

20th European Young Statisticians Meeting

Report of Contributions

Contribution ID: 1

Type: **Speaker**

Matrix Independent Component Analysis

Monday, 14 August 2017 15:30 (30 minutes)

Independent component analysis (ICA) is a popular means of dimension reduction for vector-valued random variables. In this short note we review its extension to arbitrary tensor-valued random variables by considering the special case of two dimensions where the tensors are simply matrices.

Primary author: Mr VIRTÄ, Joni (University of Turku)

Presenter: Mr VIRTÄ, Joni (University of Turku)

Contribution ID: 2

Type: **Speaker**

Nonparametric estimation of gradual change points in the jump behaviour of an Ito semimartingale

Monday, 14 August 2017 14:30 (30 minutes)

In applications the properties of a stochastic feature often change gradually rather than abruptly, that is: after a constant phase for some time they slowly start to vary. The goal of this talk is to introduce an estimator for the location of a gradual change point in the jump characteristic of a discretely observed Ito semimartingale. To this end we propose a measure of time variation for the jump behaviour of the process and consistency of the desired estimator is a consequence of weak convergence of a suitable empirical process in some function space. Finally, we discuss simulation results which verify that the new estimator has advantages compared to the classical argmax-estimator.

Primary author: Mr HOFFMANN, Michael (Ruhr-Universität Bochum)

Co-authors: Prof. DETTE, Holger (Ruhr-Universität Bochum); Prof. VETTER, Mathias (Christian-Albrechts-Universität zu Kiel)

Presenter: Mr HOFFMANN, Michael (Ruhr-Universität Bochum)

Contribution ID: 3

Type: **Speaker**

Can humans be replaced by computers in taxa recognition?

Monday, 14 August 2017 14:00 (30 minutes)

Biomonitoring of waterbodies is vital as the number of anthropogenic stressors on aquatic ecosystems keeps growing. However, the continuous decrease in funding makes it impossible to meet monitoring goals or sustain traditional manual sample processing. We review what kind of statistical tools can be used to enhance the cost efficiency of biomonitoring: We explore automated identification of freshwater macroinvertebrates which are used as one indicator group in biomonitoring of aquatic ecosystems. We present the first classification results of a new imaging system producing multiple images per specimen. Moreover, these results are compared with the results of human experts. On a data set of 29 taxonomical groups, automated classification produces a higher average accuracy than human experts.

Primary author: Dr ÄRJE, Johanna (University of Jyväskylä, Department of Mathematics and Statistics)

Co-authors: Dr RAITOHARJU, Jenni (Department of Signal Processing, Tampere University of Technology); Dr MEISSNER, Kristian (Finnish Environment Institute); Dr KÄRKKÄINEN, Salme (Department of Mathematics and Statistics, University of Jyväskylä); Dr TIRRONEN, Ville (Department of Mathematical Information Technology, University of Jyväskylä)

Presenter: Dr ÄRJE, Johanna (University of Jyväskylä, Department of Mathematics and Statistics)

Contribution ID: 5

Type: **Speaker**

AIC post-selection inference in linear regression

Monday, 14 August 2017 16:00 (30 minutes)

Post-selection inference has been considered a crucial topic in data analysis. In this article, we develop a new method to obtain correct inference after model selection by the Akaike's information criterion Akaike (1973) in linear regression models. Confidence intervals can be calculated by incorporating the randomness of the model selection in the distribution of the parameter estimators which act as pivotal quantities. Simulation results show the accuracy of the proposed method.

Primary author: Mr CHARKHI, Ali (KULeuven)

Co-author: Ms CLAESKENS, Gerda (KULeuven)

Presenter: Mr CHARKHI, Ali (KULeuven)

Contribution ID: 6

Type: **Speaker**

The Elicitation Problem

Tuesday, 15 August 2017 11:00 (30 minutes)

Competing point forecasts for functionals such as the mean, a quantile, or a certain risk measure are commonly compared in terms of loss functions. These should be incentive compatible, i.e., the expected score should be minimized by the correctly specified functional of interest. A functional is called *elicitable* if it possesses such an incentive compatible loss function. With the squared loss and the absolute loss, the mean and the median possess such incentive compatible loss functions, which means they are elicitable. In contrast, variance or Expected Shortfall are not elicitable. Besides investigating the elicibility of a functional, it is important to determine the whole class of incentive compatible loss functions as well as to give recommendations which loss function to use in practice, taking into regard secondary quality criteria of loss functions such as order-sensitivity, convexity, or homogeneity.

