International Conference

APPLIED STATISTICS

2013

ABSTRACTS and PROGRAM

September 22 - 25, 2013
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2013

International Conference

APPLIED STATISTICS 2013

ABSTRACTS and PROGRAM

2013

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2013

Organized by

Statistical Society of Slovenia

Supported by

Statistical Office of the Republic of Slovenia

ALARIX

RESULT d.o.o.

Generali SpA

The word cloud on the cover was generated using www.wordle.net. The source text included the abstracts of the talks; the fifty most common words were displayed, and greater prominence was given to words that appeared more frequently.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

311(082)

INTERNATIONAL Conference Applied Statistics (2013; Ribno)
Program and abstracts / International Conference Applied Statistics 2013, Ribno (Bled), Slovenia; organized by Statistical Society of Slovenia ; [edited by Lara Lusa and Janez Stare]. - Ljubljana : Statistical Society of Slovenia, 2013

Dostopno tudi na: http://conferences.nib.si/AS2013/AS2013-Abstracts.pdf

ISBN 978-961-93547-0-4
ISBN 978-961-93547-1-1 (pdf)
1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
268736768

Scientific Program Committee

Janez Stare (Chair), Slovenia
Vladimir Batagelj, Slovenia
Maurizio Brizzi, Italy
Anuška Ferligoj, Slovenia
Dario Gregori, Italy
Dagmar Krebs, Germany
Lara Lusa, Slovenia
Mihael Perman, Slovenia
Tamas Rudas, Hungary
Albert Satorra, Spain
Hans Waege, Belgium
Tomaž Banovec, Slovenia
Jaak Billiet, Belgium
Brendan Bunting, Northern Ireland
Herwig Friedl, Austria
Katarina Košmelj, Slovenia
Irena Križman, Slovenia
Stanislaw Mejza, Poland
Jože Rovan, Slovenia
Willem E. Saris, The Netherlands
Vasja Vehovar, Slovenia

Organizing Committee

Andrej Blejec (Chair)
Lara Lusa
Irena Vipavc Brvar
Bogdan Grmek
Anamarija Rebolj

Published by:	Statistical Society of Slovenia Vožarski pot 12 1000 Ljubljana, Slovenia
Edited by:	Lara Lusa and Janez Stare
Printed by:	Statistical Office of the Republic of Slovenia, Ljubljana
Produced using:	generbook R package
Circulation:	200

PROGRAM

		Hall 1	Hall 2
Sunday	10.30 – 11.00	Registration	
	11.00 – 11.10	Opening of the Conference	
	11.10 – 12.00	Invited Lecture	
	12.00 – 12.20	Break	
	12.20 – 13.40	Developments in Statistics	
	13.40 – 15.00	Lunch	
	15.00 – 16.20	Mathematical Statistics	Modeling and Simulation
	16.20 – 16.40	Break	
	16.40 – 17.40	Education I	Sampling Techniques and Data Collection
	19.00	Reception	
Monday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Social Science Methodology	
	11.40 – 12.00	Break	
	12.00 – 13.00	Measurement	Education II
	13.00 – 14.30	Lunch	
	14.30	Excursion	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.20	Biostatistics and Bioinformatics I	
	11.20 – 11.40	Break	
	11.40 – 12.40	Biostatistics and Bioinformatics II	
	12.40 – 15.00	Lunch	
	15.00 – 16.00	Biostatistics and Bioinformatics III	
	16.00 – 16.20	Break	
	16.20 – 17.40	Statistical Applications - Biostatistics I	
Wednesday	9.30 – 10.50	Econometrics	Statistical Applications - Biostatistics II
	10.50 – 11.10	Break	
	11.10 – 12.30	Design of Experiments	
	12.30 – 12.50	Closing of the Conference	
	13.00 – 14.00	Lunch	
	14.00 – 18.00	Workshop	

10.30–11.00 **Registration**

11.00–11.10 **Opening of the Conference**

11.10–12.00 **Invited Lecture** (Hall 1) *Chair: Andrej Blejec*

1. **Embedding astronomical computer models into principled statistical analyses**
   David A. van Dyk

12.00–12.20 **Break**

12.20–13.40 **Developments in Statistics** (Hall 1) *Chair: David A. van Dyk*

1. **Symbolic covariance matrix for interval-valued data**
   Katarina Košmelj, Jennifer Le-Rademacher and Lynne Billard

2. **Adapting a classification rule to changes of distributions using unlabelled recent data**
   Vera Hofer

3. **Check plots in field breeding experiments**
   Stanislaw F. Mejza and Iwona Mejza

4. **Multivariate statistical process control for mixed-type data based on Gower distance**
   Gaj Vidmar and Neža Majdič

13.40–15.00 **Lunch**

15.00–16.20 **Mathematical Statistics** (Hall 1) *Chair: Damjan Škulj*

1. **Optimal unbiased estimates for** $P\{X < Y\}$
   Marko Obradović, Bojana Milošević and Milan Jovanović

2. **Transmuted generalized Pareto distribution**
   Faton Merovci

3. **Alpha-skew generalized t distribution**
   Sukru Acitas, Birdal Senoglu and Olcay Arslan

4. **Random search of stable member in a matrix polytope**
   Vakif Dzhafarov, Taner Buyukkoroglu and Serife Yilmaz

15.00–16.20 **Modeling and Simulation** (Hall 2) *Chair: Gaj Vidmar*

1. **Bivariate models for time series of counts**
   Miroslav Ristić, Aleksandar Nastić and Predrag Popović

2. **Assessing the model performance of nonparametric fuzzy local polynomial regression with different bandwidth selection methods**
   Memmedaga Memmedli and Munevvere Yildiz

3. **Conceptual model of perceived quality of insurance services: structural equation model application**
   Urban Šebjan and Polona Tominc

4. **A comparison of the recent ridge parameters in modeling GDP of Turkey**
   Yasin Asar, Andan Karaibrahimoglu and Aşir Genç

16.20–16.40 **Break**

16.40–17.40 **Education I** (Hall 1) *Chair: Lara Lusa*

1. **Team teaching with student: completion of experimental study**
   Jerneja Šifrer, Zala Žvab and Matevž Bren

2. **Issues and challenges of statistical literacy and statistics education in the higher education**
   Peter Kovacs

3. **AFRICA, Managing Statistics**
   Mohamadu Sani

16.40–18.00 **Sampling Techniques and Data Collection** (Hall 2) *Chair: Katarina Košmelj*

1. **Data collection on prevalence and patterns of drug use among general Slovene population**
   Katja Rostohar and Darja Lavtar

2. **The effect of sample size and weighting procedures on final estimators of health indicators**
   Metka Zaletel, Katja Rostohar, Aleš Korosec and Darja Lavtar

3. **Regional input-output tables for the Czech Republic**
   Jaroslav Sixta, Lenka Hudrlikova and Jakub Fischer

4. **Estimation in adaptive cluster sampling using auxiliary information**
   Muhammad Noor-ul-amin

19.00 **Reception**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Janez Stare*

1. **Clustered data inference when the cluster size is potentially infomative**
   Somnath Datta

10.00–10.20 **Break**

10.20–11.40 **Social Science Methodology** (Hall 1) *Chair: Somnath Datta*

1. **Statistical approach to environmental impacts analysis: the case of greenhouse gas emissions**
   Žiga Kotnik, Maja Klun and Damjan Škulj

2. **Use of different reliability and validity methods**
   Jerneja Šifrer and Matevž Bren

3. **Underlying patterns in the profile: a multidimensional scaling approach to profile analysis.**
   Patrik P. Bratkovič

4. **Measuring and visualizing change in cross-sectional time series**
   Aleš Žiberna

11.40–12.00 **Break**

12.00–13.00 **Measurement** (Hall 1) *Chair: Aleš Žiberna*

1. **Application of discriminant analysis and multinomial logistic regression in investigation of regional disparities in Serbia**
   Valentina T. Sokolovska and Emilija Nikolić-Djorić

2. **The forward search algorithm and confirmatory factor analysis**
   Aleš Toman

3. **Vernon versus Carroll: confirmatory evidence about the structure of intellectual abilities from higher education standardized tests in Costa Rica**
   Eiliana Montero-Rojas

12.00–13.20 **Education II** (Hall 2) *Chair: Andrej Blejec*

1. **House of cards: a mixed-methods study on why behavioral science students fail or succeed an introductory statistics course**
   Floryt Van Wesel and Jan B. Hoeksma

2. **Motivation for learning statistics using SPSS. A case study at the Faculty of Organizational Sciences, University of Maribor**
   Alenka Brezavšček, Petra Šparl and Anja Žnidaršič

3. **An integrative approach to teaching statistics in social sciences: the example of EU political space**
   Jaro Lajovic

4. **Are results of university entrance-exams in Israel good predictors of success in B.A. programs?**
   Tal Shahor, Nissim Ben David and Tavor Tchai

13.00–14.30 **Lunch**

14.30 **Excursion**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Vasja Vehovar*

1. **Choosing mode of data collection for surveys to maximize data quality**
   Jon Krosnick

10.00–10.20 **Break**

10.20–11.20 **Biostatistics and Bioinformatics I** (Hall 1) *Chair: Jon Krosnick*

1. **Analysing disease recurrence with missing at risk information**
   Tomaž Štupnik and Maja Pohar Perme

2. **Effect of informative censoring on net survival estimates**
   Anamarija Rebolj Kodre and Maja Pohar Perme

3. **Methods of tied events handling in Cox proportional hazard model**
   Jadwiga Borucka

11.20–11.40 **Break**

11.40–12.40 **Biostatistics and Bioinformatics II** (Hall 1) *Chair: Janez Stare*

1. **Gaussian moralized reciprocal graph modeling for genomic data integration**
   Carel F.W. Peeters, Wessel N. van Wieringen and Mark A. van de Wiel

2. **Detecting gene-longevity association using bivariate survival and genetic data.**
   Alexander Begun, Andrea Icks and Guido Giani

3. **Boosting high-dimensional classifiers**
   Rok Blagus and Lara Lusa

12.40–15.00 **Lunch**

15.00–16.00 **Biostatistics and Bioinformatics III** (Hall 1) *Chair: Delphine Maucort-Boulch*

1. **Fractal Dimension of the Stochastic Hybrid Lee-Carter Mortality Model**
   Andrzej Szymanski, Agnieszka Rossa and Leslaw Socha

2. **Early stroke recovery efficacy: a mathematical model**
   Lavoslav Čaklović and Miljenka Jelena Jurašić

3. **Local estimation of standardized incidence ratio applied to Slovenian cancer data**
   Tina Žagar, Vesna Zadnik and Maja Primic Žakelj

16.00–16.20 **Break**

16.20–17.40 **Statistical Applications - Biostatistics I** (Hall 1) *Chair: Anamarija Rebolj*

1. **Predicting basal metabolic rate using multiple longitudinal anthropometric measurements**
   Siti Haslinda Mohd Din and Emmanuel Lesaffre

2. **Indirect treatment comparisons in meta-analysis of clinical trials of antiepileptic drugs used as add-on therapy**
   Igor Locatelli, Urban Bernat and Mitja Kos

3. **Dynamic graph generation from MS Excel using R: an implementation based on SVG and shiny**
   Črt Ahlin and Lara Lusa

4. **cartHD: R package for the estimation and validation of classification trees with binary outcomes**
   Lara Lusa and Rok Blagus

9.30–10.50 **Econometrics** (Hall 1) *Chair: Mihael Perman*

1. **Robust standard errors for panel data: a general framework**
   Giovanni Millo

2. **Insurance industry and stochastic scenario**
   Dario Galdieri

3. **An estimate of the degree of interconnectedness between European regions**
   Davide Fiaschi and Angela Parenti

4. **Are funding of pensions and economic growth directly linked? New empirical results for some OECD countries**
   Pietro Cavallini, Gaetano Carmeci and Giovanni Millo

9.30–10.50 **Statistical Applications - Biostatistics II** (Hall 2) *Chair: Gregor Gorjanc*

1. **Woody vegetation structure on land abandoned from agricultural production in South Europe**
   Andreja Radović, Thomas Wrbka, Kiril Vassilev and Vassiliki Kati

2. **Different methods for handling non-Gaussian longitudinal outcome subject to potentially random dropout**
   Ali Satty

3. **The new mixture distribution models for wind speeds**
   Murat Erisoglu, Ulku Erisoglu, Aydin Karakoca and Serkan Akogul

4. **Statistical fit of empirical distributions to GPS derived animal movement data**
   Robert Mutwiri, Thomas Achia, Henry Mwambi, Rob Slotow and Vanak Vanak

10.50–11.10 **Break**

11.10–12.30 **Design of Experiments** (Hall 1) *Chair: Rok Blagus*

1. **Regular A-optimal spring balance weighting designs under special assumptions on errors**
   Bronisław Ceranka and Małgorzata Graczyk

2. **SSD populations – statistical characteristics and practical approach in grain legumes**
   Maria Surma, Tadeusz Adamski, Zygmunt Kaczmarek, Anetta Kuczynska, Karolina Krystkowiak and Stanislaw F. Mejza

3. **Evaluation of parental forms on the basis of series of unreplicated experiments with their hybrids and standard varieties**
   Zygmunt Kaczmarek, Elzbieta Adamska, Iwona Mejza, Henryk Wos and Renata Trzeciak

4. **Multiple one-way ANCOVA when the distributions of both the covariates and the error terms are non-normal**
   Pelin Kasap and Birdal Senoglu

12.30–12.50 **Closing of the Conference**

13.00–14.00 **Lunch**

14.00–18.00 **Workshop** (Hall 1)

1. **Bayesian computation with INLA**
   Thiago G. Martins

ABSTRACTS

Invited Lecture

Embedding astronomical computer models into principled statistical analyses

David A. van Dyk

Statistics Section Imperial College, London, United Kingdom
dvandyk@imperial.ac.uk

Sophisticated and computational expensive computer models are routinely used to describe complex processes in the physical, engineering, and social sciences. In Astronomy, for example, computer models are used to describe such processes as the evolution of stars and galaxies, the formation of the Universe, and the workings and calibration of sophisticated space-based telescopes. Like a statistical likelihood, these models typically predict observed quantities as a function of a number of unknown parameters. Including them as components of a multi-level statistical model, however, leads to significant modeling, inferential, and computational challenges. In this talk, I describe how we tackle these challenges in the context of two examples: (i) a principled statistical analysis of stellar evolution and (ii) the calibration of space-based X-ray telescopes. In our stellar evolution model we must link together a number of separate computer models for various phases in the life of a star. These separate models are combined via parametric models that relate the output of one model with the inputs of the next and are themselves of direct scientific interest. In our calibration example, we employ computational simulators of the subassembly components of the detector to emulate the prior distribution of its response function. In both cases we embed the computer models into likelihood-based statistical models that allow for principled inference but present substantial computational challenges. We use a combination of sophisticated MCMC techniques and simple emulators of the computer models to tackle these challenges. This strategy allows us to apply the full force of powerful statistical tools to build, fit, check, and improve the statistical models and their computer model components. In contrast to traditional methods, we can estimate the uncertainty in our fit using principled statistical techniques, and we can ensure that the components of our overall statistical models are internally coherent.

