector. A possible consequence is represented by small sample sizes that encourage the use of non-parametric testing methods. Our proposal is based on the use of a nonparametric method that presupposes the decomposition of the problem in partial tests and consists in the combination of permutation tests. Non-parametric tests are advantageous because they don't require that the probability law underlying the data belongs to a specific parametric family of distributions so they are more flexible and robust than the so-called parametric tests. Another strength of the proposed methodology is that it is suitable to test complex hypotheses such as U-shaped or V-shaped alternatives, for example when the effect of the age of the company on the propensity towards CE is decreasing for young companies and increasing for older companies.

A detailed statistical analysis of COVID-19 worldwide effects on economic, social and health welfare

Maurizio Brizzi[1] *and Diletta Cecilia Canini*[2]

[1]Department of Statistical Science "Paolo Fortunati", University of Bologna, Bologna, Italy
[2]Department of Statistical Science "Paolo Fortunati", University of Bologna, Rimini, Italy
maurizio.brizzi@unibo.it, diletta.canini@studio.unibo.it

The pandemic of COVID-19 has undoubtely affected the welfare level of almost all world countries, and it has sensibly alterated the values of a huge number of demographic, social, economic and health-related variables. In this paper we have considered and analyzed a huge set of such variables, trying to evaluate the different effects this unexpected world crisis has induced on world countries, as well as the relationships between some important social and economic variables and the intensity of pandemic effect. Some specific statistical tools have been used, such as cograduation indices and Multiple Correspondence Analysis (MCA) have been applied to verify the interaction of variables and to identify clusters of countries which had a similar pandemic impact.

Spatial non-stationarity in the determinants of land use in Campania (southern Italy) based on the GWR model

Gennaro Punzo[1], *Rosalia Castellano*[2] *and Emma Bruno*[1]

[1]University of Naples Parthenope - Department of Economic and Legal Studies, Napoli, Italy
[2]University of Naples Parthenope - Department of Management and Quantitative Studies, Napoli, Italy
gennaro.punzo@uniparthenope.it, lia.castellano@uniparthenope.it, emma.bruno@studenti.uniparthenope.it

The progressive urban conversion of natural land in artificial areas is one of the main concerns in recent years as it has serious implications for the environment in terms of damage to ecosystems and for the social and economic well-being of a community. The problem of land transformation is particularly felt in Italy where land use patterns show a high level of territorial heterogeneity. This research aims to investigate spatial non-stationarity in the determinants of land use in Campania (southern Italy). Campania is an interesting case study for three main reasons: i) it is the third region for land use in Italy and the first in southern Italy; ii) it is the most populous region in southern Italy and the most densely populated in Italy; iii) it is characterised by a complex and varied morphological structure due to the presence of the Somma-Vesuvius volcanic massif. We perform Geographically Weighted Regression (GWR) to manage spatial non-stationarity and to provide a model that better describes the data structure. The data are taken from official sources (Ispra, Istat, SIEPI) for 2016 on all 550 Campanian municipalities. The results show the crucial role of the geomorphological, demographic, socio-economic and institutional characteristics in determining land use patterns. Spatial non-stationarity shows that land use in Campania is characterised by territorial asymmetries with the presence of areas whose land use is not aligned with the real needs of the territory. The findings suggest that: i) monitoring land use changes is the prerequisite for preserving environmental quality and ecosystem services; ii) better local institutions are needed to guide territorial planning in support of sustainable land management; iii) broader administrative planning can strengthen land management by sharing responsibility among an adequate number of local authorities.

Mixed models for anomaly detection in aggregate anti-money laundering reports

Marianna Siino and Stefano Iezzi

Financial Intelligence Unit - Bank of Italy, Rome, Italy
marianna.siino@bancaditalia.it, stefano.iezzi@bancaditalia.it

In spite of the strict international standards enforced in most countries for the purpose of fighting money laundering and terrorist financing, criminal organisations and terrorists actively engage in attempting to use financial institutions as vehicles for funnelling their own ill-gotten financial resources. Under the Italian Anti-Money Laundering Law, banks and other financial intermediaries are mandated to file aggregate anti-money laundering reports (SARAs from the Italian acronym) to Italy's central anti-money laundering authority, the so-called Financial Intelligence Unit (FIU). Differently from Suspicious Transaction Reports (STRs), SARAs are non-nominal threshold-based reports, referring to all transactions amounting to 15,000 euros or more, after aggregating them according to several classification criteria related to the customer, his/her sector of activity, the type of transaction, and, in case of cross-border wire transfers, the country of the counterpart and of his/her intermediary. The aggregate reports are filed on a monthly basis to allow Italy's FIU to carry out analysis aimed at identifying any phenomena of money laundering or terrorist financing which do not emerge from STRs. To this end, statistical and machine learning techniques are deployed in order to detect financial anomalous conducts, which is an ambitious goal due to the complexity of the phenomenon and of the available data. This study offers a contribution to the class of techniques for anomaly detection, by proposing the application of linear mixed effect models to the monthly cross-border wire transfers in SARAs. The proposed approach is applied to the cross-border wire transactions between Italy and three foreign countries in 2019. The mixed effect models are estimated with a computationally high-performance procedure that overcomes the problems involving a large number of random effects and observational units. Several model specifications have been compared and an in-sample validation through perturbation of the data has shown good preliminary results. This versatile approach, which takes properly into account the complex multi-level structure of the data, can have a more general use for monitoring any type of financial transactions for the detection of anomalies.

Applying multivariate statistical process control for mixed data to prosthetic rehabilitation after lower-limb amputation

Gaj Vidmar, Neža Majdič and Helena Burger

University Rehabilitation Institute, Ljubljana, Slovenia
gaj.vidmar@ir-rs.si, neza.majdic@ir-rs.si, helena.burger@ir-rs.si

Multivariate statistical process control (MVSPC) based on mixed data (i.e., with some variables numeric and some categorical) is a recent and little-known development. The proposed approaches include modifications of Hotelling's T-squared statistic (which is the basis for MVSCP for numeric data) and approaches based on measuring distances between mixed-data-points (e.g., Gower distance or Euclidean distance). We tried nine methods: local and global Euclidean distance, local and global Gower distance, T-squared using Gower distance with or without bootstrap, T-squared using Gower distance with bootstrap based on principal component analysis, and permutational implementations of global Gower distance and T-squared using Gower distance. The methods were applied on data from 100 patients after lower-limb amputation who had received a permanent transtibial prosthesis at the University Rehabilitation Institute in Ljubljana. The data included six nominal variables (e.g., sex and diagnosis), two ordinal (e.g., activity level) and three numeric variables (e.g., age and stump circumference). A patient was considered out-of-control if he/she returned to our outpatient clinic because of problems with the prosthesis within one year from receiving the prosthesis. Data from 50 patients were used for phase I (i.e., parameter estimation); the data from the other 50 patients were used for phase II (i.e., assessment). Statistically assigned and actual patient status were compared. The performance of the methods was assessed using ROC-curves (where pre-set type I error rate was the varying criterion), classification accuracy and Cohen's kappa coefficient. All the methods yielded above-chance agreement with the actual in-control or out-of-control status (AUC values around 0.7, all statistically significantly >0.5). The highest classification accuracy and kappa values were obtained using Local Euclidean distance and local Gower distance. Overall, the proposed methods proved to be useful and could therefore be introduced into routine healthcare quality control practice.

Biostatistics & bioinformatics

The importance of imperfect detection in biological data: Large-scale climate effects meet an Amazonian butterfly

Maja Kajin[1], *Carla Maria Penz*[2] *and Phil De Vries*[2]

[1]University of Ljubljana, Biotechnical faculty, Department of Biology, Ljubljana, Slovenia
[2]University of New Orleans, New Orleans, United States of America
majakajin@gmail.com, carla.penz@gmail.com, phil.devries@gmail.com

The aim of this presentation is to show the importance of considering imperfect detection in ecological studies, more specifically population ecology. Imperfect detection is a phenomenon common to most types of biological data and it refers to the fact that we can (almost) never detect all of the desired samples (e.g. individuals in a population). Once this imperfect detection is quantified (e.g. through capture-recapture sampling), the remaining estimates of models' real parameters become more accurate. Furthermore, by considering that some individuals might have been alive but were temporarily outside the sampling area, additional population parameters, such as temporary emigration, can be modeled. The study case shows the importance of considering imperfect detection and temporary emigration for detecting large-scale climate effects (El Niño) on population dynamics of an Amazonian butterfly.