Primary author: Dr FISSLER, Tobias (University of Bern)**Presenter:** Dr FISSLER, Tobias (University of Bern)

Contribution ID: 7

Type: **Speaker**

Predict extreme influenza epidemics

Wednesday, 16 August 2017 09:30 (30 minutes)

Influenza viruses are responsible for annual epidemics, causing more than 500,000 deaths per year worldwide. A crucial question for resource planning in public health is to predict the morbidity burden of extreme epidemics. We say that an epidemic is extreme whenever the influenza incidence rate exceeds a high threshold for at least one week. Our objective is to predict whether an extreme epidemic will occur in the near future, say the next couple of weeks.

The weekly numbers of influenza-like illness (ILI) incidence rates in France are available from the Sentinel network for the period 1991-2017. ILI incidence rates exhibit two different regimes, an epidemic regime during winter and a non-epidemic regime during the rest of the year. To identify epidemic periods, we use a two-state autoregressive hidden Markov model.

A main goal of Extreme Value Theory is to assess, from a series of observations, the probability of events that are more extreme than those previously recorded. Because of the autoregressive structure of the data, we choose to fit one of the multivariate generalized Pareto distribution models proposed in Rootzén et al. (2016a) [Multivariate peaks over threshold models. arXiv:1603.06619v2]; see also Rootzén et al. (2016b) [Peaks over thresholds modeling with multivariate generalized Pareto distributions. arXiv:1612.01773v1]. For these models, explicit densities are given, and formulas for conditional probabilities can then be deduced, from which we can predict if an epidemic will be extreme, given the first weeks of observation.

Primary author: THOMAS, Maud (Université Pierre et Marie Curie)

Co-author: Prof. ROOTZÉN, Holger (Chalmers University of Technology)

Presenter: THOMAS, Maud (Université Pierre et Marie Curie)

Contribution ID: 9

Type: **Speaker**

Controlled branching processes in Biology: a model for cell proliferation

Tuesday, 15 August 2017 14:00 (30 minutes)

Branching processes are relevant models in the development of theoretical approaches to problems in applied fields such as, for instance, growth and extinction of populations, biology, epidemiology, cell proliferation kinetics, genetics and algorithm and data structures. The most basic model, the so-called Bienaymé-Galton-Watson process, consists of individuals that reproduce independently of the others following the same probability distribution, known as offspring distribution. A natural generalization is to incorporate a random control function which determines the number of progenitors in each generation. The resulting process is called controlled branching process.

In this talk, we deal with a problem arising in cell biology. More specifically, we focus our attention on experimental data generated by time-lapse video recording of cultured in vitro oligodendrocyte cells. In A.Y. Yakovlev et al. (2008) (Branching Processes as Models of Progenitor Cell Populations and Estimation of the Offspring Distributions, *Journal of the American Statistical Association*, 103(484):1357–1366), a two-type age dependent branching process with emigration is considered to describe the kinetics of cell populations. The two types of cells considered are referred as type T_1 (immediate precursors of oligodendrocytes) and type T_2 (terminally differentiated oligodendrocytes). The reproduction process of these cells is as follows: when stimulating to divide under in vitro conditions, the progenitor cells are capable of producing either their direct progeny (two daughter cells of the same type) or a single, terminally differentiated nondividing oligodendrocyte. Moreover, censoring effects as a consequence of the migration of progenitor cells out of the microscopic field of observation are modelled as the process of emigration of the type T_1 cells.

In this work, we propose a two-type controlled branching process to describe the embedded discrete branching structure of the age-dependent branching process aforementioned. We address the estimation of the offspring distribution of the cell population in a Bayesian outlook by making use of disparities. The importance of this problem yields in the fact that the behaviour of these populations is strongly related to the main parameters of the offspring distribution and in practice, these values are unknown and their estimation is necessary. The proposed methodology introduced in M. González et al. (2017) (Robust estimation in controlled branching processes: Bayesian estimators via disparities. *Work in progress*), is illustrated with an application to the real data set given in A.Y. Yakovlev et al. (2008).