Developments in Statistics

Symbolic covariance matrix for interval-valued data

Katarina Košmelj[1], *Jennifer Le-Rademacher*[2] *and Lynne Billard*[3]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]Medical College of Wisconsin, Milwaukee, USA
[3]University of Georgia, Athens, USA
katarina.kosmelj@bf.uni-lj.si, jlerade@mcw.edu, lynneb@uga.edu

Bertrand and Goupil (2000) derived the sample variance for the interval-valued data. Billard (2008) extended this result to the bivariate case to obtain the symbolic covariance matrix for interval-data. She showed that it can be decomposed into the sum of the "covariance between" (based on the interval midpoints) and "covariance within" (based on the interval ranges). Consequently, we can assess the gain in the principal component analysis (PCA) when the input-data are intervals compared to the situation when the input-data are interval midpoints. An illustration of PCA on interval data will be presented using the meteorological data from Slovenia in 1971-2011.

Adapting a classification rule to changes of distributions using unlabelled recent data

Vera Hofer

Department of Statistics and Operations Research, University of Graz, Graz, Austria
vera.hofer@uni-graz.at

The distributions a classification problem is based on can be subject to changes over the course of time. Such changes can relate to the class prior distribution (global change) or to the conditional or unconditional feature distributions (local change). After any change the original training data comprising features and class labels are no longer representative, and thus the classifier's performance will deteriorate. In many applications the re-estimation of the classification rule is impossible due to a lack of recent labelled data. However, it is often easy to measure the predictors of new instances. Credit scoring servers as a typical example of such an application. While the features of the applicants can be measured easily, their class labels are not fully known before some time lapse.
To adapt a classification rule after changes a model is introduced that estimates global and local changes in a two-step procedure using only unlabelled recent data. The model is based on mixture distributions, where the local changes are modelled as local displacement of probability mass from the positions of the old components as given in the conditional feature distribution to the positions of the new components as given in the unconditional feature distribution. Assuming that the transfer of probability mass is carried out at a minimum of energy, local changes are estimated by solving a transportation problem for a fixed value of the class prior change. In a further step the global change is found as the minimum of the objective function values obtained from the transportation problem.
The usefulness of the proposed models is demonstrated using artificial data and a real-world dataset from the area of credit scoring.

Check plots in field breeding experiments

Stanislaw F. Mejza and Iwona Mejza

Poznan University of Life Sciences, Poznań, Poland
smejza@up.poznan.pl, imejza@up.poznan.pl

In field plant breeding trials, many times it is not possible to use a traditional experimental design that satisfies the requirement of replicating all the treatments. It results from the facts of the huge number of genotypes involved, the limited amount of seed and the low availability of resources. Then, the unreplicated designs can be used for testing genotypes. By unreplicated design we mean one in which examined genotypes are replicated only once. The use of unreplicated design may be very attractive and the only possible way to carry out an evaluation (inference) of the lines. Additionally, to control the real or potential heterogeneity of experimental units and to make statistical analysis available, control (check) plots are arranged in the trial.
Some plots (check plots) with a control variety are usually placed between the plots. The frequency of check plots with the control variety is usually between 5% and 20% of the total number of plots. There are two main problems that have to be considered in the experiment, i.e. density of check plots and arrangement of them, random or systematic.
The main tool of exploring information resulting from geometry of check plots will be based on a response surface methodology. At the beginning we will try to identify response surface characterizing the environment of experiment. The obtained (estimated) response surface will be then used to adjust the observations for genotypes. It is assumed that observation is a sum of an effect resulting from environment (soil fertility) and an effect of the genotype. The difference between forecast of the observation in the given places estimated by the response surface and the observation of the experimental field is then treated as the estimate of the genotype effect. Then the ranking of genotypes is made on the basis of those differences.

Multivariate statistical process control for mixed-type data based on Gower distance

Gaj Vidmar and Neža Majdič

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

Multivariate statistical process control (MV SPC) based on mixed-type data is a very recent and relatively little developed field. The usual approach to MV SPC addresses measurement data by constructing a Shewhart chart based on the Hotelling's T-squared statistic. We review the possibilities for MV SPC with mixed-type data (i.e., when some of the variables describing the process are numeric and some categorical) and identify three main approaches. The first approach is multivariate outlier detection for mixed data, for which several algorithms have been recently proposed in the field of data mining. The second approach is dimensionality reduction (via PCA, MDS or ICA) yielding numeric dimensions, from which T-squared control charts (or multivariate EMWA or multivariate CUSUM charts) can be constructed. The third approach is based on measuring distances between mixed-data points using the Gower distance (also known as Gower's dissimilarity coefficient, Gower's index or Gower's general coefficient of similarity) and then constructing T-squared charts, D-squared charts (based on support vector data description, SVDD) or K-squared charts (based on k-nearest neighbours data description, kNNDD), whereby the control limits for the D-squared and K-squared charts are established via bootstrapping. We present a pilot application of the Gower distance approach to health-care quality monitoring.

Mathematical Statistics

Optimal unbiased estimates for $P\{X < Y\}$

Marko Obradović , Bojana Milošević and Milan Jovanović

Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
marcone@matf.bg.ac.rs, bojana@matf.bg.ac.rs, mjovanovic@matf.bg.ac.rs

In reliability theory, one of the main problems is estimating parameter $R = P\{X < Y\}$. In practice, the variable X is called stress and Y is called strength. In this paper we shall present UMVUEs for R in different cases i.e. for different distributions of X and Y. Some of them are already existing and some are original.

Transmuted generalized Pareto distribution

Faton Merovci

Department of Mathematics, University of Prishtina, Prishtina, Kosovo
fmerovci@yahoo.com

In this article, we generalize the generalized Pareto distribution using the quadratic rank transmutation map studied by Shaw et al. (2009) to develop a transmuted generalized Pareto distribution. We provide a comprehensive description of the mathematical properties of the subject distribution along with its reliability behavior. The usefulness of the transmuted generalized Pareto distribution for modeling data is illustrated using real data.

Alpha-skew generalized t distribution

Sukru Acitas[1], *Birdal Senoglu*[2] *and Olcay Arslan*[2]

[1]Department of Statistics, Anadolu University, Eskisehir, Turkey
[2]Department of Statistics, Ankara University, Ankara, Turkey
sacitas@anadolu.edu.tr, senoglu@science.ankara.edu.tr, oarslan@ankara.edu.tr

In this study, an alternative skew version of the generalized t (GT) distribution (McDonald, J. B. and Newey, W. K., Partially adaptive estimation of regression models via the generalized t distribution, Econometric Theory, 1998, 4, 428-457) is proposed using the skewing procedure given by Elal-Olivero (Elal-Olivero, D., Alpha-skew-normal distribution, Proyecciones Journal of Mathematics, 2010, 29, 224-240). Some distributional properties of the proposed distribution are explored. Maximum likelihood estimation for the parameters of the proposed distribution is considered. In the application part of this study, Old Faithful Geyser data taken from literature, which is also skew and bimodal, is used to illustrate the modeling performance of the new distribution. The results show that proposed distribution accommodates the skewness and bimodality of the data. This data is available at R system (see also Arellano-Valle, R.B., Cortes, M. A. and Gomez, H. W., An Extension of the Epsilon-Skew Normal Distribution, Communications in Statistics - Theory and Methods, 2010, 39, 912-922).

Random search of stable member in a matrix polytope

Vakif Dzhafarov, Taner Buyukkoroglu and Serife Yilmaz

Anadolu University, Eskisehir, Turkey
vcaferov@anadolu.edu.tr, tbuyukkoroglu@anadolu.edu.tr,
serifeyilmaz@anadolu.edu.tr

A square matrix is called Hurwitz stable if all eigenvalues lie in the open left half of the complex plane. Hurwitz stable matrices play an important role in control theory. In the stabilization theory of linear systems the following questions arise: Given a matrix polytope is there a Hurwitz stable member in this polytope? If the existence is guaranteed how can we find the stable member? These problems are sometimes named as hard problems in control theory.
In this report we suggest random searching algorithm for a stable member. We consider interval type matrix polytopes as well as a family which is obtained as an affine image of a parametric box. Stochastic estimation is given and a number of illustrative examples are solved.

Modeling and Simulation

Bivariate models for time series of counts

Miroslav Ristić[1], *Aleksandar Nastić*[1] *and Predrag Popović*[2]

[1]Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia
[2]Faculty of Civil Engineering and Architecture, University of Niš, Niš, Serbia
miristic72@gmail.com, anastic78@gmail.com,
predrag.popovic@gaf.ni.ac.rs

Many processes from the real life can be seen as an evolution of a time series of counts. This paper presents different approaches for modeling bivariate autoregressive time series with nonnegative integer values. Bivariate models based on binomial and negative binomial thinning operators are developed. Their statistical measures are determined and discussed. As a data distribution differs from case to case different marginal distributions are suggested. Advantages of some models over the others are investigated. Models performances are tested on a real data. Practical importance of the models is proved.

Assessing the model performance of nonparametric fuzzy local polynomial regression with different bandwidth selection methods

Memmedaga Memmedli and Munevvere Yildiz

Anadolu University, Eskisehir, Turkey
mmammadov@anadolu.edu.tr, munevvere@hotmail.com

Fuzzy regression studies mainly have focused on fuzzy parametric regression models. In practical problems, expressing relation between explanatory variable and response variable with a certain parametric model is a very important restriction, and it can cause incorrect results. Hence, interest in nonparametric regression models have increased in recent years and for this purpose different type of models have been developed. For example k-nearest neighborhood regression, local polynomial models, kernel regression, different regression models with spline functions, etc. But, there are very few studies on fuzzy nonparametric regression models. It is necessary transform many notations and approaches of nonparametric regression models to fuzzy form for creating fuzzy nonparametric regression models. In this study, we considered the relationship between the smoothing parameter value and degree of polynomial as a simulation study in nonparametric fuzzy local polynomial regression. Besides the local linear models, local cubic models are also used in this simulation study. Fuzzy forms of cross validation and generalized cross validation criteria are developed for bandwidth selection. Performances of the models are evaluated in accordance with the selected bandwidth. Simulation results showed the degree of relationship between the bandwidth and order of polynomial in experimental way. As a result, it is obtained that the bandwidth size increases while order of polynomial increases. This leads to reduction of local fitting points and as a result overall operations decrease. Especially data with fluctuating; usage of local linear models are not convenient; in this case, it is necessary to reduce bandwidth but it increases computational complexity. Local bandwidth can increase while order of polynomial increases.

Conceptual model of perceived quality of insurance services: structural equation model application

Urban Šebjan and Polona Tominc

UM EPF, Maribor, Slovenia
urban.sebjan@gmail.com, polona.tominc@uni-mb.si

Since users are highly responsive to economic changes in the insurance market, this is the reason for studying the behavior of users of insurance services. As a result, in this paper we present a conceptual model of customer perceived service quality of insurance services. The conceptual model consists of four components, namely perception of innovation and perception of reputation of the insurance company, the perception of insurance premiums, and insurance coverage of insurance services. We have verified the conceptual model using structural equation modeling (SEM) as the assessment technique for a linear relationship between the components of a sample of 200 Slovenian users of the insurance services. This paper focuses on the operationalization of the components of perception of insurance services, on the connections between the constructs based on the standardized path coefficients, and it also assesses the reliability of the composite and the eliminated mean-variance, besides looking into convergent, discriminant and nomological validity of the constructs. Exploratory and confirmatory factor analysis showed one-dimensionality of the scales of each individual construct. We found that, (a) the more the innovation of insurance company is perceived, the more likely the positive reputation of insurance company is going to be, (b) the better reputation an insurance company has, the higher insurance premium of insurance services is observed, (c) the better reputation of insurance company there is, the higher insurance coverage of insurance services is perceived, (d) the higher insurance premium of insurance services is perceived, the higher insurance coverage of insurance services is also perceived. The highest standardized path coefficient was observed between the perceived innovation and the perceived reputation of insurance company, between the perceived reputation of insurance company and the perceived insurance premium of insurance services, and finally, between the perceived insurance premium and the perceived insurance coverage of insurance services.