Years life difference compared to the general population

Damjan Manevski , Maja Pohar Perme and Nina Ružić Gorenjec

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

damjan.manevski@mf.uni-lj.si, maja.pohar@mf.uni-lj.si, nina.ruzic.gorenjec@mf.uni-lj.si

When performing survival analysis on data with long-term follow-up, one is often interested in comparing the estimated survival to the one in the general population. In such a setting, the number of years lost/saved measure has been commonly used since it has an easy interpretation which also makes it appealing to the lay audience. Several approaches for defining and estimating the number of years lost/saved have emerged in previous literature. However, many of these proposals have not been fully defined, hence some important theoretical and practical issues need to be resolved before such a measure can be standardly used. In this work, we consider the main results from the previous literature and introduce the years life difference measure. We carefully examine the subtle differences with the previous proposals - while all the measures deal with the number of years lost (or saved), they all in fact answer different questions. A non-parametric estimator for the years life difference measure is defined which relies upon external population mortality data for calculating the population curve. The use of mortality data is common for relative survival, but its practical application is not straightforward, thus we also provide an efficient R implementation. In addition, we will consider the variance of the years life difference estimator, with bootstrap being the only reliable option so far. The practical aspect of this work will be illustrated using a motivational example on the long-term survival of elite athletes.

Comparison of clustering methods for diabetic kidney disease patients formalized through category theory

Maria Mannone, Veronica Distefano, Claudio Silvestri and Irene Poli

ECLT, Ca' Foscari University of Venice, Venice, Italy
maria.mannone@unive.it, veronica.distefano@unive.it, silvestri@unive.it, irenpoli@unive.it

Precision medicine aims to find the best individualized treatment for each patient. In particular, type-2 diabetes patients that present kidney complications (diabetic kidney disease, DKD) show relevant heterogeneity in the response to the therapeutic treatment. Aiming to develop a decision system to find the best individualized drug combination, we try to find subgroups of similar patients. Seeking a precise patients grouping, we compare two clustering methods. The first is based on the agglomerative hierarchical clustering with the Gower distance for mixed data, and the second is based on the k-medoids algorithm. The comparison of two patients (according to all their variables) with the Gower distance gives a scalar; the pairwise comparison of all patients gives a dissimilarity matrix for each time point. The k-medoids algorithm is based on a generalized distance, suitable for mixed data, and minimizes the distance between clusters. The comparison between methods is contextualized within the theoretical framework of category theory, which formalizes the idea of transformation between transformations. A category is constituted by objects (points) and morphisms (arrows) between them. Categories allow for nested comparisons. The morphisms between categories are called functors, and the comparison between functors is a natural transformation. A clustering method can be seen as a functor from a dataset equipped with distances to a partition of the dataset. We can extend this idea to the comparison of clustering methods, formalizing it as a natural transformation. We compare these methods using the DC-ren longitudinal dataset, with mixed data of DKD patients. With both methods, we build clusters of similar patients, analyzing their mean values of variables and their response to the given drugs. The theoretical contextualization can help convert theorems and former knowledge from an abstract field to an applied one, giving new insights for further research and studies.

Invited session

Data science research ethics

Marianne Huebner

Department of Statistics and Probability, Michigan State University, USA
huebner@msu.edu

New challenges arise in research as data sets have grown in size and complexity. Data from different sources and public data are used to explore research questions. Are you prepared to address the emerging ethical issues surrounding research with big data? The research process should include a pre-specified study design and analysis plan and follow a systematic approach, including careful documentation for reproducibility, with data protection measures in place. Collaborations across multiple perspectives (scientific, statistical, computational, statistical, ethical) are needed. Training programs on big data ethics for graduate students should be offered and more awareness of data ethics challenges is needed for researchers. This session aims to provide an overview of some of these challenges.

Managing research data for transparency and reusability

Scout Calvert

University Library, Michigan State University, Lansing, USA
calvert4@mail.lib.msu.edu

Said to promote reproducibility and prevent fraud, data sharing is becoming a scientific norm and an expectation from funding agencies. There's also evidence it can help the careers of researchers. But sharing data is not so easy as simply sharing files. This presentation will provide some strategies for managing data so it can be ethically shared, understood, and reused.

Data science research ethics and the challenges of inference, public data and consent

Jacob Metcalf

Data & Society Research Institute, New York, USA
jake.metcalf@datasociety.net

Data science, and the related disciplines of machine learning and artificial intelligence, are founded on the assumed availability of massive amounts of data. The scientific and economic justification for collecting and using all that data is deceptively simple: we can infer expensive- and hard-to-know data from cheap- and easy-to-know data and make predictions and automated decisions on the basis of the patterns we find. When that data is about human behavior, that inferential step is ethically fraught because it often involves data that is ubiquitous (social media, geolocation, biometrics, etc.) being used to predict traits that are from an entirely different context (race, religion, sexual preference, gender, etc.), and typically without knowledge or consent. This is a highly complex ethical challenge, yet our research ethics norms and regulations were written for a different paradigm of scientific research. In this talk, I will illustrate this dynamic with several cases of data science research ethics controversies and consider how we might establish new practices for ethical research.

Good data science practice: Moving towards a code of practice for drug development

Mark Baillie

Novartis, Basel, Switzerland
mark.baillie@novartis.com

There is growing interest in data science and the challenges that could be solved through its application. The growing interest is in part due to the promise of "extracting value from data". The pharmaceutical industry is no different in this regard reflected by the advancement and excitement surrounding data science. Data science brings new perspectives, new methods, new skill sets and the wider use of new data modalities. For example, there is a belief that extracting value from data integrated from multiple sources and modalities using advances in statistics, machine learning, informatics and computation can answer fundamental questions. These questions span a variety of themes including: disease understanding (i.e. "precision" medicine, disease endo/phenotyping, etc.), drug discovery (i.e. new targets and therapies), measurement (i.e. multi-omics, digital biomarkers, software as a medical device, etc.), and drug development (i.e. dose-exposure-response, efficacy, safety, compliance, etc.). By answering these fundamental questions, we can not only increase knowledge and understanding but more importantly inform decision making; accelerating drug and medical device development through data-driven prioritisation, precise measurement, optimised trial design and operational excellence. However, with the promise of data science, there are also several obstacles to overcome, especially if data science is to live up to this promise and deliver a positive impact. These obstacles include consensus on a common understanding of the very definition of data science, the relationship between data science and existing fields such as statistics and computing science, what should be involved in the day to day practices of data science, and what is "good" practice. The talk will explore these issues with the aim of opening a dialogue on good data science practice.

Modeling & simulation

Estimating the conditional distribution in functional regression problems

Thomas Kuenzer[1], *Siegfried Hörmann*[1] *and Gregory Rice*[2]

[1]Graz University of Technology, Graz, Austria
[2]University of Waterloo, Waterloo, Canada
kuenzer@tugraz.at, shoermann@tugraz.at, grice@uwaterloo.ca

We consider the problem of consistently estimating the conditional distribution $P(Y \in A \mid X)$ of a functional data object $Y = (Y(t) : t \in [0,1])$ given covariates X in a general space, assuming that Y and X are related by a functional linear regression model. Two natural estimation methods for this problem are proposed, based on either bootstrapping the estimated model residuals, or fitting functional parametric models to the model residuals and estimating $P(Y \in A \mid X)$ via simulation. We show that under general consistency conditions on the regression operator estimator, which hold for certain functional principal component based estimators, consistent estimation of the conditional distribution can be achieved, both when Y is an element of a separable Hilbert space, and when Y is an element of the Banach space of continuous functions on the unit interval. The latter results imply that sets A that specify path properties of Y that are of interest in applications can be considered, such as the maximum of the curve. Our methods have numerous applications in the context of constructing prediction sets, quantile regression and VaR estimation. Compared to direct modelling these curve properties using scalar-on-function regression, modelling the whole response distribution and extracting the curve properties in a second step allows us to harness the full information contained in the functional data to fit the regression model and achieve better results. We study the proposed methods in several simulation experiments and real data analysis of electricity price curves and show that they outperform both the non-parametric kernel estimator and functional binary regression.

Univariate goodness-of-fit tests for randomly censored data: tests' adaptation versus data transformation

Marija Cuparić and Bojana Milošević

University of Belgrade - Faculty of Mathematics, Belgrade, Serbia
marijar@matf.bg.ac.rs, bojana@matf.bg.ac.rs

Recently, several approaches for adaptation of goodness of fit tests for censored data have been proposed. This paved the way for the bunch of goodness of tests for such data. However, those tests usually depends on censoring distribution which is unknown in practice, and the application of resampling procedures is indispensable, but computationally expensive, step toward obtaining p-values. Here, we present an imputation procedure that can serve as an alternative approach to adaptation proposal. Additionally, we illustrate proposal on several characterization based exponentiality tests proposed so far.