Primary author: Ms MINUESA ABRIL, Carmen (University of Extremadura)

Co-authors: Dr DEL PUERTO GARCÍA, Inés María (University of Extremadura); Dr GONZÁLEZ VELASCO, Miguel (University of Extremadura)

Presenter: Ms MINUESA ABRIL, Carmen (University of Extremadura)

Contribution ID: 10

Type: **Speaker**

Multilevel Functional Principal Component Analysis for Unbalanced Data

Monday, 14 August 2017 16:30 (30 minutes)

Functional principal component analysis (FPCA) is the key technique for dimensionality reduction and detection of main directions of variability present in functional data. However, it is not the most suitable tool for the situation when analyzed dataset contains repeated or multiple observations, because information about repeatability of measurements is not taken into account. Multilevel functional principal component analysis (MFPCA) is the modified version of FPCA developed for data observed at multiple visits. The original MFPCA method was designed for balanced data only, where for each subject the same number of measurements is available. In this article we propose the modified MFPCA algorithm which can be applied for unbalanced functional data; that is, in the situation where a different number of observations can be present for every subject. The modified algorithm is validated and tested on real-world sleep data.

Primary author: ROŠŤÁKOVÁ, Zuzana (Institute of Measurement Science, Slovak Academy of Sciences)

Presenter: ROŠŤÁKOVÁ, Zuzana (Institute of Measurement Science, Slovak Academy of Sciences)

Contribution ID: 11

Type: **Speaker**

Mallows' Model Based on Lee Distance

Thursday, 17 August 2017 15:30 (30 minutes)

In this paper the Mallows' model based on Lee distance is considered and compared to models induced by other metrics on the permutation group. As an illustration, the complete rankings from the American Psychological Association election data are analyzed.

Primary authors: Prof. STOIMENOVA, Eugenia (Institute of Information and Communication Technologies and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 25A, 1113 Sofia, Bulgaria); Mr NIKOLOV, Nikolay (Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 8, 1113 Sofia, Bulgaria)

Presenter: Mr NIKOLOV, Nikolay (Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 8, 1113 Sofia, Bulgaria)

Contribution ID: 12

Type: **Speaker**

Parameter Estimation for Discretely Observed Infinite-Server Queues with Markov-Modulated Input

Tuesday, 15 August 2017 14:30 (30 minutes)

The Markov-modulated infinite-server queue is a queueing system with infinitely many servers, where the arrivals follow a Markov-modulated Poisson process (MMPP), i.e. a Poisson process with rate modulating between several values. The modulation is driven by an underlying and unobserved continuous time Markov chain $\{X_t\}_{t \geq 0}$. The inhomogeneous rate of the Poisson process, $\lambda(t)$, stochastically alternates between d different rates, $\lambda_1, \dots, \lambda_d$, in such a way that $\lambda(t) = \lambda_i$ if $X_t = i, i = 1, \dots, d$.

We are interested in estimating the parameters of the arrival process for this queueing system based on observations of the queue length at discrete times only. We assume exponentially distributed service times with rate μ , where μ is time-independent and known. Estimation of the parameters of the arrival process has not yet been studied for this particular queueing system. Two types of missing data are intrinsic to the model, which complicates the estimation problem. First, the underlying continuous time Markov chain in the Markov-modulated arrival process is not observed. Second, the queue length is only observed at a finite number of discrete time points. As a result, it is not possible to distinguish the number of arrivals and the number of departures between two consecutive observations.

In this talk we show how we derive an explicit algorithm to find maximum likelihood estimates of the parameters of the arrival process, making use of the EM algorithm. Our approach extends the one used in Okamura et al. (2009), where the parameters of an MMPP are estimated based on observations of the process at discrete times. However, in contrast to our setting, Okamura et al. (2009) do not consider departures and therefore do not deal with the second type of missing data. We illustrate the accuracy of the proposed estimation algorithm with a simulation study.

Reference: Okamura H., Dohi T., Trivedi K.S.
(2009).

Markovian Arrival Process Parameter Estimation With Group Data.
IEEE/ACM Transactions on Networking.
Vol. 17, No. 4, pp. 1326–1339

Primary authors: Dr KNAPIK, Bartek (Vrije Universiteit Amsterdam); Ms SOLLIE, Birgit (Vrije Universiteit Amsterdam); Prof. DE GUNST, Mathisca (Vrije Universiteit Amsterdam); Prof. MANDJES, Michel (Universiteit van Amsterdam)

Presenter: Ms SOLLIE, Birgit (Vrije Universiteit Amsterdam)

Contribution ID: 13

Type: **Speaker**

Some recent characterization based goodness of fit tests

Tuesday, 15 August 2017 13:00 (30 minutes)

In this paper some recent advances in goodness of fit testing are presented. Special attention is given to goodness of fit tests based on equidistribution and independence characterizations. New concepts are described through some modern exponentiality tests. Their natural generalizations are also proposed. All tests are compared in Bahadur sense.