A comparison of the recent ridge parameters in modeling GDP of Turkey

Yasin Asar[1], *Andan Karaibrahimoglu*[1] *and Aşir Genç*[2]

[1]Necmettin Erbakan University, Konya, Turkey
[2]Selcuk University, Konya, Turkey
yasar@konya.edu.tr, akara@konya.edu.tr, agenc@selcuk.edu.tr

Wide lands and underground sources are not the only wealth of the countries. Some indicators are calculated to put forth the exact power of countries. In order to compare the wealth of nations, one of the most important indicators is Gross Domestic Products (GDP). We have modeled GDP by cost components (at 1987 & 1998 prices) between the years 1968 and 2010, closely concerning the economy of Turkey in parallel to the growth and trends in the world. We have explained GDP using some parameters in the axes of foreign trade and production by multiple linear regression. But we have found that the model has a collinearity problem resulted by diagnose. And also we have detected the sources of collinearity problem. There are many methods and new approaches in the remedy of multicollinearity problem detected in the model. Multicollinearity is a big problem in the multiple regression in which the predictor variables are themselves highly correlated. Ridge regression is one of the widely used remedial methods of the multicollinearity problem. In this study, we will compare some the new approaches in selecting the parameter k for the solution proposed by Hoerl, A. E., Kennard, R. W. (Technometrics, 1970a, 12:55–67), Kibria, B. M. G. (Communications in Statistics, 2003, B(2)32:419–435), Khalaf, G., Shukur, G. (Communications in Statistics, 2005, A34:1177–1182) and Dorugade, A.V., Kashid, D.N.,(International Journal of Applied Mathematical Sciences, 2010, 4 (9), 447–456) and determine the best parameter of them. We will also make a simulation study to evaluate the performance of these estimators based on the mean squared error (MSE) criterion.

Education I

Team teaching with student: completion of experimental study

Jerneja Šifrer, Zala Žvab and Matevž Bren

University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
jerneja.sifrer@fvv.uni-mb.si, zala.zvab@gmail.com,
matevz.bren@fvv.uni-mb.si

Team teaching with student has become every year's practice with the undergraduate students' class of Statistics at the Faculty of Criminal Justice and Security, University of Maribor. In year 2012/2013 we continued with an experimental study performed in the academic years 2010/11 and 2011/12. As shown in previous experiments, team teaching offers many advantages, primarily daily evaluation of teaching methods and students' achievements, and improves communication between professor and students. In our contribution we will show the comparison of all three years of experimental study: comparison of students' outcomes, the results of quantitative analysis of students' questionnaire on theirs experience on team teaching, and qualitative analysis on several benefits to students, student teacher and professor. Hypothesis tested are that team teaching contributes to the better students outcomes, more collaboration and every day students' work and more positive attitude of students towards the subject, where team teaching is being used.

Issues and challenges of statistical literacy and statistics education in the higher education

Peter Kovacs

Department of Statistics and Demography University of Szeged, Szeged, Hungary
kovacs.peter@eco.u-szeged.hu

Statistical literacy is very important to describe and understand phenomena in real life. The relevance of statistical literacy is indicated by several definitions and researches which are well established in the statistical literature. A very important question is what universities could do for statistical literacy? In order to improve statistical literacy, the institutions of higher education have three target groups: students, academic staff, and external actors (citizens, secondary school students, etc.). In the presentation I sum up the Hungarian activities for developing statistical literacy and statistical education within the framework of the International Statistical Literacy Project.
In the presentation I review the issues of teaching of statistic, the expected outcome concerning the society without the expected outcomes of the training programs, the expected level of statistical literacy of the non-statistician staff. I examine the kind of skills and knowledge which have added values to the efficiency of increasing statistical literacy concerning researchers and teachers. These include knowing the students' knowledgebase, special features of Y-Z generation, pedagogic and teaching methods, skills and ability of the teachers, information about the teaching methods trends and researches as well as, the attitude and the knowledgebase of the teacher. Finally I suggest activities by which the universities could improve of statistical literacy.

AFRICA, Managing Statistics

Mohamadu Sani

Accra Polytechnic, Accra, Ghana
sanimotot@yahoo.com

This work is as a result of over 4 years of intensive research on the way institutions in Africa manages statistics to influence positively on the lives of their people. The work considered as a case study six African countries namely Ghana, Zimbabwe, Kenya, Sudan, Nigeria, and South Africa over the past decade and how they have used their previous population censuses and other statistical data to influence the lives of their indigenes. The work seeks to unravel the myth as to whether statistical management in Africa is a mere formality or real. Also to what extent does statistics assist in planning people's livelihood in Africa? The work also tells readers the success story of a country like South Africa which has financially and logistically strengthened its statistical institutions. Finally, the author makes a strong case for the need for countries in Africa to consider having autonomous institutions responsible for dealing with issues on statistical management since this can to a large extent minimize considerably the influence of politicians on such institutions. The experience of the Ghana statistical service recently in the hands of opposition politicians when it recently declared Ghana a middle income earning state was also adequately looked at in this work.

Sampling Techniques and Data Collection

Data collection on prevalence and patterns of drug use among general Slovene population

Katja Rostohar and Darja Lavtar

National Institute of Public Health, Ljubljana, Slovenia
Katja.Rostohar@ivz-rs.si, Darja.Lavtar@ivz-rs.si

The National Institute of Public Health implemented a Survey on the use of tobacco, alcohol and other drugs 2011-2012, which was conducted among 15,200 people aged between 15 and 64 years. An extensive survey of approx. 180 issues was responded by almost half of the people invited: around 40% online, 30% by telephone and 30% through personal interviews. The questions were mostly about tobacco, alcohol, drug use and attitudes related to them. Some of the risk behaviours studied in the research are negatively characterized in the society or even prohibited (e.g. binge drinking or heroin use) and have low prevalence rates. In the study we looked at how we can provide adequate data quality regarding accessibility of respondents. In order to reach individuals selected in the sample, three different modes of interview were offered to them. The majority of respondents to the survey were younger than 24; they were mostly women, and some of the respondents were still attending school. We also found out that the largest share of item non-response was among those surveys completed online, while this proportion is very low in telephone and personal interviews. The data was weighted according to the population margins. We compared the proportions of different risk behaviours according to gender, age categories, educational attainment and labour status. We found out that the proportions vary according to the selected demographic criteria, as well as regarding the mode of interviewing used in the survey. In the end we will present the advantages and disadvantages of different modes of interviews in the assessment of undesirable and risk behaviours in the population.

The effect of sample size and weighting procedures on final estimators of health indicators

Metka Zaletel, Katja Rostohar, Aleš Korosec and Darja Lavtar

National Institute of Public Health, Ljubljana, Slovenia
Metka.Zaletel@ivz-rs.si, Katja.Rostohar@ivz-rs.si,
Ales.Korosec@ivz-rs.si, Darja.Lavtar@ivz-rs.si

Due to low and medium prevalence rates of some diseases and behaviour patterns, sample sizes of health surveys are usually high in comparison to other social surveys. Therefore financial and organizational management of the survey causes high burden on the conducting institution which consequently leads to low frequency of these surveys and slow reaction to rapid changes in the society. On the basis of simulations, we have analyzed the impact of sample size to achieve the acceptable level of accuracy of the final estimators and also the impact of weighting procedures to the representativeness of the estimators. The survey on the use of alcohol, tobacco and other drugs, conducted by National Institute of Public Health in 2011/12, was used as the empirical data with the large sample and as the starting point for the simulations. Based on the results of simulations for expected data variability we estimated the minimum sample size needed for different estimators of key variables, assuming the acceptable level of precision. Therefore we will present calculations of the smallest samples that are acceptable for final estimators of health indicators. At the same time we analyzed the effect of weighting procedures on the final precision of the indicators for different sample sizes for some of the key indicators. We used the empirical data where the weighting procedures are partially defined by sampling design and the sampling frame. The impact of weight varies according to the sample size. We will demonstrate the dynamics of the impact of weighting as well as the possible application and recommendations will be shown. In the conclusions we will present the application of the above findings in the future surveys in Slovenia.

Regional input-output tables for the Czech Republic

Jaroslav Sixta, Lenka Hudrlikova and Jakub Fischer

University of Economics in Prague, Prague, Czech Republic
sixta@vse.cz, lhudrlikova@gmail.com, fischerj@vse.cz

Compilation of Regional Input-Output Tables (RIOTs) is a quite demanding issue and therefore it is very rarely done even the methodology is known. RIOTs offer many possibilities of analyses ranging from a very simple structural analysis to complicated input-output models. The connection of the region with regional economic data provides the much higher possibilities of the practical use of such models. On the contrary to national Symmetric Input-Output tables, RIOTs comprise interregional export and import and it is the most difficult issue. The links to other regions including other countries are very difficult to directly record and some simplifying methods have to be used. Finally, a balanced approach between production method, expenditure method and income method of estimation of gross domestic product is found. The paper describes the methodology and data prepared for the Czech Republic.

Estimation in adaptive cluster sampling using auxiliary information

Muhammad Noor-ul-amin

Universite de Bourgogne, Dijon, France
nooramin.stats@gmail.com

The ratio estimators are often considered when the relationship between variable of interest and auxiliary variable is linear. It is often the case in practical surveys that variable of interest has non-linear relationship with auxiliary variable. In this study, we proposed a ratio estimator in adaptive cluster sampling using the transformed populations. The estimator is based on two auxiliary variables when one auxiliary variable has linear relationship, while other auxiliary variable has non-linear relationship with study variable. The performance of proposed estimator is evaluated by comparing with estimators available in literature. The populations have been generated using the stochastic models. The results from simulation studies have signify the adequate performance of the proposed estimator.

Invited Lecture

Clustered data inference when the cluster size is potentially infomative

Somnath Datta

University of Louisville, Louisville, United States of America
somnath.datta@louisville.edu

We discuss how to extend parametric and nonparametric inference procedures when the classical assumption of independence is violated due to clustering. Clustered data arise in a number of practical applications where observations belonging to different clusters are independent but observations within the same cluster are dependent. While making adjustments for possible cluster dependence, one should also be aware of the "informative cluster size" phenomenon which occurs when the size of the cluster is a random variable that is correlated to the outcome distribution within a cluster, often through a cluster specific latent factor. We demonstrate the correct inference procedures under various scenarios.

Social Science Methodology

Statistical approach to environmental impacts analysis: the case of greenhouse gas emissions

Žiga Kotnik[1], *Maja Klun*[1] *and Damjan Škulj*[2]

[1]Faculty of Administration, University of Ljubljana, Ljubljana, Slovenia
[2]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
ziga.kotnik@fu.uni-lj.si, maja.klun@fu.uni-lj.si,
damjan.skulj@fdv.uni-lj.si

Global warming is a global problem. Air emissions contribute to the degradation of flora and fauna, irrespective of their origin. A ton of CO2 emissions produced in the European Union or in the U.S. causes the same damage to the environment as a ton of CO2 emissions produced in China or India. In the long run the environmental damage sustained by the state depends on the level of global emissions. The global nature of the problem calls for a coordinated global co-operation if the risk of catastrophic climate change is to be tackled effectively. Therefore, the realization of international environmental agreements is essential for the efficient and effective overall coordination since they set up concrete and realistic environmental targets. Notwithstanding the fact that some countries have already introduced higher environmental standards, others must accede to international cooperation as well, since only this can be a significant impact in the process of climate change.
The main issue we adress in our research is whether economically more advanced economies use environmental expenditues more effective and efficient in achieving environmental objectives in comparison with economically less developed EU economies. This issue was estimated and generalized by a specifically tailored statistical model and tested on the selected EU Member States. Our analysis demonstrates the influence between environmental indicators, i.e. environmental expenditures on environmental impacts. After careful literature based selection of environmental indicators different econometrical models were tested using a regression analysis. In the empirical analysis we test the difference between new and old EU Member States. We assume that the new EU Member States will show notable progress in effective and efficient use of environmental expenditures due to their slower implementation of the EU environmental regulations in their pre-accession period.

Use of different reliability and validity methods

Jerneja Šifrer and Matevž Bren

University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
jerneja.sifrer@fvv.uni-mb.si, matevz.bren@fvv.uni-mb.si

This contribution is based on the construct "why people obey the law", developed by university professor at New York University Tom R. Tyler. Several follow-up studies, adopting Tyler's theory found out that Tyler's process-based model of policing have left basic measurement questions unanswered. With these findings we prepared questionnaire, which was used in pilot study among 479 students. The purpose of this pilot study was to test the construct for its reliability and validity. We defined and assessed different methods used for determining reliability and validity of the construct. We assessed reliability with three different coefficients: Cronbach's coefficient alpha, coefficient theta with the use of principal components analysis, and coefficient omega with the use of factor analysis. For validity assessment we perform exploratory factor analysis in SPSS to determine the dimensionality of individual multi-item theoretical variable, and we perform confirmatory factor analysis with AMOS to determine the ability of a predefined factor model to fit an observed data set. For testing the model adequacy we used hi-square goodness-of-fit test. We found out that different results emerge using different methods. We also showed that Cronbach's alpha as a measure of homogeneity only partly determines reliability of the instrument and that the validity of the construct can be improved with the factor analysis. In accordance with our findings we proposed more reliable and valid instrument.