Robust mixture regression modeling for heterogeneous data sets

Fatma Zehra Doğru[1] *and Olcay Arslan*[2]

[1]Giresun University, Giresun, Turkey
[2]Ankara University, Ankara, Turkey
fatma.dogru@giresun.edu.tr, oarslan@ankara.edu.tr

Modeling skewness and heavy-tailedness in heterogeneous data sets is a challenging problem especially in regression analysis. To do so, this study aims to propose mixture regression modeling based on the shape mixtures of skew Laplace normal (SMSLN) distribution for modeling skewness and heavy-tailedness simultaneously. This newly proposed model will be an alternative to the mixture regression model based on the shape mixtures of skew-t-normal (SMSTN) distribution. The SMSLN distribution given by Doğru and Arslan (2019, 2021) is a flexible extension of the skew Laplace normal distribution and has also an extra shape parameter that enables controlling skewness and kurtosis. On the other hand, skewness and heavy-tailedness can be modeled by skew t, skew t normal, or SMSTN (Tamandi et al. (2019)) distributions. Unlike the SMSTN distribution, the SMSLN distribution has fewer numbers of parameters to be estimated, and hence it is computationally less intensive than the SMSTN distribution. We give the expectation-maximization (EM) algorithm to obtain the maximum likelihood (ML) estimators for the parameters of interest. The performances of the proposed estimators are demonstrated with a simulation study and a real data example as the "Pinus Nigra tree" data set. Results are also compared with the results obtained from the mixture regression model based on the SMSTN distribution.

New class of goodness-of-fit tests based on independence-type characterizations

Katarina Halaj[1], *Bojana Milošević*[2], *Marko Obradović*[2] *and Maria Dolores Jiménez-Gamero*[3]

[1]Faculty of Transport and Traffic Engineering, University of Belgrade, Belgrade, Serbia
[2]Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
[3]Dpto. Estadística e Investigación Operativa, Universidad de Sevilla, Seville, Spain
k.halaj@sf.bg.ac.rs, bojana@matf.bg.ac.rs, marcone@matf.bg.ac.rs, dolores@us.es

We present a new class of characterization-based test statistics which can be used for testing goodness-of-fit with several classes of null distributions. The resulting tests are consistent against fixed alternatives. Some limiting and small sample properties of the test statistics are explored. In comparison with common universal goodness-of-fit tests, the new tests exhibit a very competitive behavior. The handiness of the proposed tests is demonstrated through several real data examples.

Modeling complex histograms

Herwig Friedl

Graz University of Technology, Graz, Austria
hfriedl@tugraz.at

Data is available from black and white C-SAM images of wafer structures. The statistical analysis is based on the corresponding multimodal histograms of the greyscales. The objective is to draw conclusions on both, the quality of the wafers as also on the contrast of the images. A heterogeneous mixture of gamma densities together with a uniform component has been applied to enable such a two-fold failure analysis.

Determining factor impacting electronic fitness tracker usage for health and wellness management via predictive analytics

Sinjini Mitra

California State University, Fullerton, Fullerton, USA
smitra@fullerton.edu

Increasingly, wearable technologies such as smartwatches and electronic fitness trackers are becoming popular on a daily basis, from individuals to organizational wellness programs. As with any digital technology, the use of such trackers varies considerably among consumers based on several factors. In this paper, we present some initial analysis from a sample of 145 individuals to determine how fitness devices affect personal health and wellness in a college-age population dominated by non-traditional or underrepresented students such as ethnic minorities, first-generation students, among others. We found that a range of factors, from demographic background (like gender, ethnicity) to health conditions and goals to technology and social media usage and experiences predict the level of physical and fitness activities individuals perform on a daily basis. Moreover, we also explore how perceptions about health/fitness and motivation to lead a healthier lifestyle potentially changed with the use of fitness trackers. Finally, additional analyses were performed to understand potential privacy and security concerns that people may have with the data collected by the electronic fitness devices, technology issues, and other challenges associated with these devices, and how those experiences and attitudes pose obstacles to more widespread adoption of this technology among the general population.

Invited lecture

On censoring (with a nod towards causality)

Jan Beyersmann

Ulm University, Ulm, Germany
jan.beyersmann@uni-ulm.de

Survival or time-to-event analysis is a key discipline in biostatistics, currently put to prominent use in trials on treatment of and vaccination against COVID-19. A defining characteristic is that participants have varying follow-up times and outcome status is not known for all individuals. This phenomenon is known as censoring. If time-to-event and time-to-censoring are entirely unrelated, it is rather easy to see that hazards remain identifiable from censored data, and hazard estimators may subsequently be transformed to recover probability statements. However, COVID-19 treatment and vaccination trials are just two of the many examples where event and censoring times are related. Luckily, the modern counting process approach to survival analysis finds that hazards remain identifiable under rather general "independent censoring" mechanisms, including those encountered in COVID-19 trials. Given the theoretical and practical relevance of censoring, it is rather disturbing to find—as I will demonstrate—that there is a Babylonian confusion on "independent censoring" in the textbook literature. Unfortunately, censoring processes as in COVID-19 trials are two examples where the textbook literature often goes haywire. It is a small step from this mess to misinterpretations of both hazards and censoring. On the other hand, there currently is a very active debate about the use of hazards spearheaded by causal reasoning. In a nutshell, the worry is that hazards are conditional quantities which renders causal conclusions impossible ("collider bias"). I will argue that causal reasoning somewhat overfocusses on interventional "do(no censoring)" effects (which is not what identifiability of hazards is about) and that the collider bias issue disappears from a functional point of view, but that hazards remain rather subtle quantities. Time permitting, I will illustrate matters with a causal g-computation-/Aalen-Johansen-type analysis of clinical hold in a randomized clinical trial.

Invited session

Blockmodelling

Aleš Žiberna

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
ales.ziberna@fdv.uni-lj.si

Blockmodeling is a technique for finding clusters of units that are equivalent based on some notion of equivalence (stochastic, structural, regular, generalized, etc.) and therefore occupy similar position in the network. It also deals with determining ties among these clusters. The presentations in this invited session will focus on different aspect of blockmodeling. The first presentation will compare methods for blockmodeling dynamic networks via Monte Carlo simulation study, the second one discusses generalized direct blockmodeling of larger networks, while the third one uses an adaptation of stochastic blockmodeling to disentangle homophily, community structure and triadic closure in networks.

Blockmodeling dynamic networks: A Monte Carlo simulation study

Marjan Cugmas and Aleš Žiberna

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
marjan.cugmas@fdv.uni-lj.si, ales.ziberna@fdv.uni-lj.si

Social network analysis methodology is essential for studying the relationships among units when networks operationalise such relationships. For example, suppose the aim is to identify groups of equivalent units (according to their links) and the links among the groups so obtained, for which a researcher can apply blockmodeling. Moreover, suppose that several networks are observed regarding the same units at different points in time. In this case, specific blockmodeling approaches are available for use. An overview of some of these blockmodeling approaches is to be provided in this presentation, while differences among them are to be highlighted. Alongside this general overview, a Monte Carlo simulation study is described that empirically evaluates the differences among these blockmodeling approaches. Various factors are considered in this study, such as blockmodel type, blocks' densities, the stability of groups in time, local network mechanisms, and network size. The study results indicate that while separate analyses of networks at different time points prove sufficient in some cases, the use of blockmodeling for dynamic networks improves the results in particular other cases. General guidelines on the use of one approach or another will be given.

Disentangling homophily, community structure and triadic closure in networks

Tiago Peixoto

Central European University, Vienna, Austria
peixotot@ceu.edu

Network homophily, the tendency of similar nodes to be connected, and transitivity, the tendency of two nodes being connected if they share a common neighbor, are conflated properties in network analysis, since one mechanism can drive the other. Here we present a generative model and corresponding inference procedure that is capable of distinguishing between both mechanisms. Our approach (https://arxiv.org/abs/2101.02510) is based on a variation of the stochastic block model (SBM) with the addition of triadic closure edges, and its inference can identify the most plausible mechanism responsible for the existence of every edge in the network, in addition to the underlying community structure itself. We show how the method can evade the detection of spurious communities caused solely by the formation of triangles in the network, and how it can improve the performance of link prediction when compared to the pure version of the SBM without triadic closure.