Primary author: Dr MILOŠEVIĆ, Bojana (Faculty of Mathematics)

Presenter: Dr MILOŠEVIĆ, Bojana (Faculty of Mathematics)

Contribution ID: 14

Type: **Speaker**

Confidence regions in Cox proportional hazards model with measurement errors

Thursday, 17 August 2017 16:00 (30 minutes)

Cox proportional hazards model with measurement errors in covariates is considered. It is the ubiquitous technique in biomedical data analysis. In Kukush et al. (2011) [Journal of Statistical Research **45**, 77-94] and Chimisov and Kukush (2014) [Modern Stochastics: Theory and Applications **1**, 13-32] asymptotic properties of a simultaneous estimator $(\lambda_n; \beta_n)$ for the baseline hazard rate $\lambda(\cdot)$ and the regression parameter β were studied, at that the parameter set $\Theta = \Theta_\lambda \times \Theta_\beta$ was assumed bounded.

In Kukush and Chernova (2017) [Theory of Probability and Mathematical Statistics **96**, 100-109] we dealt with the simultaneous estimator $(\lambda_n; \beta_n)$ in the case, where the Θ_λ was unbounded from above and not separated away from 0. The estimator was constructed in two steps: first we derived a strongly consistent estimator and then modified it to provide its asymptotic normality.

In this talk, we construct the confidence interval for an integral functional of $\lambda(\cdot)$ and the confidence region for β . We reach our goal in each of the three cases: (a) the measurement error is bounded, (b) it is normally distributed, or (c) it is a shifted Poisson random variable. The censor is assumed to have a continuous pdf. In future research we intend to elaborate a method for heavy tailed error distributions.

Primary author: Ms CHERNOVA, Oksana (Taras Shevchenko National University of Kyiv)

Presenter: Ms CHERNOVA, Oksana (Taras Shevchenko National University of Kyiv)

Contribution ID: 15

Type: **Speaker**

Joint Bayesian nonparametric reconstruction of dynamical equations

Thursday, 17 August 2017 09:30 (30 minutes)

We propose a Bayesian nonparametric mixture model for the joint full reconstruction of m dynamical equations,

given m observed dynamically-noisy-corrupted chaotic time series. The method of reconstruction is based on the Pairwise Dependent Geometric Stick Breaking Processes mixture priors (PDGSBP) first proposed by Hatjispyros et al. (2017). We assume that

each set of dynamical equations has a deterministic part with a known functional form i.e.

$x_{ji} = g_j(\vartheta_j, x_{j,i-1}, \dots, x_{j,i-l_j}) + \epsilon_{x_{ji}}, \quad 1 \leq j \leq m, \quad 1 \leq i \leq n_j.$ under the assumption that the noise processes $(\epsilon_{x_{ji}})$

are independent and identically distributed for all j and i from some unknown zero mean process $f_j(\cdot)$. Additionally, we assume that a-priori we have the knowledge that the processes $(\epsilon_{x_{ji}})$ for

$j = 1, \dots, m$ have common characteristics, e.g. they may have common variances or even have similar tail behavior etc. For a full reconstruction, we would like to jointly estimate the following quantities $(\vartheta_j) \in \Theta \subseteq \text{cal} R^{k_j}, \quad (x_{j,0}, \dots, x_{j,l_j-1}) \in \text{cal} X_j \subseteq \text{cal} R^{l_j},$ and perform density estimation to the noise processes.

Our contention is that whenever there is at least one sufficiently large data set, using carefully selected informative borrowing-of-strength-prior-specifications we are able to reconstruct those dynamical processes that are responsible for the generation of time series with small sample sizes; namely sample sizes that are inadequate for an independent reconstruction. We illustrate the joint estimation process for the case $m = 2$, when the two time series are coming from a quadratic and a cubic

stochastic process of lag one and the noise processes are zero mean normal mixtures with common components.

Primary authors: Mr MERKATAS, Christos (Department of Mathematics, University of the Aegean, Greece); Prof. HATJISPYROS, Spyridon (Department of Mathematics, University of the Aegean, Greece)

Presenter: Mr MERKATAS, Christos (Department of Mathematics, University of the Aegean, Greece)

Contribution ID: 16

Type: **Speaker**

Viterbi process for pairwise Markov models

Thursday, 17 August 2017 10:00 (30 minutes)

My talk is based on ongoing joint work with my supervisor Jüri Lember.