Underlying patterns in the profile: a multidimensional scaling approach to profile analysis.

Patrik P. Bratkovič

Department of Psychology, University of Ljubljana, Ljubljana, Slovenia
patrik.bratkovic@gmail.com

Profile analysis via Multidimensional scaling (PAMS) is a relatively unknown technique for finding underlying patterns in responses to psychological tests. Much of it could be attributed to that the technique did not have any objective way of choosing which variables would be included in a profile. In recent years, bootstrapping of the variables removed the subjectivity, but some questions still remain. This contribution focuses on presenting PAMS, how it is used, and potential areas of improvement.
A real data example with previously collected data will be used to outline the steps of the procedure, the results, and the steps that are being taken to assure that the extracted profiles are as accurate as possible. The data is from a Big Five questionnaire, and earlier research has found that women consistently score higher than men in on the agreeableness, extraversion and openness scales while slightly lower on emotional stability. While the empirical results indicate that men scored higher on the BFQ this time around, the set up hypothesis that PAMS would be able to find a profile similar to previous results, as well as finding that women were generally better aligned with that prototypical profile could not be rejected.

Measuring and visualizing change in cross-sectional time series

Aleš Žiberna

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

The talk will focus on several options for measuring and visualizing the change in several time series measured on a sample of countries. The main goal of the research is to come up with ways of measuring and visualizing change that supports cross-country comparisons, however in such a way that changes in two countries can be determined to be similar even if the levels of variables and therefore also changes are vastly different (e.g. number of soldiers in France and Slovenia) and if the main ”changes” appear in different (yet close) time points and are of different lengths. The goal is to obtain visualization that supports comparisons of “profiles” across countries and find transformations or similarity measures that support clustering in terms of similarities of those profiles. In addition, we want to construct at least two indices of change. One should measure total change across several variables, while at the same time it should not be sensitive to changes that are “reversed” in the next time point. The second index should measure change in a certain direction, either specified (individually for each variable) as up/down change or towards a certain direction.

Measurement

Application of discriminant analysis and multinomial logistic regression in investigation of regional disparities in Serbia

Valentina T. Sokolovska[1] *and Emilija Nikolić-Djorić*[2]

[1]Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia
[2]Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia
valentina.sokolovska25@gmail.com, emily@polj.uns.ac.rs

Based on our previous research, it is determined that a part of the population without education in total population is a demographic characteristic, according to which the regions in Serbia most diverge. Based on this indicator, great similarities are found between the regions of Šumadija and West Serbia, and East and South Serbia, in contrast to Belgrade and Vojvodina, and East and South Serbia. Considering the fact that this research is conducted based on the data from census in 2002, the goal of this thesis is to analyse regions of Serbia according to the census of 2011, using the same indicators. Time distance of ten years would show if there are any changes in demographic and economic indicators of the Serbian population and if regional divergences grow or diminish and by which factor they do so. For this thesis, the methods of descriptive discriminatory analysis are implemented, as well as multinomial logistic regression.

The forward search algorithm and confirmatory factor analysis

Aleš Toman

University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
ales.toman@fmf.uni-lj.si

The forward search algorithm is an iterative and graphical method for data exploration and robust parameter estimation. It helps us to identify observations in the sample that have disproportionately large impact on the subsequent statistical inference.
The algorithm splits the data into a clean and outlier-free small subset called the basic set and a bigger subset of potential outliers and influential observations. It proceeds with introducing new observations to the basic set until all observations are included. Several estimates and indices can be monitored during the whole procedure to reveal the true structure of the data.
There are three possible ways to apply the algorithm in the context of confirmatory factor analysis:
1) We can identify observations that are distant from the bulk of the data, i.e. outliers. 2) We can identify the observations with disproportionately large impact on covariance matrix estimate. 3) We can identify the observations with large impact on factor model parameters, i.e. influential observations.
The structure and the performance of the algorithm will be demonstrated on a simple example.

Vernon versus Carroll: confirmatory evidence about the structure of intellectual abilities from higher education standardized tests in Costa Rica

Eiliana Montero-Rojas

University of Costa Rica, San José, Costa Rica
eilianamontero@gmail.com

Two hierarchical models were tested empirically about the factor configuration implied by two alternative theories of the structure of intellectual abilities, the Three Stratum Theory by John B. Carroll and the verbal-educative and spatial-perceptual-practical model by Philip E. Vernon. Data came from 5 standardized tests used at the University of Costa Rica for selection and diagnostic purposes at the entrance level. The sample size was 131 students. Using third order confirmatory factor analyses, the hierarchical factorial configurations inferred from both theories were estimated and their fit was compared, in order to conclude which was more empirically plausible for these data. The five tests are: fluid intelligence (in the line of Cattell's Factor G), quantitative abilities, writing, reasoning in verbal contexts and, reasoning in Math contexts. All of them use the multiple choice format. Parceling was used to build the indicators for each of the constructs implied for both theories, also contributed to achieve more statistical power. Content based parcels showed better fit than parcels composed randomly within each test Models based on Vernon theory showed, with one exception, consistently better fit in Chi-square, RMSEA, GFI, AGFI, AIC y BIC indicators, compared with those based on Carroll's, even though the differences can be considered relatively small. Additionally, the plausibility of two alternative hypotheses was tested under both theories, regarding the items for the two tests of reasoning. The hypothesis that grouped these items in thematic or content constructs along with the quantitative and writing tests show better fit in both theories than the grouping under a general ability or reasoning construct. The best model in terms of consistency and fit was the configuration implied by Vernon with thematic or content grouping for the reasoning tests, with the following fit indexes: Chi-Square=268.04 (P=0.035), RMSEA=0.037, GFI=0.85, AGFI=0.82, RMR=0.063, PGFI=0.70, NFI=0.90, NNFI=0.97.

Education II

House of cards: a mixed-methods study on why behavioral science students fail or succeed an introductory statistics course

Floryt Van Wesel and Jan B. Hoeksma

Department of Methods, Faculty of Psychology and Education, VU University Amsterdam, Amsterdam, The Netherlands
f.van.wesel@vu.nl, j.b.hoeksma@vu.nl

Statistics is a major component in behavioral science education. However, many students underestimate its importance and fail its courses. Although modern teaching strategies (such as web-based applets and animations) have facilitated a better understanding of the basic principles of probability, introductory statistics still is a stumbling block for many of these students. This mixed-methods study investigates students' learning behavior and experiences leading to success or failure in an introductory statistics course. For this purpose ten freshmen university psychology and pedagogic students in the Netherlands were examined over a twelve-week period during their first statistics course. Qualitative data were collected weekly during two focus-group meetings and quantitative data consisted of grades obtained for tests and assignments during the course. Findings show that the image 'statistics' has amongst these students plays a dominant role in how they experience the course: A more positive image is related to higher grades, more time spend studying, a positive attitude towards the course matter, and less drop out and course failure. More interestingly, the major process identified concerning failure or success in the course follows the principle of building a house of cards: When one card is misplaced the whole house collapses, i.e., when they don't fully understand one of the basic principles of probability, they lose track of the material, subsequently reducing their self confidence and motivation, resulting in drop out or failure. From these findings recommendations are formulated on how to consolidate the house of cards with the aim to reduce course failure.

Motivation for learning statistics using SPSS. A case study at the Faculty of Organizational Sciences, University of Maribor

Alenka Brezavšček, Petra Šparl and Anja Žnidaršič

Faculty of Organizational Sciences, University of Maribor, Kranj, Slovenia
alenka.brezavscek@fov.uni-mb.si, petra.sparl@fov.uni-mb.si,
anja.znidarsic@fov.uni-mb.si

Educators need to know how to motivate students to learn statistics and to use statistical software, which can provide the practical skills necessary to analyze data and make informed decisions. Using a sample of students from the Faculty of Organizational Sciences, University of Maribor, we will perform an investigation of students' motivation to learn statistics using the statistical package SPSS. In the research, the survey which is composed of two questionnaires from other studies was used. In the paper, the obtained results will be presented. The validity of the questionnaire will be analyzed. The differences in perceived usefulness of SPSS between the first and the second-degree students as well as the dependence of attitude toward statistics on various factors related to using SPSS will be discussed.

An integrative approach to teaching statistics in social sciences: the example of EU political space

Jaro Lajovic

ro sigma research and statistics, Ljubljana, Slovenia
jaro.lajovic@rosigma.si

In teaching statistics an important challenge is the common approach that is primarily focused on ”how” (i.e. on formal foundations of the methodology) while paying less attention to ”why” (i.e. to showing the practical benefits of the methodology). This challenge can be particularly pronounced in human sciences studies where students may not feel attracted to delving into mathematical backgrounds.
Important points in overcoming this obstacle - and thus stressing the ”why” - include: using appropriate, real-life data sets; employing illustrative ways of analytical results representation (primarily visualisation techniques); and using appropriate (i.a. easily accessible) software environment.
We present an R software module (prepared for students of social sciences) integrating: (a) a real, up-to-date, interesting data set; (b) intuitively interpretable visualisations; (c) an open-source, free program environment; and (d) online access to web databases, to provide students with insight into the EU political space using/introducing the methods of hierachical clustering and principal components analysis.
The module is based on (a) the freely accessible Chapel Hill Expert Survey database on the EU political parties; (b) the R rgl library for 3D representations within (c) the R programming package (enhanced with the ade4TkGUI and FactoMineR libraries), the module being implemented as an additional drop-down menu in the R GUI; complemented with (d) on-line access to the European Election Database and Inter-Parliamentary Union Database.

Are results of university entrance-exams in Israel good predictors of success in B.A. programs?

Tal Shahor, Nissim Ben David and Tavor Tchai

The Academic College of Emek Yezreel, Emek Yezreel, Israel
tals@yvc.ac.il, nissimb@yvc.ac.il, tchait@yvc.ac.il

This study examines the relationship between the scores of the psychometric test and its various parts, against the actual success of undergraduate economics majors of the Israeli Emek Yezrael Academic College. The psychometric test examined in this study consists of three parts: a quantitative part, a Hebrew part, and a foreign language (English) part. The final score is made up of the weighted average of the three sections. Until now, admission to the college's economics department has been based upon this weighted score. It has been suggested, however, that in the case of the economics department, the quantitative part alone may allow for a better prediction of the chances of success. This study aims to test this hypothesis. Examination of the validity of psychometric tests is usually conducted on samples which have undergone prior selection (selected sample). Therefore the coefficients of the regression may be bias. To tackle this problem, we employed in this study several circumventing methods: first, we used the fact that the college's economics department does enroll some students who haven't met the admissions criteria and have been accepted as exceptions. Using dummy variables, we compared the level of success of those students with the levels of success of standard students. Additionally, we also applied in the study a 'propensity score matching' technique, in which we examined how final score values of similar 'student pairs' relate to their psychometric scores, while keeping their matriculation scores fixed. The results indicate no significant difference between the predictions of the weighted scores and the quantitative scores. In both cases, the coefficient is positive and significant, albeit small. It was also found that students with lower psychometric scores can also succeed in their studies, and therefore, it is possible to apply more flexibility into admission requirements

Invited Lecture

Choosing mode of data collection for surveys to maximize data quality

Jon Krosnick

Stanford University, Stanford, United States of America
krosnick@stanford.edu

Survey researchers can now collect data in one of four modes: face-to-face interviewing, telephone interviewing, self-administered paper questionnaires, and self-administered computer-presented questionnaires. It might seem that question-answering will be cognitively equivalent in all four modes, but psychological theory anticipates differences between the modes in response accuracy for a variety of reasons. This lecture will begin with a review of this theoretical background and will then review the findings of empirical comparing data collected in various modes in order to make recommendations about mode selection. Part of the discussion will compare the findings of online surveys of the general public done with probability samples, recruited through RDD phone calls or face-to-face visits to respondents' homes (as has been done by organizations in the U.S. Sweden, and Germany, and is being done in Norway and perhaps other nations as well) vs. online surveys done with non-probability samples of people who volunteer to answer questionnaires occasionally for money or gifts or prizes (in response to online ads or email invitations sent to non-representative groups of people). The quality of data obtained from these various methods will be contrasted.

Biostatistics and Bioinformatics I

Analysing disease recurrence with missing at risk information

Tomaž Štupnik[1] *and Maja Pohar Perme*[2]

[1]Department of Thoracic Surgery, Univerzitetni klinični center Ljubljana, Ljubljana, Slovenia
[2]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
tomaz.stupnik@kclj.si, maja.pohar@mf.uni-lj.si

When analysing time to disease recurrence, we sometimes stumble over data where we have absolutely no idea whether the studied patients are still alive or not. We only know their interval-censored times to disease recurrence. This may be due to a benign nature of a disease where patients are only seen at recurrences, for example in esophageal achalasia - with an approximately 50% recurrence rate at 10 years. Another example may be a poorly designed national registry of a benign disease or medical device implantation, with insufficient patient identifiers to obtain their dead/alive status.
When the average time to disease recurrence is long enough in comparison to the expected survival of the patients, the statistical analysis of such data may be significantly biased. Under the assumption that the expected survival of an individual is not influenced by the disease itself, we may want to reduce this bias by using the general population mortality tables. We show why the intuitive solution of simply censoring the patients with their expected survival time does not give unbiased estimates. We provide an alternative framework that allows for unbiased estimation of the usual quantities of interest in survival analysis. Our results are supported by simulations and by real data examples.