Generalized direct blockmodeling of large valued networks

Carl Nordlund

Institute for Analytical Sociology, Linköping University, Norrköping, Sweden
carl.nordlund@liu.se

Essentially a data reduction technique for networks, blockmodeling allow for the identification of nodes that are equivalent in some meaningful sense, and how these sets of nodes relate to each other. Contrary to community detection methods, blockmodeling is agnostic about the kind of underlying anatomy that may exist. What is provided, however, is the specific notion of equivalence that should apply: whereas structural equivalence is the most rudimentary form of equivalence, implying that actors have identical ties to alters, regular and generalized equivalence increase the complexity and variety of the kinds of relational patterns we are looking for in a network. Blockmodeling heuristics can be separated into indirect and direct approaches. The indirect uses proxy measures of equivalence, followed by hierarchical clustering to identify sets of actors that are equivalent. Suitable also for valued networks, the indirect approach is limited to the more rudimentary form of structural equivalence. The direct approach is not constrained to structural equivalence, but its computationally intensive search algorithms confides it to small networks (<50). Additionally, due to its direct matching between empirical networks and ideal binary ties, the direct approach also struggles with valued networks. This paper presents novel approaches for direct blockmodeling of both valued and large networks. For the former, a weighted-correlation-based measure of fit is introduced which allows for direct blockmodeling of valued networks using the standard set of ideal binary blocks, without any a priori transformation or dichotomization of the valued relations. For the latter, a hybrid sequential indirect-direct approach is proposed, where an indirect approach is used to, first, reduce a network to a structural block image that subsequently is used as input to a direct weighted-correlation-based approach.

Statistical applications II

Spatial statistical modeling of air pollution

Marek Brabec

Institute of Computer Science, The Czech Academy of Sciences, Prague, Czech Republic
mbrabec@cs.cas.cz

We will present several approaches to the problem of large-scale spatial statistical modeling of selected air pollutants with the special twist of presence of known spatial heterogeneity brought in by the patchwork of background and urban areas with substantially different autocorrelation properties of the spatial field. This is solved in the current large ARAMIS (Air quality research assessment and monitoring integrated system, SS02030031) project sponsored by the Technology Agency Czech Republic. In the modeling, we start from the idea of urban increment field used in various numerical models in Atmospheric sciences. Then we formulate several additive statistical models with background and urban increment components (plus several other regression terms correcting for known nuisance covariates). In fact, the correction terms include output from numerical air pollution modeling (in particular CAMx and Symos) to correct for non-stationarity caused by physical sources. Our approaches to model identification/estimation will include both frequentist and Bayesian strategies. In particular, we use penalized component GAM (Generalized Additive Model) based on low rank implementation of Gaussian processes approximately corresponding to traditional geostatistical covariance models (Wood 2017) and then also Bayesian spatially-varying coefficient model (Finley, Banerjee 2019). We will illustrate our modeling framework in detail on large scale measurement data from professional measurement network run by the Czech Hydrometeorological Institute.

The impact of outliers on the IV and 2SLS estimators in the linear regression model with endogeneity

Aleš Toman

University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia
ales.toman@ef.uni-lj.si

In a linear regression model, endogeneity (i.e., a correlation between some explanatory variables and the error term) makes the classical OLS estimator biased and inconsistent. When instrumental variables (i.e., variables correlated with the endogenous explanatory variables but not with the error term) are available to partial out endogeneity, the IV and 2SLS estimators are consistent and widely used in practice. The effect of outliers on the OLS estimator is carefully studied in robust statistics, but surprisingly, the effect of outliers on the IV and 2SLS estimator has received little attention in previous research. Existing work has mainly focused on the robust estimation of the variable cross-covariance matrices that are later used in IV and 2SLS estimators. In this presentation, we use the forward search algorithm to investigate the effect of outliers (and other contamination schemes) on various aspects of the IV-based estimation process. The algorithm begins the analysis with a subset of observations that does not contain outliers and then increases the subset by adding one observation at a time until all observations are included and the entire sample is analyzed. Contaminated observations are included in the subset in the final iterations. During the process, various statistics and residuals are monitored to detect the effects of outliers. We use simulation studies to examine the effect of known outliers occurring in the (i) dependent, (ii) exogenous or (iii) endogenous exploratory, or (iv) instrumental variable. Summarizing the results, we propose and implement a method to identify outliers in a real data set where contamination is not known in advance.

The impact of missing data imputation procedures on the data topology

Blagoje Ivanović[1], *Katarina Halaj*[2], *Bojana Milošević*[1], *Danijel Subotić*[3] *and Mirjana Veljović*[1]

[1]Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
[2]Faculty of Transport and Traffic Engineering, University of Belgrade, Belgrade, Serbia
[3]ASW Inženjering D.O.O, Belgrade, Serbia
blagoje_ivanovic@matf.bg.ac.rs, k.halaj@sf.bg.ac.rs, bojana@matf.bg.ac.rs, danijel.subotic@asw.eu, mirjana_veljovic@matf.bg.ac.rs

Non-responses in surveys, non-recorded data, limitations of measuring devices, time limitations, etc. usually result in data incompleteness. Since most statistical models and machine learning procedures are not designed for incomplete data, many different imputation procedures have been proposed so far. In this work, we review several most commonly used parametric and nonparametric missing data imputation procedures and compare their performance from different angles, including the impact on the underlying topological structure. The latter will be achieved by examining the relative change in persistency homology diagrams of true and imputed data sets. All imputation procedures are tested on many artificially generated data clouds with specific shapes, as well as on several real datasets.

Improving the representativeness of non-probability samples: A case study of two web surveys

Ana Slavec

InnoRenew CoE, Izola, Slovenia
ana.slavec@innorenew.eu

Web surveys, even for purposes of scientific data collection, are commonly based on non-probability samples as this saves costs and other resources. Unlike probability sampling procedures, non-probability sampling does not enable the generalisation of results from sample to the population. Since certain users are more likely to volunteer to participate, non-probability samples often have a certain selection bias. The representativeness of non-probability sampling designs can be improved with measures such as trying to spread the sample recruitment as broadly as possible by combining several recruitment channels. This contribution presents the case study of two web surveys in Slovenia that were based on large convenience samples, first on the topic of COVID-19 protective measures and the second on topic of COVID-19 vaccination. In both cases, we run a parallel survey where the same questionnaire was administered to members of an online market research panel that is representative of the Slovenian population. Based on the comparison of results of the two convenience samples to the respective panel samples we estimate how biased they are and discuss possible approaches to improve their representativeness.

Toward unified criteria for assessing construct validity in quantitative, qualitative and mixed methods research

Joca Zurc[1] *and Anuška Ferligoj*[2]

[1]University of Maribor, Faculty of Arts, Maribor, Slovenia
[2]University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
joca.zurc@um.si, anuska.ferligoj@fdv.uni-lj.si

Validation frameworks and validity criteria increase the meaning and usefulness of data and findings of empirical research. Thus, the interest for appraising research validity is presented as long as researching itself. In the recent 60 years, from defining validity as the three types of validation procedures—content, criterion, and construct (Cronbach & Meehl, 1955)—an extensive amount of valuable works were contributed to the validity issue in quantitative, qualitative and mixed methods research. However, despite many attempts to systemize the field, we are still facing the diversity of terms, frameworks and criteria of assessing the meaning of data and inferences across all methodological traditions, rapidly appearing in studies of mixing quantitative and qualitative approaches. Therefore, our study aimed to contribute to the crucial discussion on developing the unified validity assessment criteria as core standards in mixed methods research across different disciplines and research designs. Based on a systematic literature review and experts' interviews with nine international mixed methods scholars, developers of the field, our study revealed ten essential criteria presenting the principles of the construct validity and leading to a new validity assessing framework in mixed methods research. The findings seem to be in line with the construct validity framework of Messick (1995) and the validation framework in mixed methods research of Dellinger & Leech (2007). We did, however, elucidate the structure of the framework and heterogeneity between the criteria. Furthermore, our study presented the three unique criteria indispensables in apprising construct validity in a mixed methods study. Hence, the integration of quantitative and qualitative approaches, engagement of differences and believability upgrade the existing frameworks by emphasizing specific features of the mixed methods methodology that should be considered in their validity assessment. Future studies on empirically testing and evaluation of practical use of the new framework should be encouraged.