We consider a Markov chain $Z = \{Z_k\}_{k \geq 1}$ with product state space $\mathcal{X} \times \mathcal{Y}$, where \mathcal{Y} is a finite set (state space) and \mathcal{X} is an arbitrary separable metric space (observation space). Thus, the process Z decomposes as $Z = (X, Y)$, where $X = \{X_k\}_{k \geq 1}$ and $Y = \{Y_k\}_{k \geq 1}$ are random processes taking values in \mathcal{X} and \mathcal{Y} , respectively. Following cite{pairwise,pairwise2,pairwise3}, we call the process Z a \textit{pairwise Markov model}. The process X is identified as an observation process and the process Y , sometimes called the \textit{regime}, models the observations-driving hidden state sequence.

Therefore our general model contains many well-known stochastic models as a special case: hidden Markov models, Markov switching models, hidden Markov models with dependent noise and many more. The \textit{segmentation} or \textit{path estimation} problem consists of estimating the realization of (Y_1, \dots, Y_n) given a realization $x_{1:n}$ of (X_1, \dots, X_n) . A standard estimate is any path $v_{1:n} \in \mathcal{Y}^n$ having maximum posterior probability:

$\$v_{1:n} = \operatorname{argmax}_{y_{1:n}} P(Y_{1:n} = y_{1:n} | X_{1:n} = x_{1:n}). Any such path is called Viterbi path and we are interested in the behaviour$

We show that under some conditions the infinite Viterbi path indeed exists for almost every realization $x_{1:\infty}$ of X , thereby defining an infinite Viterbi decoding of X , called the *Viterbi process*. This is done through construction of *barriers*. A barrier is a fixed-sized block in the observations $x_{1:n}$ that fixes the Viterbi path up to itself: for every continuation of $x_{1:n}$, the Viterbi path up to the barrier remains unchanged. Therefore, if almost every realization of X -process contains infinitely many barriers, then the Viterbi process exists.

Having infinitely many barriers is not necessary for existence of infinite Viterbi path, but the barrier-construction has several advantages. One of them is that it allows to construct the infinite path *piecewise*, meaning that to determine the first k elements $v_{1:k}$ of the infinite path it suffices to observe $x_{1:n}$ for n big enough. Barrier construction has another great advantage: namely, the process $(Z, V) = \{(Z_k, V_k)\}_{k \geq 1}$, where $V = \{V_k\}_{k \geq 1}$ denotes the Viterbi process, is under certain conditions regenerative. This can be proven by, roughly speaking, applying the Markov splitting method to construct regeneration times for Z which coincide with the occurrences of barriers. Regenerativity of (Z, V) allows to easily prove limit theorems on paths. In fact, in a special case of hidden Markov model this regenerative property has already been known to hold.

Primary author: Mr SOVA, Joonas (University of Tartu)

Presenter: Mr SOVA, Joonas (University of Tartu)

Contribution ID: 17

Type: **Speaker**

E-optimal approximate block designs for treatment-control comparisons

Thursday, 17 August 2017 14:00 (30 minutes)

We study E -optimal block designs for comparing a set of test treatments with a control treatment. We provide the complete class of all E -optimal approximate block designs and we show that these designs are characterized by simple linear constraints. Employing the provided characterization, we obtain a class of E -optimal exact block designs with unequal block sizes for comparing test treatments with a control.

Primary author: Mr ROSA, Samuel (Comenius University in Bratislava)

Presenter: Mr ROSA, Samuel (Comenius University in Bratislava)

Contribution ID: 18

Type: **Speaker**

Information criteria for structured sparse variable selection

Thursday, 17 August 2017 14:30 (30 minutes)

In contrast to the low dimensional case, variable selection under the assumption of sparsity in high dimensional models is strongly influenced by the effects of false positives.

The effects of false positives are tempered by combining the variable selection with a shrinkage estimator, such as in the lasso, where the selection is realized by minimizing the sum of squared residuals regularized by an ℓ_1 norm of the selected variables. Optimal variable selection is then equivalent to finding the best balance between closeness of fit and regularity, i.e., to optimization of the regularization parameter with respect to an information criterion such as Mallows's Cp or AIC. For use in this optimization procedure, the lasso regularization is found to be too tolerant towards false positives, leading to a considerable overestimation of the model size. Using an ℓ_0 regularization instead requires careful consideration of the false positives, as they have a major impact on the optimal regularization parameter. As the framework of the classical linear model has been analysed in previous work, the current paper concentrates on structured models and, more specifically, on grouped variables. Although the imposed structure in the selected models can be understood to somehow reduce the effect of false positives, we observe a qualitatively similar behavior as in the unstructured linear model.