Effect of informative censoring on net survival estimates

Anamarija Rebolj Kodre and Maja Pohar Perme

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
anamarija.rebolj@mf.uni-lj.si, maja.pohar@mf.uni-lj.si

Estimation of relative survival probability has become one of the the most basic steps when reporting cancer survival statistics. Its aim is to evaluate the burden of a disease using the observed survival data and population mortality tables. It has been recently proven that the only generally appropriate measure of cancer burden is net survival, which estimates the probability of surviving cancer in a hypothetical situation when cancer is the only cause of death. In practice, the age structure of exposed cohort may change in calendar time. This may happen since both the incidence of cancer and the size of the population subgroups may change due to the long follow-up time. Censoring due to the end of study thus becomes informative. The correction of the net survival estimator to overcome the problem of informative censoring is proposed. Its properties will be illustrated on simulated and real data.

Methods of tied events handling in Cox proportional hazard model

Jadwiga Borucka

Warsaw School of Economics, Warsaw, Poland
jadwiga.borucka@gmail.com

The aim of the current paper is to present and compare methods that have been developed to handle tied events in Cox proportional hazard model estimation. Cox proportional hazard model is one of the most popular methods used in 'time-to-event' analysis. It was introduced by Cox in 1972 as the first model in the class of semiparametric models used in survival analysis. The model is based on several restrictive assumptions. Among others, it assumes that there are no tied events. Ideally, time variable is assumed to be continuous, however in fact time is usually being measured in hours, days, years, etc. thus it is nothing unusual to observe two events happening at the same time (such events are called 'tied events'). So far several methods have been developed that enable modification of partial likelihood function proposed by Cox to estimate the model in case of tied events existence.
There are four most commonly used methods in order to adjust partial likelihood function for tied events purpose. There are: exact expression derived by Kalbfleisch and Prentice (1980), Breslow approximation (1974), Efron approximation (1977) and discrete model. The exact method assumes that ties result from lack of precision in measuring survival time and any of possible d! arrangements of d tied events at the moment t can happen with exactly the same probability. Breslow and Efron approximations are based on the partial likelihood function modification while discrete model replaces the proportional hazard model with the discrete logistic model (mathematical formulas will be included in the full presentation). All these methods are available in SAS PHREG procedure as option TIES in MODEL statement. In order to compare these methods, the same model was estimated 40 times, 10 times per each method. Dataset used for analysis comes from Krall, Uthoff and Harley (1975) who analyzed data for 65 patients from study on multiple myeloma. Results include fit statistics of each model and estimation time (real time and CPU) calculated as mean out of 10 estimations using each method.
On the basis of fit statistics it can be stated that discrete model gives the best results, however its estimation time is approximately 60% longer than for exact method which gives quite similar results in terms of fit statistics and even 40 times longer than Breslow method and 30 times longer as compared with Efron approximation. Estimation with the use of Breslow and Efron methods gives relatively quick results however fit statistics differ from those obtained with exact or discrete method to a large extent. Thus, in case of datasets with tied events, exact or discrete method should be chosen, however depending on the size of the dataset, number of tied events and time that can be spent on the given analysis, an appropriate approximation might be used instead.

Biostatistics and Bioinformatics II

Gaussian moralized reciprocal graph modeling for genomic data integration

Carel F.W. Peeters, Wessel N. van Wieringen and Mark A. van de Wiel

VU University Medical Center, Amsterdam, the Netherlands
cf.peeters@vumc.nl, w.vanwieringen@vumc.nl, mark.vdwiel@vumc.nl

Graphical modeling refers to a class of probabilistic models that utilizes graphs to express conditional (in)dependence relations between random variables. In the Gaussian case, conditional independence between a pair of variables corresponds to zero entries in the precision matrix. Hence, model selection efforts in Gaussian graphical models focus on uncovering the pattern of zeroes in the precision matrix. Since their inception graphical models have found widespread use in a plethora of fields, among which is the modeling of gene regulatory networks. It is on such networks that we focus, oft implying high-dimensional settings, i.e., situations in which the ratio of observations to variables is unsuitable for classical approaches.
Most graphical modeling efforts set out to graphically model an unstructured precision matrix, i.e., all variables are treated equally. Our methodological objective is to infer the (sparse) graph pattern from a collection of Gaussian random variables in which (a) certain subsets enjoy differential treatment and (b) reciprocality is allowed. That is, we infer (sparse) graphs from model-structured precision matrices. The Markov properties embodied by these graphs connect to the moralization of reciprocal graphs implied by nonrecursive simultaneous equation systems having both endogenous and exogenous variables. The framework we propose intends to be attractive in terms of scalability. Our interconnected applied goal is to bring the resulting framework to bear on integrative graphical modeling, focusing on realistic regulatory networks by the joint graphical modeling of data from multiple genomic platforms. The genomic platforms we consider regard messenger RiboNucleic Acid (mRNA) and microRNA expression data.

Detecting gene-longevity association using bivariate survival and genetic data.

Alexander Begun, Andrea Icks and Guido Giani

Institute of Biometrics and Epidemiology, German Diabetes Center, Düsseldorf, Germany
alexander.begun@ddz.uni-duesseldorf.de,
Andrea.Icks@ddz.uni-duesseldorf.de,
Guido.Giani@ddz.uni-duesseldorf.de

To identify gene-longevity association, information on genotype frequencies for two or more age groups is needed. Significant differences in the gene or allele frequencies in distinct age groups can indicate that respective genes are involved in life span determination. Survival data included in the basic "gene frequency" method allow the estimation of the hazard functions and relative risks for different genotypes. An extension of the relative risk model makes it possible to study the cohort effect and the effect of antagonistic pleiotropic. We compare parameter estimates and the power, calculated using the uni- and bivariate models applied to the bivariate genetic and longevity data.

Boosting high-dimensional classifiers

Rok Blagus and Lara Lusa

Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
rok.blagus@mf.uni-lj.si, lara.lusa@mf.uni-lj.si

Boosted classifiers combine the votes of many weak classifiers that perform only slightly better than a random guess to form a strong classifier. A lot of research in the low dimensional setting (the number of samples exceeds the number of variables) showed the effectiveness of boosting and also its surprising resistance to over-fitting. Boosting was used also in the setting where the number of variables exceeds the number of samples (high-dimensional data) and was not as efficient as in the low-dimensional setting; however, the reasons for its poor performance were not well explained. Another issue that to our knowledge has not been addressed yet is how boosting performs in the high-dimensional setting if data are class-imbalanced, i.e. when the number of samples in each class is different. We extensively study these problems by using simulated and real datasets and by comparing different boosting and bagging techniques.

We first explain the reasons for the poor performance of boosting in the high-dimensional setting and based on our findings we propose a straightforward, yet efficient, modification of the most standard AdaBoost.M1 algorithm that can perform well also in the high-dimensional setting. We show that AdaBoost.M1 algorithm with complex classifiers is inappropriate for high-dimensional data, unless the size of the training set is sufficiently large. The sufficient sample size is determined by the ratio of the number of variables and the number of samples and the magnitude of the between class difference; in our simulation settings the smallest number of samples where AdaBoost.M1 performed well was 200 and the required sample size was larger when the difference and/or the number of variables were larger. The performance of boosting in the high-dimensional setting can be improved if the cross-validated error rate is used when constructing the ensembles, or by using stumps (trees with only two terminal nodes) as weak classifiers.

Researchers should be careful when applying boosting to class imbalanced data as this can increase the class imbalance bias. This issue can be avoided by using boosting on previously down-sized training set, or by using more complex ensembles that combine boosting with bagging. In our analyses EasyEnsemble with stumps or with our modification of the AdaBoost.M1 algorithm yielded the best overall performance when the level of class-imbalance was not extreme, while a more straightforward ensemble combining the results of classification trees using many smaller and balanced training sets achieved a better performance when the class-imbalance was very large.

Biostatistics and Bioinformatics III

Fractal Dimension of the Stochastic Hybrid Lee-Carter Mortality Model

Andrzej Szymanski[1], *Agnieszka Rossa*[1] *and Leslaw Socha*[2]

[1]University of Lodz, Lodz, Poland
[2]Cardinal Stefan Wyszynski University in Warsaw, Warsaw, Poland
anszyman@math.uni.lodz.pl, agrossa@uni.lodz.pl,
leslawsocha@poczta.onet.pl

The analysis of mortality is one of the basic problems faced by mathematical demography. The crude age-specific period mortality rates $m_x(t)$ can be defined for each group x and for each year t. The mortality rates can be predicted by means of the so-called Lee-Carter stochastic mortality model (LC). On the other hand, it can be seen that the LC model may be expressed as a solution of the stochastic differential equation of the Black-Sholes type. When the performance of a mortality model is tested with empirical data, then its parameter estimates are found not to be constant in time – their values depend on the period under study. Therefore it seems to be reasonable to take into account various methods of switching times, not only the Janic-Ledwina adaptive test.
Our aim in the present paper is to substitute the method of Janic-Ledwina by the methods of fractal theory. The most popular parameter among fractal characteristics is fractal dimension, which can characterize so called strange attractor of the dynamical system, i.e. a set with a fractal structure towards which a solution of the system evolves over time. We try to estimate fractal dimension for the stochastic mortality model (EHLC) using fuzzy techniques. The next extension of the EHLC model is Milevsky and Promislov model. We will characterize the scalar hybrid generalized Milevsky-Promislov model by fractal dimension and estimate the model using the mortality data of Poland.

Early stroke recovery efficacy: a mathematical model

Lavoslav Ćaklović[1] *and Miljenka Jelena Jurašić*[2]

[1] Department of Mathematics, Faculty of Natural Sciences, Zagreb, Croatia
[2]Department of Neurology, University Hospital Center, Sestre milosrdnice, Zagreb, Croatia
caklovic@math.hr, mjjurasic@gmail.com

Stroke is the second overall leading cause of death, and the most important cause of adulthood disability in Croatia. Outcome evaluation is needed for better patient care and recovery. The aim of this study was to explore the new Multicriteria Decision Model – the Potential Method (PM) – in post-hoc treatment efficacy evaluation as well as determination of the most important factors that influence it.

One year stroke patient database was evaluated and several stroke scales analyzed: prestroke Barthel index (preBI), NIHSS, poststroke Barthel index (postBI) and modified Rankin scale (mRS). The new Efficiency scale (Eff) was constructed using ordinal weighted aggregation on the incomplete preference graphs generated by preBI, NIHSS, postBI and mRS, with the purpose to measure early post-stroke recovery efficacy. Eff scale has excellent properties. As the dead-living (D-L) discriminator Eff dominates NIHSS (AUC(Eff) = 0.989, AUC(NIHSS) = 8.23), and its deviance measure in logit regression model has a very high value of 78.5%.

A detailed Eff study in combination with age and gender was done, showing that women have significantly better recovery until 60 years. The main difficulty in all analysis was missing data, therefore factors' aggregations were done. The factors were: stroke manifestations (headache, epilepsy, loss of consciousness), dead-living, and stroke severity (minor, moderate, moderate-severe, severe) using NIHSS. A hierarchical clusterization was performed over factors' combinations (with Eff as the distance) and comparison to the results of the Canonical Correspondence Analysis over stroke symptoms with stroke manifestations as covariates. The chi-square proportion showed the one-dimensional feature of our model (81% accountable in first dimension).

CCA and NIHSS scale are well attuned and appear to measure the same phenomenon advocating that stroke manifestations and their combinations may serve as stroke severity predictors. Eff scale, also, shows that moderate-severe and severe stroke recovery is equally slow and difficult.

Local estimation of standardized incidence ratio applied to Slovenian cancer data

Tina Žagar, Vesna Zadnik and Maja Primic Žakelj

Institute of Oncology Ljubljana, Ljubljana, Slovenia
tzagar@onko-i.si, vzadnik@onko-i.si, mzakelj@onko-i.si

The objective of the approach named "local estimation of standardized incidence ratio (SIR)" is to produce maps using geocoded data. Most methods in public health field are based on data aggregated to geographical areas and only few on point data.
The whole study area (Slovenia in our case) is covered with grid points 1 km apart. Circle is centred at each of the fine grid locations and the cases and population data occurring within the circle are determined. Indirect age standardization is applied and SIR is calculated for each grid point covering the study area producing a smooth map of relative risk. The circle radius is not fixed in advance but is changing from intergrid distance (1 km) to chosen maximum radius with step 1 km until predetermined minimum population limit is reached. This way, for each grid point the calculated SIRs are based on (approximately) the same number of persons at risk giving more stable estimates. The drawback of the local SIR estimation is that the geocoded data are not always available. The benefits are that the arbitrary administrative areas are ignored and, above all, such high resolution maps can reveal more local risk patterns helping to identify areas where further investigation is needed.
Descriptive presentations of the cancer burden are enhanced by cancer maps making them an important tool in public health research. The cancer data were provided by the Cancer Registry of Republic of Slovenia and the geographic coordinates for cancer cases as well as for population were provided by the Central Population Register. Several cancer maps were produced and compared to some frequently used mapping approaches.