Statistical approximations to the Ising model on fractal lattices

Andrej Srakar

Institute for Economic Research (IER) and University of Ljubljana, Ljubljana, Slovenia
andrej.srakar@ier.si

The Ising spin glass is a one-parameter exponential family model for binary data with quadratic sufficient statistic. Bhattacharya and Mukherjee (2017) showed that given a single realization from this model, the maximum pseudolikelihood estimate of the natural parameter is $\sqrt{(a_N)}$-consistent at a point whenever the log-partition function has order a_N in a neighborhood of that point. The exact solutions of the Ising model in one and two dimensions are well known, but much less is known about solutions on fractal lattices. In an important contribution, Codello, Drach and Hietanen (2015) constructed periodic approximations to the free energies of Ising models on fractal lattices of dimension smaller than two using a generalization of combinatorial method of Feynman and Vdovichenko. We generalize their approach to fractal lattices of dimensions 2 and greater than 2, in particular of Koch curve variety. To this end we combine combinatorial optimization and transfer matrix approaches, referring to earlier works of Andrade and Salinas (1984). We compute approximate estimates for the critical temperatures and compare them to more usual Monte Carlo estimates. Referring to Codello et al., we compute the correlation length as a function of the temperature and extract the relative critical exponent. The method allows generalizations to any fractal lattice, as well as concrete solutions to approach solutions for other non-translationally invariant lattices (e.g. those with random interactions). We illustrate applications of our results on synthetic and real-world data.

Invited session

Statistical analysis of the COVID-19 outbreak: Methods and application

Fabio Divino

University of Molise, Campobasso, Italy
fabio.divino@unimol.it

Since the end of 2019, beginning in China, the entire planet has experienced the spread of the COVID-19 outbreak due to the infection by the Sars-Cov-2 virus. The world wide emergency has interested many fields of biomedical sciences, such as virology or immunology, but also the fields of biostatistics and epidemiology. The possibility and capability to develop statistical models able to describe and predict the evolution of the epidemic curves have played a central role in the management of the emergency. Indeed, the contributions of many statisticians in supporting the efforts of the public health agencies and institutions have been fundamental for decoding the different patterns of the pandemic across time, in order to properly support the decision makers. In this Section, the three invited speakers, all with recognized records in this field, will present their recent results and developments, at both methodological and applicative levels.

COVID-19 in Slovenia, from a success story to disaster: What lessons can be learned?

Damjan Manevski

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

damjan.manevski@mf.uni-lj.si

During the first wave of the COVID-19 pandemic (spring 2020), Slovenia was among the least affected countries in Europe. During the second wave (autumn 2020), the situation became drastically worse with high numbers of deaths per number of inhabitants ranking Slovenia among the most affected countries. This was true even though strict non-pharmaceutical interventions (NPIs) were enforced to control the progression of the epidemic. Using a semi-parametric Bayesian model (developed for the purpose of this study) we explore if and how the changes in mobility, their timing and the activation of contact tracing can explain the differences in the epidemic progression of the two waves. To fit the model we use data on daily numbers of deaths, patients in hospitals, intensive care units, etc. and allow transmission intensity to be affected by contact tracing and mobility (data obtained from Google Mobility Reports). Our results imply that though differences between the two waves cannot be fully explained by mobility levels and contact tracing, implementing interventions at a similar stage as in the first wave would keep the death toll and the health system burden low in the second wave as well. On the other hand, sticking to the same timeline of interventions as observed in the second wave and focusing on enforcing a higher decrease in mobility would not be as beneficial. According to our model, the "dance" strategy, i.e. first allowing the numbers to rise and then implementing strict interventions to make them drop again, has been played at too late stages of the epidemic. In contrast, a fixed strategy of reducing the mobility by 15–20% compared to the pre-COVID level would suffice to keep the epidemic under control. A very important factor in this result is the presence of contact tracing, without it, the reduction in mobility needs to be substantially larger. The flexibility of our proposed model allows similar analyses to be conducted for other regions even with slightly different data sources for the progression of the epidemic; the extension to more than two waves is straightforward. The model could help policymakers worldwide make better decisions regarding the timing and severity of the adopted NPIs.

From data to modelling: Why statistics is fundamental to manage the epidemic

Antonello Maruotti[1], *Alessio Farcomeni*[2], *Fabio Divino*[3], *Giovanna Jona-Lasinio*[4], *Gianfranco Lovison*[5], *Pierfrancesco Alaimo Di Loro*[4] *and Marco Mingione*[4]

[1]Libera Università Maria Ss Assunta, Roma, Italy
[2]University of Rome "Tor Vergata", Rome, Italy
[3]University of Molise, Campobasso, Italy
[4]University of Rome "La Sapienza", Rome, Italy
[5]University of Palermo, Palermo, Italy
a.maruotti@lumsa.it, alessio.farcomeni@uniroma2.it, fabio.divino@unimol.it, giovanna.jonalasinio@uniroma1.it, gianfranco.lovison@unipa.it, pierfrancesco.alaimodiloro@uniroma1.it, marco.mingione@uniroma1.it

In epidemic challenges the statistician has a key role to play: informing policy decisions, tracking changes, evaluating risks. The proposed methods, based ensamble approaches and/or spatio-temporal models, aim at monitoring and forecasting the main indicators describing the evolution of COVID-19, quantifying the impact on human's health and on the health system. A parallel aim is that of informing best policy practices. The outcomes will be real-time risk-based indicators, to identification of areas at-risk with guarantees on correctness of alarms, prediction of pressure on the health system. Formally, we introduce an extended generalised logistic growth model for discrete outcomes, in which spatial and temporal dependence are dealt with the specification of a network structure within an Auto-Regressive approach. A major challenge concerns the specification of the network structure, crucial to consistently estimate the canonical parameters of the generalised logistic curve, e.g. peak time and height. We compared a network based on geographic proximity and one built on historical data of transport exchanges between regions. Parameters are estimated under the Bayesian framework, using Stan probabilistic programming language. The proposed approach is motivated by the analysis of both the first and the second wave of COVID-19 in Italy, i.e. from February 2020 to July 2020 and from July 2020 to December 2020, respectively. We analyse data at the regional level and, interestingly enough, prove that substantial spatial and temporal dependence occurred in both waves, although strong restrictive measures were implemented during the first wave. Accurate predictions are obtained, improving those of the model where independence across regions is assumed.

Reproducibility in COVID-19 experience: Pitfalls and challenges

Clelia Di Serio

Centro Universitario di Statistica per le Scienze Biomediche, Universitá Vita Salute San Raffaele, Milano, Italy

diserio.clelia@unisr.it

When dealing with biomedical data retrieved under emergency conditions, main statistical features of study design are dismissed, no matter how big the data collection can be. From COVID-19 experience we learned one major scientific lesson concerning the importance of "quality" rather than "quantity" in collecting observational data to enhance new knowledge in medicine. Despite the huge amount of COVID-19 publications in a very short period, it has been clearly seen that good statistical and computational tools cannot overcome poor quality of data. Understanding the COVID-19 data generating process results fundamental for answering to crucial scientific questions that nowadays remains still unsolved such as those concerning prevalence, immunity, transmissibility and susceptibility. This type of data suffer from many limitations for gathering robust clinical conclusions due to unmeasured confounders, measurement errors, and bias selection effects. Each of these characteristics represents a source of uncertainty, often ignored or assumed to be random, that may limit the degree of reproducibility and lead to paradoxical conclusions in assessing the role of risks factors. In fact, new paradigms and new designs schemes must be investigated to make inferential conclusion meaningful and informative when dealing with case series studies and data such those collected during COVID-19 emergency.

Network analysis

Exploring the effect of extreme anchor labeling on research findings

Vanja Erčulj[1] *and Anže Mihelič*[2]

[1]Faculty of Criminal Justice and Security, Ljubljana, Slovenia
[2]Faculty of criminal justice and security, Ljubljana, Slovenia
vanja.erculj@fvv.uni-mb.si, anze.mihelic@fvv.uni-mb.si

The Likert(-type) scale assumes that the strength of an individual's attitude is linear and thus can be measured. The strength of agreement with statements is mostly measured on a five- or seven-point scale with response anchors labeled from "Strongly disagree" to "Strongly agree". Such anchor labeling provides verbal symmetry and balance to the scale and ensures that intervals between anchors are as close to equally distanced as possible in measuring attitudes. For respondents, however, using verbally symmetric anchor labeling may result in interpretation difficulties. This is particularly pronounced in translations of anchor labels to languages where the phrase "strongly (dis)agree" is nearly never used. Therefore, extreme values are commonly translated verbally asymmetrical but semantically symmetrical, for example: from "Sploh se ne strinjam" to "Se popolnoma strinjam" in Slovene and from "Stimme überhaupt nicht zu" to "Stimme voll und ganz zu" in German. To explore the difference between different extreme anchor labeling, we have conducted an online experiment with two different extreme anchor labeling: verbally symmetric ("Strongly disagree" to "Strongly agree") and verbally asymmetric ("Not agree at all" to "Completely agree") in Slovene. Five constructs of Protection-motivation theory with 26 items were measured. The comparison of the two scales included skewness and kurtosis of individual items, confirmatory factor analysis results, mean values and variances of composite scores (calculated as arithmetic means of items measuring each construct), and results of multiple linear regression. Slightly higher variability of composite scores was found in the verbally asymmetrical group, and the means of the composite score for one construct statistically significantly differed, suggesting a larger perceived distance between extremities of the verbally asymmetrical scale. Slight differences in the results of the multiple linear regression model were observed. Our findings suggest that verbally asymmetrical translation of extreme anchors seems to be slightly superior.

Weighting in non-compensatory composite indices: The weighted Mazziotta-Pareto index

Matteo Mazziotta and Adriano Pareto

Italian Institute of Statistics, Rome, Italy
mazziott@istat.it, pareto@istat.it

Composite indices (also known as composite indicators) are very popular tools for assessing and ranking countries and other geographical areas in terms of development, environmental performance, sustainability, and other complex phenomena that are not directly measurable with a single indicator. The Mazziotta-Pareto Index (MPI) and its variant Adjusted MPI (AMPI) is a composite index for summarizing a set of indicators that are assumed to be not fully substitutable. It is based on a non-linear function which, starting from the simple arithmetic mean of the normalized indicators, introduces a penalty for the units with unbalanced values of the indicators. This methodology is often applied to the calculation of both non-compensatory composite indices of "positive" phenomena, such as well-being and sustainable development, and "negative" phenomena, such as poverty. In the MPI and AMPI, all components are assumed to have equal importance, which may not be the case. In this work, a weighted version of the two indices (WMPI and WAMPI, where W stands for "Weighted") is proposed, when a set of weights is available. In practice, since the MPI and AMPI are based on the calculation of the mean and standard deviation of the normalized values, for each unit, we calculate the weighted mean and weighted standard deviation of the normalized values. The weighted coefficient of variation can then be obtained simply by dividing the weighted standard deviation by the weighted mean. Finally, the two standard formulas can be applied. Some numerical examples are also shown, in order to assess the effect of different weighting schemes on the results.

Quality of mixed methods research in intervention studies: Preliminary results

Gizela Kopač and Valentina Hlebec

University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
gizela.kopac@gmail.com, valentina.hlebec@fdv.uni-lj.si

Mixed methods approach has become very popular in intervention research. Researchers can find basic procedures, practical guidance and mixed methods appraisal tools to follow while realizing a mixed methods in intervention research. Our objective is to examine the use of mixed methods approach in intervention research. We obtained a list of intervention studies in database Springerlink and sampled every fifth study to obtain a sample of 200 interventions which included mixed methods approach. We constructed the conceptual model and divided it into three sections: (i) topic; (ii) checklist items; and (iii) item description. The model contains five topics: research, intervention, quantitative methods, qualitative methods and mixed methods. Then we used this conceptual model to assess mixed methods research in intervention studies. This presentation presents preliminary results of this assessment.

Mixed field of mixed methods: Bibliographic analysis

Daria Maltseva[1], *Stanislav Moiseev*[1] *and Joca Zurc*[2]

[1]Higher School of Economics, Moscow, Russia
[2]University of Maribor, Faculty of Arts, Maribor, Slovenia
dmaltseva@hse.ru, smoiseev@hse.ru, joca.zurc@um.si

Mixed methods research as an intensively emerging methodological field being developed over the past 30 years, which has a broad extension in studies across diverse scientific areas and disciplines. However, there is a little attention given to the research field itself. How the mixed methods research was developed from the beginning until today? Who are the most important pioneers and scientists working in the field? What kind of research interest and themes were addressed in the mixed methods studies? What are the main journals which promote the field development? It is important to answer these questions and to define the characteristics the state-of-the-art of the mixed methods research and methodological development itself. This paper aims to answer these questions by providing a quantitative analysis of the field of mixed methods research that reveals connections between authors, publications and journals from the middle of 20th century to 2018. We collected all available sources from Web of Science (Core Collection) using the keywords "mixed method", "mixed research", and their variation. The data consists of 16,347 papers found by this research query, and 488,696 works, which are being cited by these papers in the reference lists. Using the program WoS2Pajek, we transformed these data into a collection of networks: one-mode citation network and different two-mode networks, including works and authors, works and keywords, and works and journals. This permitted us to get the information on the patterns of publications over time, to distinguish the most important publications, journals and authors in the field, look at the authors' collaboration practices, and get the idea of the topic structure of the field. By performing a "main path" analysis, we traced the most important stages in the evolution of the field, and identified the most relevant body of knowledge that it developed over time, which could be viewed as the main corpus of knowledge for any newcomer in the field. The obtained findings can be used as guidelines for implementing mixed methods research in the future, contribute to a common methodological language, and will be helpful for different users such as researchers, funders and reviewers. The complexity of the mixed methods research and its novelty versus the traditional quantitative and qualitative approaches indicates that the structuring of the field development deserve a special attention

among the many other mixed methods open issues.

Using a predictive model to map the Russian information operation networks

Sachith Dassanayaka, Dimitri Volchenkov and Ori Swed

Texas Tech University, Lubbock, United States
sachith-eranga.dassanayaka@ttu.edu, dimitri.volchenkov@ttu.edu, ori.swed@ttu.edu

Information operations by foreign adversaries pose a meaningful threat to democratic processes. Given the increased frequency of this type of threat, understanding those operations is paramount in the effort of combating their influence. Building on existing scholarship on the inner functions within those influence networks on social media, we suggest a new approach to map those type of operations. Using Twitter content identified as part of Russian influence network we created a predictive model to map the network operations. We classify accounts type based on their authenticity function for a sub-sample of accounts and trained AI to identify similar patterns of behavior across the network. Our method model attains 88% prediction accuracy for the test set. We validate our predicted results set by comparing the similarities with the 3 million Russian troll tweets dataset. The result indicates 81% similarity between the two datasets. The predictive and validation results suggest that our neural network model can use to identify the tweets actors.

Obtaining closed form Bayes factors from summary statistics in common experimental designs

Thomas Faulkenberry

Tarleton State University, Stephenville, Texas, USA
faulkenberry@tarleton.edu

Consider the common scenario where one wishes to test for differences among group means. In a Bayesian framework, the goal is to assess the relative evidence between two competing models: $\mathscr{H}_0$, where all group means are equal, and $\mathscr{H}_1$, where at least one group mean is different from the others. In this talk, I will discuss recent work on developing methods for computing Bayes factors directly from summary statistics in common experimental designs. The Bayes factor, defined as the ratio of marginal likelihoods for the two competing models, represents the factor by which the prior odds for one model over the other is updated after observing data. Particularly, I will discuss a choice of prior distribution that yields Bayes factors with simple, closed form structure. These results allow for a number of nice applications which I will discuss, including a web application that applied researchers can use to measure the evidential value of their own data.

Invited lecture

The seven deadly sins of big data – (and how to avoid them)

Richard De Veaux

Williams College, Williamstown, USA
rdeveaux@williams.edu

Organizations, from government to industry are accumulating vast amounts of data, nearly continuously. Big data and artificial intelligence promise the moon and the stars, "solving previously unsolvable problems". There is certainly a lot of hype. There's no doubt that there are insights to be gained from all these data, but is it as easy as the hype claims? What are the challenges? Much can go wrong in the data analysis cycle, even for trained professionals. In this talk I'll discuss a wide variety of case studies from a range of industries to illustrate the potential dangers and mistakes that can frustrate problem solving and discovery — and that can unnecessarily waste resources. My goal is that by seeing some of the mistakes I (and others) have made, you will learn how to take advantage of data insights without committing the "Seven Deadly Sins".