Statistical Applications - Biostatistics I

Predicting basal metabolic rate using multiple longitudinal anthropometric measurements

Siti Haslinda Mohd Din and Emmanuel Lesaffre

Erasmus University Medical Centre, Rotterdam, The Netherlands
s.bintimohddin@erasmusmc.nl, e.lesaffre@erasmusmc.nl

Longitudinal profiles are usually analyzed as response, but in our application they serve as predictor. More specifically, the evolutions of 22 anthropometric measurements over 3 years (every 6 months) are used as predictor to predict basal metabolic rate (BMR) recorded in a Malaysian study on 70 boys and 69 girls. To measure BMR a laborious procedure is required, and hence the research question was how well it could be predicted from body measurements that are easier to record. To this end we propose a joint modeling approach whereby the latent anthropometric evolutions , measured by their random intercept and slope, are estimated and at the same time used as predictors for BMR. While there are other possible approaches, such as classical and ridge regression, joint modeling has less biased and more efficient estimates of regression coefficients (J. Ibrahim et. al., Journal of Clinical Oncology, 2010, 28, 2796-2801). Furthermore, joint modeling deals better with missing longitudinal covariates than the standard approaches. In our work, we applied joint modeling in a Bayesian context involving Markov Chain Monte Carlo (MCMC) techniques. Thus our approach fits jointly linear mixed models on selected three anthropometric measurements and a classical linear regression to predict BMR.
Joint modeling is nowadays a popular approach to fit longitudinal and survival outcomes at the same time. Despite its attractiveness, diagnostic tools are lacking to verify the statistical assumptions made in the joint model. For instance, detecting influential observations in a joint modeling exercise is useful to discover whether subjects or observations have an unduly large impact on the estimation process. In this presentation Bayesian diagnostic checks, such as tools for detection outlying observations/subjects, are presented.

Indirect treatment comparisons in meta-analysis of clinical trials of antiepileptic drugs used as add-on therapy

Igor Locatelli[1], *Urban Bernat*[2] *and Mitja Kos*[1]

[1]Faculty of Pharmacy, University of Ljubljana, Ljubljana, Slovenia
[2]KRKA d.d., Novo mesto, Slovenia
Igor.Locatelli@ffa.uni-lj.si, urban.bernat@krka.biz,
Mitja.Kos@ffa.uni-lj.si

The majority of people with epilepsy have a good prognosis and their seizures can be well controlled with the use of a single antiepileptic drug. But up to 30% of patients develop refractory epilepsy. The purpose of this research is to evaluate the comparative effectiveness of antiepileptic drugs, when used as add-on treatment.
A systematic review of clinical trials of antiepileptic drugs indicated for add-on treatment was performed. Only randomized placebo controlled trials were included in meta-analysis. A 50% or higher reduction in total seizure frequency during a period of 12-24 weeks (study duration period) was set as antiepileptic treatment efficacy parameter. Meta-analysis was performed in Winbugs and MIX 1.7 software. A random effects model with indirect treatment comparison was applied by combining Der-Simonian and Laird method and Bucher method for indirect comparisons. The influence of dose-effect relationship was also included in meta-analysis model.
In total 21 trials were selected for meta-analysis. In total 6 antiepileptic drugs were used in these trials as add-on treatment, namely; eslicarbazepine (3 studies), lacosamide (3 studies), pregabaline (6 studies), retigabine (2 studies), tiagabine (3 studies), and zonisamide (4 studies). All 6 antiepileptic drugs were significantly more effective than placebo. Meta-analysis of direct comparison versus placebo showed that relative risk (95% confidence interval) for seizure frequency reduction was 1.8 (1.5 – 2.3) for eslicarbazepine, 1.7 (1.4 – 2.0) for lacosamide, 2.8 (2.1 – 3.9) for pregabaline, 2.1 (1.7 – 2.6) for retigabine, 3.2 (2.1 – 5.0) for tiagabine, and 1.8 (1.5 – 2.3) for zonisamide. The influence of dose-effect relationship was found for pregabaline and eslicarbazepine. Indirect comparison meta-analysis revealed that pregabalin 600 mg and tiagabine were the most effective antiepileptic drugs.

Dynamic graph generation from MS Excel using R: an implementation based on SVG and shiny

Črt Ahlin[1] *and Lara Lusa*[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
crt.ahlin@gmail.com, lara.lusa@mf.uni-lj.si

Spreadsheet programs are often used for data entry and manipulation. Although they usually include some datasome data plotting capabilities, customized and interactive graphs are not easily obtained. R statistical environment includes graphical facilities for data display that are extremely flexible. It recently became rather straightforward to create interactive graphics and web applications with R, using Scalable Vector Graphics (SVG) and the shiny package.
We developed some MS Excel spreadsheets that can be used for the generation of interactive plots using R. The spreadsheets embed macros written in MS Visual Basic for Applications (VBA), which are used to read the data and set the parameters that define some of the plots characteristics. To manage the connection between Excel and R we use the RExcel add-in for Excel together with the statconn server and rcom package, all of which were developed within the statconn project. The end user of medplot therefore does not need to be familiar with R.
The implemented plots can be used to visualize the results of repeated tests and the presence and intensity of symptoms for a cohort of subjects. The graphs contain dynamic components: one offers additional data about a plotted point via moving the mouse cursor (via SVG graphics), the other allows the user to input additional filters for plotting (via shiny R package).

cartHD: R package for the estimation and validation of classification trees with binary outcomes

Lara Lusa and Rok Blagus

Institute for Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
lara.lusa@mf.uni-lj.si, rok.blagus@mf.uni-lj.si

Classification trees are popular classifiers that are often used in medicine, computer science and many other fields where the aim is to define a rule that can be used to predict the class membership of new samples using the values of some predictor variables. Their appeal lies in their flexibility and in the simplicity of the derived rules, which can be easily presented graphically.
When many potential predictor variables are available, deriving the classification rule can be time consuming, as the fitting algorithm selects the best variables and their best splitting points, univariatly trying all of them. The computational burden is further aggravated when the classification tree needs to be fitted multiple times on subsets of the complete data, or with different weights. For example, boosting methods that use classification trees as their base classifiers vary the weights of the samples at each iteration, based on the classification results obtained in the previous steps. Also cross-validation, a resampling methods commonly used to estimate the predictive accuracy of the classifiers, fits the classifier multiple times on subsets of the complete data. While boosting and cross-validation algorithms are fairly easy to implement, often the available implementations are far from optimal from the computational point of view and do not take advantage of the fact that some quantities can be derived from the fit obtained on the complete data and thus evaluated only once.
We developed cartHD, an R package that reduces the computational burden encountered in when using the classification trees with boosting, cross-validation and multiple downsizing. The functions outperform the existing alternatives available in R when the number of variables is large. The package includes functions that can be used to plot the classification trees and is freely available on GitHub.

Econometrics

Robust standard errors for panel data: a general framework

Giovanni Millo

R&D Dept. Assicurazioni Generali, Trieste, Italy
giovanni_millo@generali.com

A comprehensive, modular and flexible framework is described for estimation of robust standard errors in panel data. Heteroskedasticity and autocorrelation robust estimators in the tradition of White (1980) and Arellano (1987) are brought together with the "SCC" mixing-fields based estimator of Driscoll and Kraay (1998), the unconditional "PCSE" estimator of Beck and Katz (1995), and the double-clustering approach of Cameron et al. (2011) and Thompson (2011), trying to bring together the applied literatures in macroeconometrics, finance, political science and accounting by demonstrating the common features of these apparently different approaches. The covariance estimators are integrated in the R package "plm" and allow robust specification and restriction testing over a number of different panel models.

Insurance industry and stochastic scenario

Dario Galdieri

Assicurazioni Generali, Trieste, Italy
dario_galdieri@generali.com

In the last years insurance industry has been focused on the importance of properly managing options and guarantees embedded in insurance contracts. Interest rates are in this period very low, which means that minimum interest rate guarantees have moved from being far out-of-the money to expiring in the-money.Furthermore, insurance firms operating within the European Union will be subject to the Solvency II directive, which places new demands on insurance companies. In particular the valuation of assets and liabilities now needs to be market consistent.One way to accomplish a market consistent valuation is through the use of an economic scenario generator (ESG), which creates stochastic scenarios of future asset returns.An assessment of how well the models applicable to different type of assets available in the ESG market capture prices of instruments traded on the market is made and I will show which are the instruments used to validate the stochastic scenarios in terms and market consistency and risk neutrality.

An estimate of the degree of interconnectedness between European regions

Davide Fiaschi and Angela Parenti

University of Pisa, Pisa, Italy
dfiaschi@ec.unipi.it, aparenti@ec.unipi.it

This paper provides a methodology to estimate the degree of economic interconnectedness across different regions, and applies such methodology to a sample of 199 European NUTS2 regions in the period 1980-2008. The first step consists in the estimate of a panel of volatility growth rates in each period for each region, under the assumption that regional income growth rates follow an autoregressive process; in the second step, via a generalized variance decomposition analysis, this panel is used to estimate the connectedness matrix (called adjacent matrix in network literature), which measures the interconnectedness between regional income, conditioned to how many periods ahead we allow the shocks to propagate across regions. The estimated connectedness appears very heterogeneous and not symmetric; the own connectedness is not very relevant (at most 14% of the variance is due to shocks arising and remaining in the same region, i.e. true idiosyncratic component of regional shocks); there exists a periphery of regions with high interconnectedness inward and outward; and, finally, the regions with the highest negative net value of connectedness are in the core of Europe (this means that they are net source of shocks for the other regions). Moreover, the country component is not very relevant (at most 20% of total variance is due to shocks arising in regions belonging to the same country), but very heterogeneous across countries (from 1% in Luxembourg to 20% in United Kingdom). Finally, the comparison of this connectedness matrix with some spatial matrices generally used in spatial econometrics reveals that an exogenous spatial matrix, as, e.g., a first-order or a second-order matrix, are far from representing the actual interconnectedness between European regions.

Are funding of pensions and economic growth directly linked? New empirical results for some OECD countries

Pietro Cavallini[1], *Gaetano Carmeci*[2] *and Giovanni Millo*[3]

[1]Belluno, Italy
[2]DEAMS, University of Trieste, Trieste, Italy
[3]Generali Research and Development, Trieste, Italy
pietro0cavallini@gmail.com, gaetano.carmeci@econ.units.it, Giovanni_Millo@Generali.com

We empirically test on a panel of OECD countries the hypothesis of a direct and positive link between funding of pensions and economic growth, which is based on the idea that richer pension systems can accelerate the development of the financial system and thus promote a more efficient capital allocation. We follow Davis and Hu (2008) [Davis and Hu (2008), Does funding of pensions stimulate economic growth?, Journal of Pension Economic and Finance, Cambridge University Press, vol. 7 (02), 221-249] in estimating a modified Cobb-Douglas production function, where pension fund assets are treated as a shift factor, but we criticize their results from an econometric point of view, since both the Dynamic OLS and Mean Group (MG) estimators are inadequate in the case of cross-sectional correlated residuals. Indeed, we find a highly significant level of correlation in the MG residuals across countries that we attribute to common global shocks driving per capita outputs. Therefore we adopt a more general approach suitable to the presence of a multifactor error structure. Our results exclude the existence of a long run cointegration relationship between autonomous (or total) pension fund assets and per capita output for our panel of OECD countries, unless, in contrast to the conclusion of the cross-sectional dependence test, we ignore it and assume independence of residuals.
Our results exclude the existence of a long run cointegration relationship between autonomous pension fund assets and per capita output for our panel of OECD countries, unless, in contrast to the conclusion of the cross-sectional dependence test, we ignore it and assume independence of residuals.

Statistical Applications - Biostatistics II

Woody vegetation structure on land abandoned from agricultural production in South Europe

Andreja Radović[1], *Thomas Wrbka*[2], *Kiril Vassilev*[3] *and Vassiliki Kati*[4]

[1]University of Zagreb, Faculty of Science, Zagreb, Croatia
[2]University of Vienna, Vienna, Austria
[3]Bulgarian Academy of Sciences and Arts, Sofia, Bulgaria
[4]University of Patras, Agrinio, Greece
andreja.radovic@biol.pmf.hr, thomas.wrbka@univie.ac.at, kiril5914@abv.bg, vkati@cc.uoi.gr

Agriculture, accounting for almost half of total European area, is one of the main driving forces of current distribution and ecological structure of biological communities. Socio-economic changes over the past few decades have led to major changes in the way of processing agricultural land. These changes can be classified as: 1) intensification of agricultural practices in the areas of good quality and 2) the abandonment of agricultural production in less attractive areas. Diverse studies provided evidence for subsequent propulsion of woody and shrub vegetation after land abandonment. The main research task in this study was revealing 3D woody plant structure on areas where the abandonment of agriculture occurs in south Europe and to determine whether the effect of the abandonment of agricultural production is consistent across the region.
Data on woody vegetation 3D structure and composition was collected on 72 randomly selected plots of 1km X 1km area, located in SE Europe (Greece, Bulgaria, Croatia and Albania). All 1x1km plots were covered by more than 50% by arable land in the period as far in the past data were available for each country, ranging from 1950ties for Greece to 1970ties for Albania The plots chosen reflect the land abandonment gradient in terms of woody vegetation cover, after the following four categories: 1: 0-25%, 2:25-50%, 3:50-75%, and 4:>75%. Our study tested how does 3D diversity and functional diversity of woody plants differ across countries in the region. Finally, we extracted the optimal set of landscape metrics that is best correlated with woody vegetation diversity.