Workshop

SHARE dataset and analysis of nonstandard data: Examples and applications

Andrej Srakar

Institute for Economic Research and School of Economics and Business, University of Ljubljana, Ljubljana, Slovenia

andrej.srakar@ier.si

Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database, covering 27 European countries and Israel, with micro data on health, socio-economic status and social and family networks of about 140,000 individuals aged 50 or older. Due to its large scale and data quality, it has become subject to numerous applications in past fifteen years, in terms of statistics one could easily describe it as a "toy for statisticians". The workshop will focus on usages of SHARE data with methods from the analysis of nonstandard data, developed for large datasets, largely symbolic and functional data analysis, which are being ever more frequently used in statistics and econometrics. We will present statistical theory behind those methods and use many of them as implemented in R to present interesting and novel applications with SHARE data, based on some of the most recent functional, interval and histogram regressions, clustering and dimensionality reduction methods. The participants should be familiar with R programming and have its latest version installed on their computers. The use of RStudio is also encouraged.

POSTERS

Statistical causal analysis of food quality

Zelimir Kurtanjek

University of Zagreb, Zagreb, Croatia
zelimir.kurtanjek@gmail.com

Applied is statistical evaluation of causal relations between food molecular analytics and food quality tests. Two sets of experimental data of wine and bread quality. The wine data set includes 7500 samples organoleptic wine preferences and 12 biochemical and physicochemical features. The bread data set includes 42 wheat samples and 45 chemical, physical and biochemical properties: indirect quality parameters (6), farinographic parameters (7), extensographic parameters (5), baking test parameters (2) and reversed phase-high performance liquid chromatography (RP-HPLC) of gluten proteins (25). Studied are causality effects of the wheat features and two technical baking quality parameters. The causal effects are evaluated from causal directed acyclic graphs (DAG) and application of Pearl d-separation algorithm to eliminate covariate confounding. The causal graphs are generated by deductive causalities by use of the field knowledge and inductive causalities evaluated statistical from observed experimental data. The causalities are estimated as point estimates by linear regression on population levels of the corresponding sample data sets. The linear estimates are compared to nonlinear causality effects by partial dependence plots of the corresponding boosted random forest decision models. Applied is open source DoWhy software available on GitHub. The causalities are discussed from view point of food quality monitoring, technology production monitoring and control, and potential genetic improvements of the cultivars.

On discriminant analysis using bivariate exponential distributions

Georg Mbaeyi and Chijioke Nweke

Alex Ekueme Federal University, Abakaliki, Nigeria
george.chinanu@funai.edu.ng, cj_nweke@yahoo.com

This study focused on obtaining allocation rules when the assumption of normality is violated. More specifically, when available data is of the bivariate exponential distributions. Both simulated and two sets of real-life data were used to demonstrate the applicability and performance of the derived allocation rules.

Convergence results for solution of stochastic hard-soft constraints convex feasibility problem

Chijioke Joel Nweke[1], *Akaninyene Udo Udom*[2] *and George Chinanu Mbaeyi*[1]

[1]Department of Mathematics and Statistics, Alex Ekwueme Federal University Ndufu-Alike, Ebonyi State, Nigeria

[2]Department of Statistics, University of Nigeria Nsukka, Enugu State, Nigeria

cj_nweke@yahoo.com, chijiokenweke20@gmail.com, chinanu20@gmail.com

This work considers the stochastic convex feasibility problem involving hard constraints (that must be satisfied) and soft constraints (whose proximity function should be minimized) in Hilbert space. Convergence in quadratic mean and almost surely was proved for the result of the solution. An alternating projection involving 1-lipschitzian and firmly non-expansive mapping was adopted.

Nonlinear random forest classification using copula mutual information

Ayyub Sheikhi[1] *and Radko Mesiar*[2]

[1]Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran

[2]Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, STU Bratislava, Slovakia, Bratislava, Slovakia

sheikhy.a@uk.ac.ir, radko.mesiar@stuba.sk

In this work we use a copula mutual information approach to select the most important features for a random forest classification. Based on associated copula mutual information between these features we carry out this feature selection. We then embed the selected features to a random forest algorithm to classify a label-valued outcome. We investigate statistical properties of the proposed classification algorithm. Our Algorithm enables us to select most relevant features when the features are not necessarily connected by a linear function; also, we can stop the classification when we reach the desired level of accuracy. We apply this method on a simulation study as well as a real data set of COVID-19 and for a diabetes dataset.

Network-based point pattern analysis of traffic accidents in City of Cape Town, South Africa

Christo du Toit , Sulaiman Salau and Sebnem Er

University of Cape Town, Cape Town, South Africa
DTTCHR015@myuct.ac.za, sulaiman.salau@uct.ac.za, Sebnem.Er@uct.ac.za

A road traffic accident (RTA) can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment. RTAs have a large social and economic impact on livelihoods in South Africa. In 2018, there were 12 921 RTA related fatalities recorded in South Africa of which 1 064 occurred in the Western Cape Province. In order to effectively reduce the number and the injury-severity of RTAs in South Africa, and ultimately increase road safety, a better understanding of the spatial distribution of road traffic accidents is needed. In this research paper, spatial point pattern analysis of the intensity of geocoded road traffic accidents that occurred between January 2015 and December 2017 in City of Cape Town is conducted. A network based kernel density estimation is implemented using accidents that are constrained on a linear network of roads and the locations of hot spots and the significance of these were determined using network based nearest neighbour distances.

Statistical machine learning for medicinal plant leaves classification

P.G. Jayani Laskshika

University of Sri Jayewardenepura, Nugegoda, Sri Lanka
jayanilakshika76@gmail.com

Medicinal plants are usually identified by practitioners based on years of experience through sensory or olfactory senses. Automatic ways to identify medicinal plants are useful. The main objective of the research is to develop an automatic algorithm to classify medicinal plants using medicinal plant leaves. We refer to our medicinal plant classification algorithm as MEDIPI which is divided into offline phase and online phase. The classification algorithm is trained in the offline phase. In the online phase, the pre-trained classification model is used to real-time leaf image classification for general users. Our classification algorithm operates on the features extracted from the image leaves. First, leaf images are processed by means of sequence of image processing steps. The main image processing steps involve Convert original image to RGB image, Gray scaling, Gaussian smoothing, Binary thresholding, Remove stalk, Closing holes, and Resize image which used to remove undesired distortion. The second stage is to extract features from plant leaf images. We introduced 52 computationally efficient features to classify plant species which are mainly classified into four groups as shape, color, texture, and scagnostics. Length, area, monotonocity are some of them. Next, we trained our algorithm using random forest, gradient boosting, and extreme gradient boosting. The model trained with random forest algorithm provides the highest accuracy. Our algorithm works as a hierarchical classification system which contains 3 levels. The first level classifies images according to the shape. The second level classifies according to the edge types. The bottom level classifies the plant species. We used high dimensional visualization approaches to visualize what is happening inside the trained algorithm and provides transparency to our black-box model. The MEDIPI algorithm yields accurate results to the state-of-the existing techniques in the field for medicinal plants classification. MedLEA is an open-source repository R software established by us.

Accuracy of space-time Moran's I, a dynamic-time dependence spatial autocorrelation detection for spatial panel data with time trend

Rahma Fitriani, Si Darmanto and Zerlita Fakhda Pusdiktasari

University of Brawijaya, Malang, Indonesia
rahmafitriani@ub.ac.id, darman_stat@ub.ac.id,
zerlitafahdha@gmail.com

Spatial panel data are cross section of observations, each is associated with a position in space, repeated over several time periods. At one point in time, nearby observations tend to be similar. The degree of similarity is defined as spatial autocorrelation. Moran's I is a common index to measure the degree of the spatial autocorrelation at one point in time. However, when an apparent trend is observed in the time series of each spatial unit, there might be a time lagged on the spatial effect, such that it fails to detect the contemporaneous spatial autocorrelation in each time unit spatial data. Motivated by this issue, a component which accommodates the dynamic-time dependence nature of the spatial autocorrelation must be accommodated in the Moran's I index for the spatial panel data, which is the main objective of this study. The weight matrix is modified to capture the dynamic. The accuracy of the proposed index is analyzed based on a simulation study. The proposed index works, especially when the degree of the contemporaneous spatial autocorrelation is high. It also succeeds in detecting the dynamic spatial autocorrelation of the number of East Java's Covid-19 cases.

Time series clustering based on time-varying Hurst exponent

Alex Babiš and Beata Stehlíková

Comenius university, Faculty of mathematics, physics and informatics, Bratislava, Slovakia
alexbabis96@gmail.com, stehlikova@fmph.uniba.sk

In our work we deal with clustering of exchange rates based on time-varying estimate of Hurst exponent. Hurst exponent characterize long-range depen-dece of time series, either time series is trending or mean-reverting or be-have completely random. Firstly, we fitted ARIMA-GARCH models to ev-ery time series to reduce biasness of rescaled range analysis method used for estimation of Hurst exponent. We only considered models with good resid-uals, meaning no autocorrelation or ARCH effect was present in residuals. The final model was chosen by Bayesian information criterium. Hurst ex-ponent was estimated on residuals from models by means of rolling window approach. Given time-varying Hurst exponent clustering was employed to capture structure of exchange rate market given response of each individual exchange rate to specific information.