Different methods for handling non-Gaussian longitudinal outcome subject to potentially random dropout

Ali Satty

University of Kwazulu-Natal, Pietermaritzburg, South Africa
alisatty1981@gmail.com

The present paper compares and contrasts several statistical methods for analyzing incomplete non-Gaussian longitudinal outcomes when the underlying study is subject to dropout. We focus here on binary outcomes. Similar results will likely apply to other data types as well but should, ideally, be the subject of additional research. The methods that are considered include weighted generalized estimating equations (WGEE), multiple imputation after generalized estimating equations (MI-GEE) and generalized linear mixed models (GLMM). The paper aims to explore the performance of the above methods in terms of handling dropouts that are missing at random (MAR). The methods are compared on simulated data. The correlated binary variables are generated from a logistic marginal model. Dropouts are generated under several different dropout rates and sample sizes. The comparison will be made through the evaluation of bias, accuracy and mean square error. MI-GEE was considerably robust, doing better than all the other methods in terms of the small and large sample sizes, regardless of the dropout rates.

The new mixture distribution models for wind speeds

Murat Erisoglu[1], *Ulku Erisoglu*[1], *Aydin Karakoca*[1] *and Serkan Akogul*[2]

[1]Necmettin Erbakan University, Konya, Turkey
[2]Selcuk University, Konya, Turkey
merisoglu@konya.edu.tr, ugokal@konya.edu.tr, akarakoca@konya.edu.tr, serkanakogul@gmail.com

The probability density function of wind speed is important in numerous wind energy applications. In practice, it is substantially important to describe the variation of wind speeds for optimizing the design of the systems resulting in less energy generating costs. Therefore, the appropriate distribution modeling of wind speed is very important. A variety of probability density functions have been used in literature to describe wind speed distributions such as Weibull, Gamma, Exponential, etc. If the frequency distribution of wind speed has heterogeneous structure, the standard probability distributions are not enough to modeling of wind speeds. In such cases, the mixture distribution models are used to modeling of wind speeds in the specialized literature on wind energy. The purpose of this paper is to show that the mixture of the same and different distributions of Weibull, Gamma, and Exponential distributions is the appropriate distribution for the wind speeds.

Statistical fit of empirical distributions to GPS derived animal movement data

Robert Mutwiri[1], *Thomas Achia*[1], *Henry Mwambi*[1], *Rob Slotow*[2] *and Vanak Vanak*[3]

[1]School of Mathematics, Statistics and Computing, University of KwaZulu Natal, Pietermartzburg, South Africa
[2]School of Life Sciences, University of KwaZulu Natal, Durban, South Africa
[3]Ashoka Trust for Research in Ecology and the Environment, Bangalore, India
robmathenge@gmail.com, achiat@ukzn.ac.za, mwambih@ukzn.ac.za, slotow@ukzn.ac.za, abivanak@gmail.com

A wide range of studies in movement ecology have used heavy-tailed distributions to analyze the distribution of step lengths of animal datasets collected through radio and GPS telemetry devices. Many of these studies use simple fitting methods to estimate parameters and test the goodness-of-fit of such distributions, leading to misleading conclusions. In this study we review and compare the suitability of candidate alternative heavy-tailed distributions: truncated power law, log-normal, exponential, gamma and Weibull distributions considered jointly with other circular distributions to decompose step lengths and turn angles into movement states in more complex models such as state space and hidden Markov models. We present a principled statistical approach for discerning and quantifying animal movement step lengths with applications to elephant movement data based on maximum likelihood and Kolmogorov Smirnov (KS) goodness-of-fit tests. We parameterize and compare the performance of these models at six temporal resolutions (30 minutes, 1 hour, 2 hours, 5 hours,12 hours and 24 hours) using data from one African elephant (Loxodonta africana). Based on the KS test, the elephant movement at 1, 2, 5, 12 and 24 hours follow a power law distribution and supports the findings of previous studies. However, the Weibull and lognormal distributions find support at 12 hours and 24 hours respectively while exponential distribution is not supported at all resolutions. In summary, our findings strongly support the need for testing alternative distributions when analyzing animal movement data and demonstrates the importance of heavy-tailed distributions in movement ecology theory.

Design of Experiments

Regular A-optimal spring balance weighting designs under special assumptions on errors

Bronisław Ceranka and Małgorzata Graczyk

Poznań University of Life Sciences, Poznań, Poland
bronicer@up.poznan.pl, magra@up.poznan.pl

The problems linked with an A-optimal spring balance weighting design under special assumptions: the errors are correlated and have equal variances are discussed. The topic is focus on the determining the lowest bound of the trace of inverse of the information matrix. The constructing method of the optimal design, based on the incidence matrices of balanced incomplete block designs, is presented.

SSD populations – statistical characteristics and practical approach in grain legumes

Maria Surma[1], *Tadeusz Adamski*[1], *Zygmunt Kaczmarek*[1], *Anetta Kuczynska*[1], *Karolina Krystkowiak*[1] *and Stanislaw F. Mejza*[2]

[1]Institute of Plant Genetics PAS, Poznan, Poland
[2]University of Life Science, Poznan, Poland
Msur@igr.poznan.pl, Tada@igr.poznan.pl, Zkac@igr.poznan.pl, Akuc@igr.poznan.pl, Kkry@igr.poznan.pl, Smejza@up.poznan.pl

In breeding of pea and lupins homozygous lines are needed to develop new cultivars. Generally, homozygosity can be attained by selfing successive generations or by hapoidization of hybrids using different in vitro techniques and doubling chromosomes in haploid plants. Pisum and Lupinus species are known to be recalcitrant to in vitro culture. In spite of numerous studies focused on the production of doubled haploids (DH), no important results have been noted to date. For shortening the breeding cycle, the single seed descent technique (SSD) in combination with in vitro culture of immature embryos may be an alternative to the DH system. Theoretical considerations showed that in the absence of gene linkage the frequency of recombinants in the DH and SSD populations is the same and as the result the means and variances of quantitative traits are expected to be identical. In the presence of linkage between genes conditioning metrical traits the frequency of recombinants in the SSD populations is expected to be higher than in DH lines. The aim of the present studies was to establish in vitro conditions for the culture of pea and lupin embryos as the first step in research aimed at shortening generation cycles in the SSD technique. The data for shoot and root growth during culture duration were statistically processed using 3-factor multivariate analysis of variance and related multivariate techniques. For each species hypotheses of no differences between media, temperature regimes and cultivars, as well corresponding interactions were tested. Based on statistical results in vitro conditions for embryo culture of pea and lupins were established. The study was supported by NATIONAL, MULTI-YEAR PROGRAM 2011-2015 “Improvement of domestic sources of plant protein, their production, economy and feeding technologies”, funded by the Polish Goverment and Polish Ministry of Agriculture and Rural Development.

Evaluation of parental forms on the basis of series of unreplicated experiments with their hybrids and standard varieties

Zygmunt Kaczmarek[1], *Elzbieta Adamska*[1], *Iwona Mejza*[2], *Henryk Wos*[3] *and Renata Trzeciak*[1]

[1]Institute of Plant Genetics PAS, Poznan, Poland
[2]Poznan University of Life Sciences, Poznan, Poland
[3]Plant Breeding Company Strzelce Ltd. Group PBAI, Breeding Division Borowo, Strzelce, Poland
Zkac@igr.poznan.pl, Eada@igr.poznan.pl, Imejza@up.poznan.pl, m_luty@hr-strzelce.pl, Rtrz@igr.poznan.pl

The statistical methods used for the analysis of results of a series of unreplicated experiments with the same set of genotypes conducted in incomplete block designs in several environments are presented. The analysis is based on Scheffe-type mixed model for observations. The various possibilities offered by the assumed model and the methods derived from it are indicated. In particular it is shown how inferences concerning individual genotypes i.e. standard varieties and hybrids obtained from the line x tester crossing system can be drawn. The line x tester system is of interest for breeders dealing with estimation and testing of general (g.c.a.) and specific (s.c.a.) combining abilities of parental forms. In the situation, when the observations in series of experiments concerns line x tester hybrids, a special attention is given to estimation and hypothesis testing problems concerning the genotype x environment interaction. Methods of statistical analysis presented in the paper allow to get information concerning the behaviour of individual hybrids as well as the combining ability effects of their parents in various environmental conditions, including assessment of stability and adaptability. The estimators of interesting genetic effects are given in the form of comparisons (contrasts) among hybrids. The methods presented in the paper are illustrated by its application in the analysis of results of a series of 4 experiments of winter oilseed rape, in which 3 standard varieties were replicated and $10x6 = 60$ MS line x restorer hybrids were unreplicated. The analysis described here is confirmed to data concerning grain yields. Calculations were made by means of the statistical program SERGEN.
The study was supported by project MR61 from Ministry of Agriculture and Rural Development in Poland

Multiple one-way ANCOVA when the distributions of both the covariates and the error terms are non-normal

Pelin Kasap[1] *and Birdal Senoglu*[2]

[1]Department of Statistics, Ondokuz Mayis University, Samsun, Turkey
[2]Department of Statistics, Ankara University, Ankara, Turkey
pelin.kasap@omu.edu.tr, senoglu@science.ankara.edu.tr

In this study, we obtain the explicit estimators of the parameters in multiple one-way ANCOVA model when the distribution of the covariates and the error terms are generalized logistic (GL) and long-tailed symmetric (LTS), respectively (Islam, M.Q. and Tiku, M.L., Multiple linear regression model with stochastic design variables, Journal of Applied Statistics, 2010, 37(6), 923-943; Kasap, P., Stochastic ANCOVA: Statistical Inference, Unpublished Ph.D. thesis, 2011, Ankara University, Ankara, Turkey). In the estimation procedure, we use the methodology known as the modified maximum likelihood (MML) originated by Tiku (Tiku, M.L., Estimating the mean and standard deviation from a censored normal sample, Biometrika, 1967, 54, 155-165; Tiku, M.L., Estimating the parameters of normal and logistic distributions from a censored normal sample, Austral. J. Stat., 1968, 10, 64-74). Based on these estimators we propose a new test statistic for testing the linear contrasts. Efficiencies and the robustness properties of the proposed estimators are compared with the traditional least squares (LS) estimators via Monte-Carlo simulation study for two covariates case. Simulation results show that the MML estimators are more efficient and robust than the LS estimators. A real life data taken from the literature is analyzed at the end of the study.