A-optimal designs for cubic polynomial models with mixture experiments in three components

Mahesh Kumar Panda

Central University of Odisha, Odisha, Koraput, India
mahesh2123ster@gmail.com

This article obtains A-optimal minimum support designs for the three different forms of cubic polynomial mixture models, i.e., full cubic, cubic without 3-way effect, and special cubic mixture models in three ingredients. The necessary and sufficient conditions for the proposed designs have been verified by the celebrated equivalence theorem.

Comparison of likelihood ratio statistics for a familial DNA search in subdivided populations: Simulation studies from Thailand

Monchai Kooakachai

Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand

Monchai.K@chula.ac.th

In a familial DNA search, the goal is to infer genetic relationships among forensic DNA samples. Likelihood ratio statistic has been commonly used for the test of hypotheses corresponding to a familial DNA search. It is known to be optimal in a single population framework. However, in subdivided populations, e.g., in the form of racial groups, the likelihood ratio calculation needs to be adjusted since allele frequencies generally do differ among human populations. In this work, we investigated performance of two likelihood ratio statistics for a kinship testing based on Type I error and power. The first one is the classical likelihood ratio with a single set of allele frequency. For this approach, we assumed a homogeneous population, i.e., a population substructure does not exist in the framework. The second statistic is the weighted average of the likelihood ratios under the single population scenarios based on prior probabilities. This is defined by utilizing the allele frequencies from each subpopulation. With simulation studies on Thai population, the latter statistic was found to be better as we found about five and eleven percent increases in statistical power for testing parent-child and full sibling relationships, respectively. This indicates that population substructure should be included in the familial DNA search.

An improved - more robust spatial outliers detection method

Zerlita Fahdha Pusdiktasari, Rahma Fitriani and Eni Sumarminingsih

University of Brawijaya, Malang, Indonesia
zerlitafahdha@gmail.com, rahmafitriani@ub.ac.id, eni_stat@ub.ac.id

Spatial outlier is an object that significantly deviates from its surrounding neighbors. Average Difference Algorithm (AvgDiff) is one of spatial outliers' detection methods, which accommodates spatial information in the calculation of the degree of outlierness (DO). However, AvgDiff has the possibility of swamping effects, due to the non-robust nature of average that is used in the algorithm. Other drawback of AvgDiff, is that it does not use statistical tests to determine the status of an object, whether it is an outlier or not. It chooses top m outliers, which are m objects with largest DO. In this case, researchers needs a priori information to define how many objects they want to detect as outliers. In practice, the researchers would never know how many spatial outliers present in the data. This study aims to propose a more robust spatial outliers' detection method, particularly to reduce the swamping effect. It is done by changing the average to median (which is more robust) in the calculation of scores which represents the neighbors' conditions. Statistical test is used to determine the status of an object. Simulation is conducted to analyze the swamping effect and the accuracy of the method. The result confirms, in the absence of a priori information about the number of outliers contained in the data, the proposed method has a lower level of swamping effect than AvgDiff.

On discriminant analysis with some bivariate exponential distributions

George Mbaeyi and Chijioke Nweke

Alex Ekwueme Federal University, Ndufu-Alike, Nigeria
george.chinanu@funai.edu.ng, cj_nweke@yahoo.com

This study focused on obtaining allocation rules when the assumption of normality is violated in discriminant analysis. More specifically, when available data is of the bivariate exponential distributions. Both simulated and real-life data were used to demonstrate the applicability and performance of the allocation rules.

Spatio-temporal model for categorical data: An application to analyzing rainfall levels

Anagh Chattopadhyay[1] *and Soudeep Deb*[2]

[1]Indian Statistical Institute, Kolkata, India
[2]Indian Institute of Management, Bangalore, India
anagh72@gmail.com, soudeep@iimb.ac.in

The problem of rainfall prediction, for both short and long term future horizon, is an essential and important research question in meteorological studies. It is often of primary interest to analyze and forecast the rainfall level as a categorical variable (binary to denote rainfall or not; multiple categories such as no, low, high etc.), and not as a continuous type variable. In this paper, we propose a new spatio-temporal model to deal with this problem, for rainfall is a phenomena that depends on both spatial proximity and temporal autocorrelation. Our model is defined through a hierarchical structure for the latent variable which corresponds to the probit link function. The mean structure of the proposed model is designed to capture the trend, the seasonal pattern as well as the lagged effects of various environmental variables (temperature,wind speed, pressure, humidity). The covariance structure of the model is defined as an additive combination of a zero-mean spatio-temporally correlated process and a white noise process. The parameters associated with the space-time process enable us to analyse the effect of proximity of two points with respect to space or time, and its influence on the overall process. For implementation, we employ a complete Bayesian framework to get estimates for the parameters in the model. Using appropriate priors on the parameters, we use the concepts of Gibbs sampling to sample from the posterior distribution. Convergence is ensured by borrowing strength from the Gelman-Rubin statistic. Our method is implemented on an Australian dataset, which consists of daily data from 49 locations and 4 years. We find that the lagged environmental variables have a significant effect in determining rainfall levels. The proposed model is found to provide good forecasting results as well. In fact, through an extensive comparative study, we discover that our approach has superior predictive accuracy than other existing methods in the literature.

Estimation of parameters of extended Weibull distribution

Arbër Qoshja, Artur Stringa and Frederik Dara

University of Tirana, Tirana, Albania
arber.qoshja@fshn.edu.al, artur.stringa@fshn.edu.al, frederik.dara@fshn.edu.al

In this article, a generalization of the new Weibull distribution is derived from the modified Lehmann Type II-G class of distributions. Also, we describe different methods of estimation for the unknown parameters of the model. These methods include maximum likelihood, least squares, weighted least squares, Cramer-von Mises, maximum product of spacings, Anderson-Darling and right-tail Anderson-Darling methods. Numerical simulation experiments are conducted to assess the performance of the so obtained estimators developed from these methods. The mean square error is used as the criterion for comparison.

A proposed Bayesian method for the parameter estimation of COGARCH (1,1) model via Lindley's approximation

Yakup Ari

Alanya Alaaddin Keykubat University, Alanya, Türkiye
yakup.ari@alanya.edu.tr

This study aims to propose a Bayesian estimation method for the parameters of the COGARCH (1,1) model using Lindley's approximation, which is an explicit solution to the ratio of integrals. The COGARCH (1,1) model has three parameters: prior distributions are assumed gamma and uniform distributions respectively to satisfy the stationarity conditions. The simulation study is applied to compare the Bayes' estimates under square error loss function with the Pseudo-maximum likelihood estimates. The simulation study is done using Lévy jumps derived from Compound Poisson Process and for different sample sizes which are 2000, 5000, and 10,000. In the simulation study, parameter estimates were compared according to the expected risk values and no significant difference was found between the methods. In addition to the simulation study, for illustrative purposes, the daily USD/TRY exchange rate volatility between the period 2018 and 2021 was predicted by the COGARCH model, and the model's parameters were estimated by Lindley's Approximation and maximum likelihood methods. In conclusion, all estimators have performed almost the same under square error loss functions. It is observed that as expected the expected risk of each estimator decreases as the sample size increases. The error difference in their relative performance tends to get smaller and smaller with the increase in sample size. The prior distribution of the parameter eta should be assumed vague prior or another distribution since uniform distributed prior makes the Bayes estimate equal to the maximum likelihood estimate.

An improved ratio-product-ratio class of estimators for finite population mean

Singh Deepak and Yadav Rohini

Amity Institute of Applied Sciences, Amity University, Noida, India
deepaksingh2112@gmail.com, rohiniyadav.ism@gmail.com

This study proposes an improved ratio-product-ratio class of estimators, which is efficient to linear regression estimator for estimating the population mean utilizing accessible auxiliary information. All the properties like bias and mean squared error are studied under large sample approximation. The new family has been developed by the power transformation which makes the family members further efficient to some existing family of estimators. It has been demonstrated theoretically that the proposed family of estimators at the optimum value of constants is efficient to the usual sample mean, ratio, product, linear regression, and some recently proposed estimators. Empirically it has been shown that the proposed family is efficient to its subfamilies, linear regression estimator, and to some established estimators.

INDEX OF AUTHORS

Index of Authors