Workshop

Bayesian computation with INLA

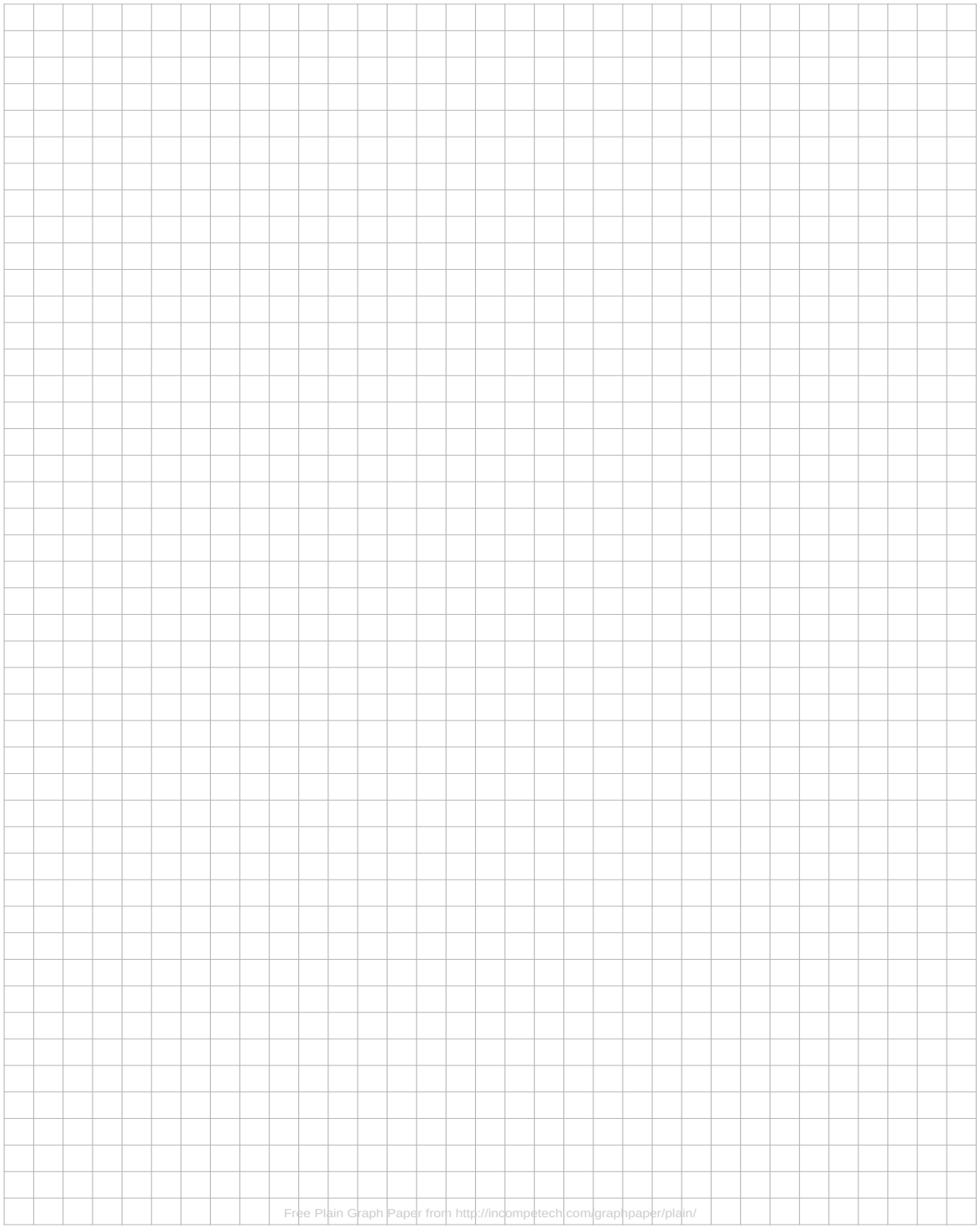
Thiago G. Martins

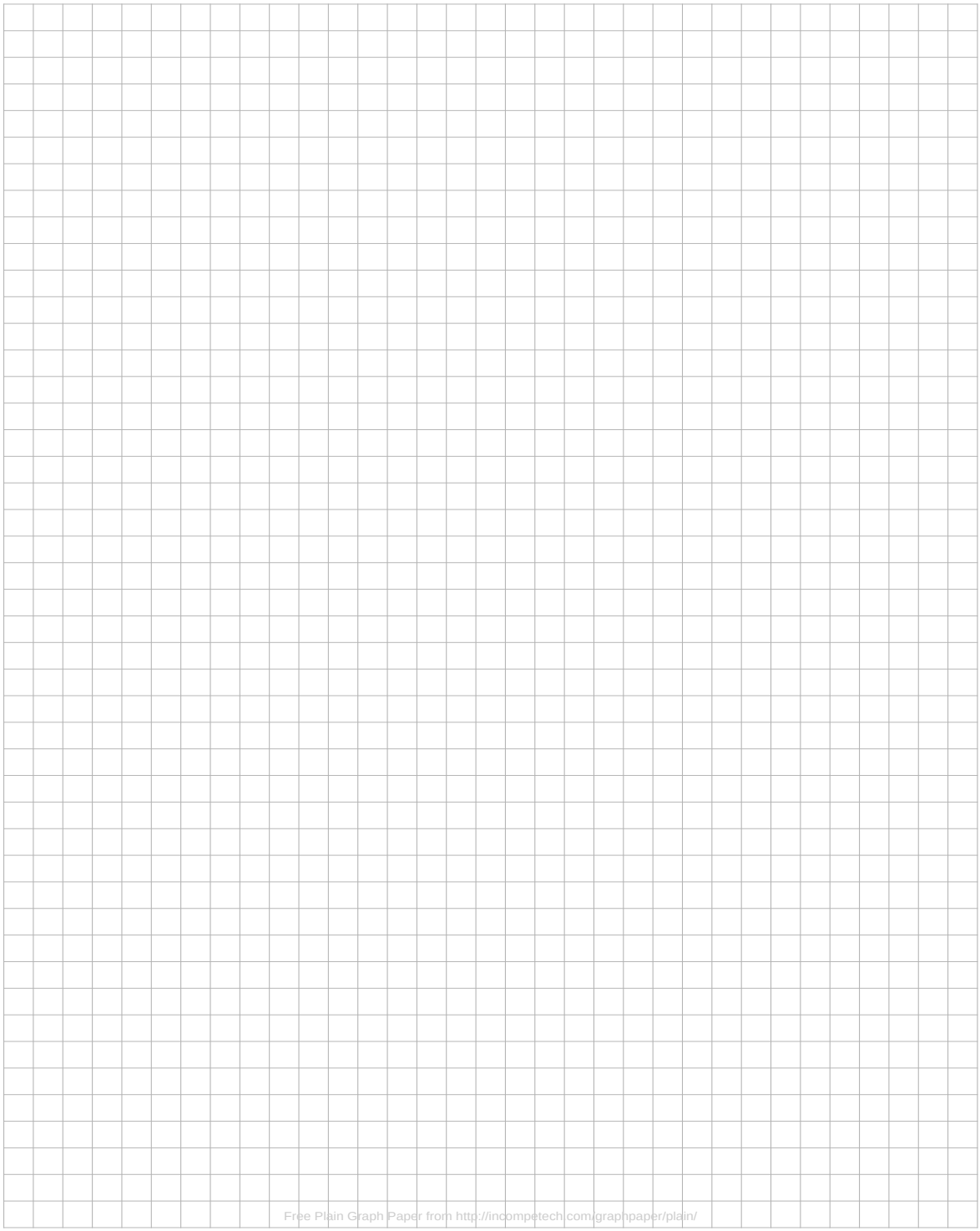
Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway
Thiago.Guerrera@math.ntnu.no

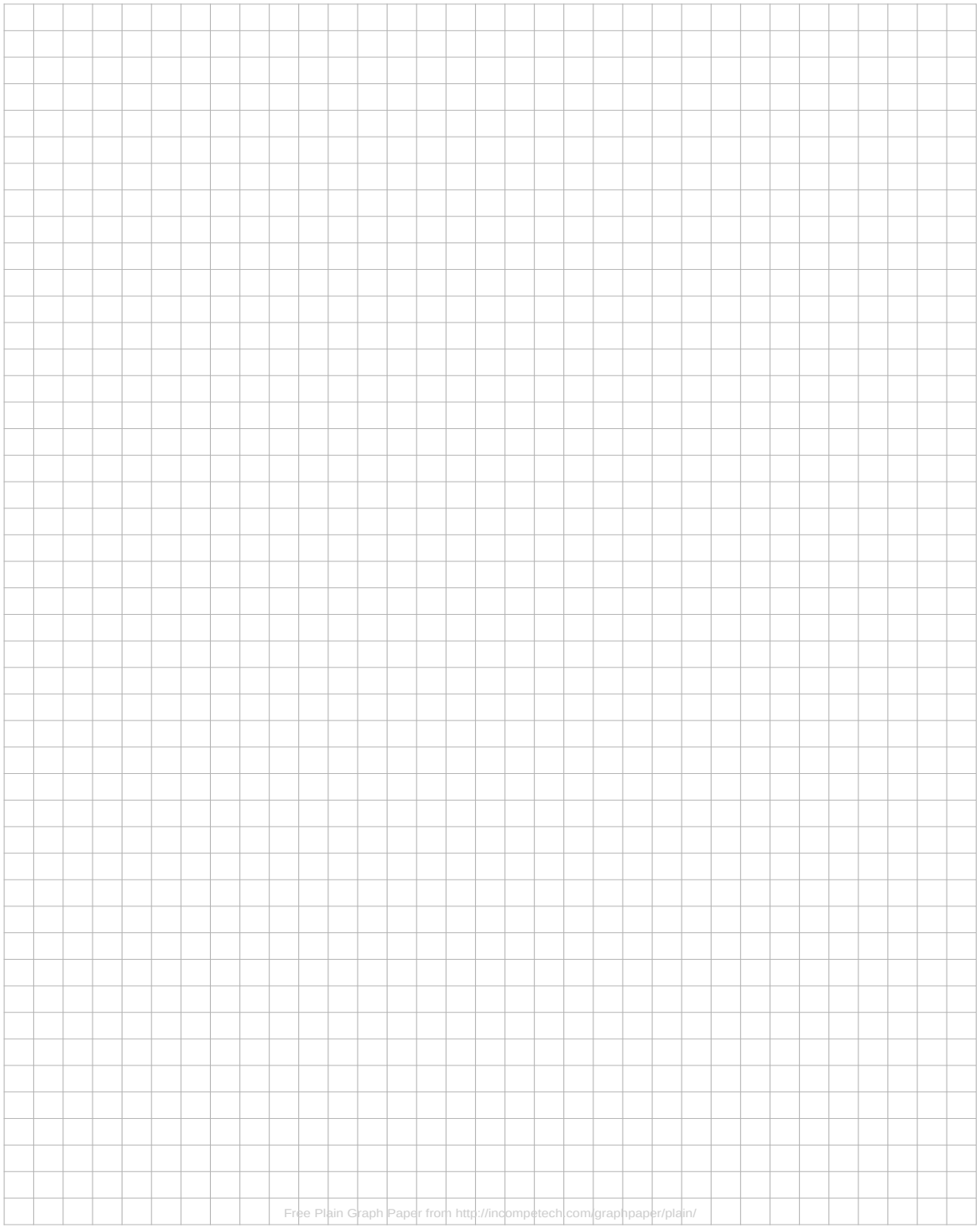
In this tutorial, I will discuss approximate Bayesian inference for a class of models named latent Gaussian models (LGM). LGM's are perhaps the most commonly used class of models in statistical applications. It includes, among others, most of (generalized) linear models, (generalized) additive models, smoothing spline models, state space models, semiparametric regression, spatial and spatiotemporal models, log-Gaussian Cox processes and geostatistical and geoadditive models. The concept of LGM is intended for the modeling stage, but turns out to be extremely usefull when doing inference as we can treat models listed above in a unified way and using the same algorithm and software tool. Our approach to (approximate) Bayesian inference, is to use integrated nested Laplace approximations (INLA). Using this tool, we can directly compute very accurate approximations to the posterior marginals. The main benefit of these approximations is computational: where Markov chain Monte Carlo algorithms need hours or days to run, our approximations provide more precise estimates in seconds or minutes. Another advantage with our approach is its generality, which makes it possible to perform Bayesian analysis in an automatic, streamlined way, and to compute model comparison criteria and various predictive measures so that models can be compared and the model under study can be challenged.
I will introduce the class of latent Gaussian models, describe the "big picture" of the INLA algorithm and introduce the r-INLA package. This is intended to be an applied tutorial, so I will focus on the use of the package on a variety of examples rather than the theory and implementation details behind INLA.

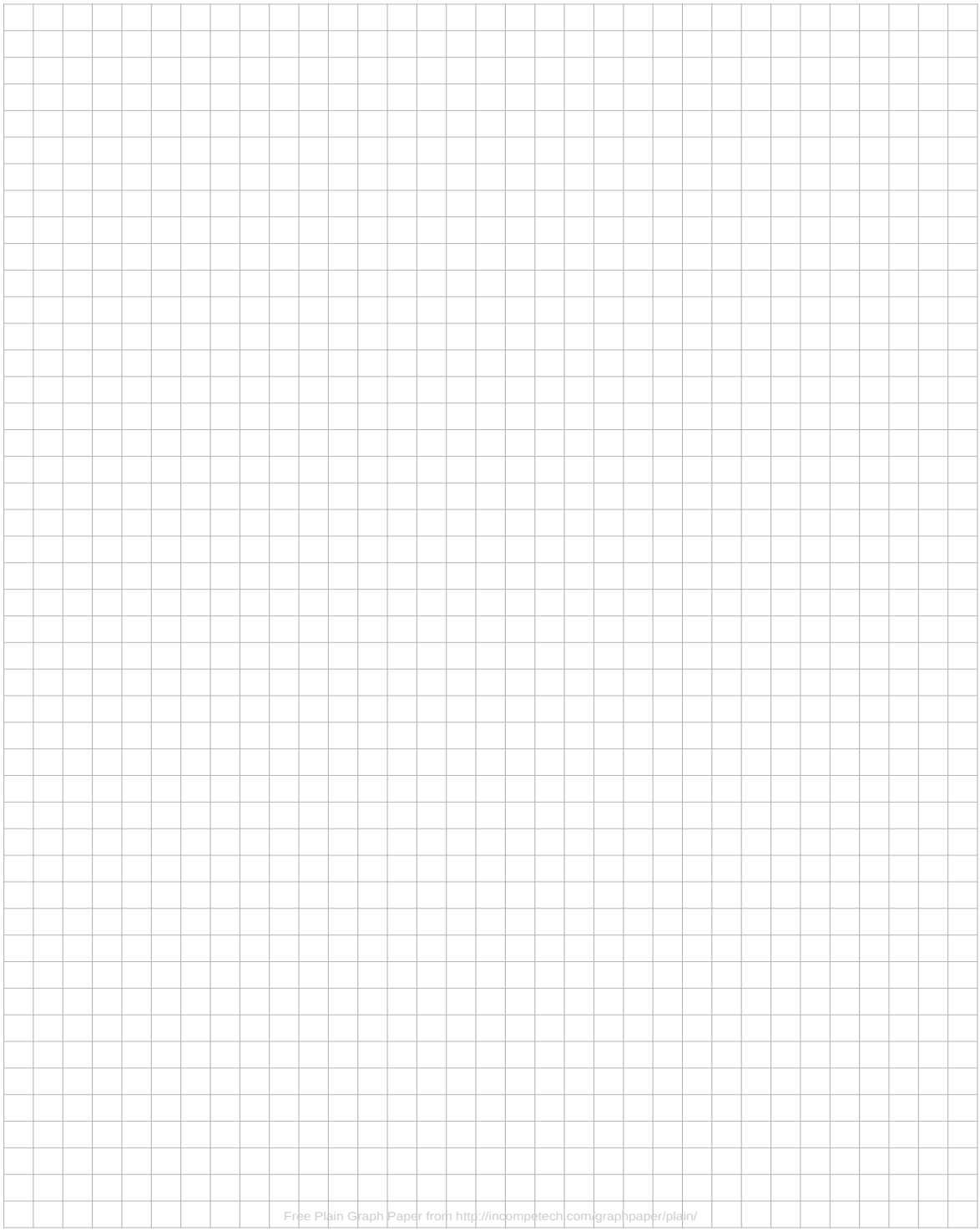
INDEX OF AUTHORS

Index of Authors


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/

SUPPORTED BY

RESULT