Primary author: Mr MARQUIS, Bastien (Université Libre de Bruxelles)

Co-author: Mr JANSEN, Maarten (Université Libre de Bruxelles)

Presenter: Mr MARQUIS, Bastien (Université Libre de Bruxelles)

Contribution ID: 19

Type: **Speaker**

Simulating and Forecasting Human Population with General Branching Process

Friday, 18 August 2017 11:30 (30 minutes)

The branching process theory is widely used to describe a population dynamics in which particles live and produce other particles through their life, according to given stochastic birth and death laws. The theory of General Branching Processes (GBP) presents a continuous time model in which every woman has random life length and gives birth to children in random intervals of time. The flexibility of the GBP makes it very useful for modelling and forecasting human population. This paper is a continuation of previous developments in the theory, necessary to model the specifics of human population, and presents their application in forecasting the population age structure of Bulgaria. It also introduces confidence intervals of the forecasts, calculated by GBP simulations, which reflect both the stochastic nature of the birth and death laws and the branching process itself. The simulations are also used to determine the main sources of risk to the forecast.

Primary author: Dr TRAYANOV, Plamen (Sofia Univeristy "St. Kliment Ohridski")

Presenter: Dr TRAYANOV, Plamen (Sofia Univeristy "St. Kliment Ohridski")

Contribution ID: 20

Type: **Speaker**

Theoretical and simulation results on heavy-tailed fractional Pearson diffusions

Friday, 18 August 2017 10:30 (30 minutes)

We define heavy-tailed fractional reciprocal gamma and Fisher-Snedecor diffusions by a non-Markovian time change in the corresponding Pearson diffusions. We illustrate known theoretical results regarding these fractional diffusions via simulations.

Primary author: Mr PAPIĆ, Ivan (Department of Mathematics, J.J. Strossmayer University of Osijek)

Co-authors: Prof. SIKORSKII, Alla (Department of Statistics and Probability, Michigan State University); Prof. ŠUVAK, Nenad (Department of Mathematics, J.J. Strossmayer University of Osijek); Prof. LEONENKO, Nikolai N. (School of Mathematics, Cardiff University)

Presenter: Mr PAPIĆ, Ivan (Department of Mathematics, J.J. Strossmayer University of Osijek)

Contribution ID: 21

Type: **Speaker**

Copula based BINAR models with applications

Friday, 18 August 2017 11:00 (30 minutes)

In this paper we study the problem of modelling the integer-valued vector observations. We consider the BINAR(1) models defined via copula-joint innovations. We review different parameter estimation methods and analyse estimation methods of the copula dependence parameter. We also examine the case where seasonality is present in integer-valued data and suggest a method of deseasonalizing them. Finally, an empirical application is carried out.

Primary author: BUTEIKIS, Andrius (Faculty of Mathematics and Informatics, Vilnius University)

Presenter: BUTEIKIS, Andrius (Faculty of Mathematics and Informatics, Vilnius University)

Contribution ID: 22

Type: **Speaker**

Fréchet means and Procrustes analysis in Wasserstein space

Tuesday, 15 August 2017 15:30 (30 minutes)

We consider three interlinked problems in stochastic geometry: (1) constructing optimal multicouplings of random vectors; (2) determining the Fréchet mean of probability measures in Wasserstein space; and (3) registering collections of randomly deformed spatial point processes. We demonstrate how these problems are canonically interpreted through the prism of the theory of optimal transportation of measure on \mathbb{R}^d . We provide explicit solutions in the one dimensional case, consistently solve the registration problem and establish convergence rates and a (tangent space) central limit theorem for Cox processes. When $d > 1$, the solutions are no longer explicit and we propose a steepest descent algorithm for deducing the Fréchet mean in problem (2). Supplemented by uniform convergence results for the optimal maps, this furnishes a solution to the multicoupling problem (1). The latter is then utilised, as in the case $d = 1$, in order to construct consistent estimators for the registration problem (3). While the consistency results parallel their one-dimensional counterparts, their derivation requires more sophisticated techniques from convex analysis. This is joint work with Victor M. Panaretos

Primary author: Dr ZEMEL, Yoav (Ecole polytechnique fédérale de Lausanne)

Co-author: Prof. PANARETOS, Victor (Ecole polytechnique fédérale de Lausanne)

Presenter: Dr ZEMEL, Yoav (Ecole polytechnique fédérale de Lausanne)

Contribution ID: 23

Type: **Speaker**

Efficient estimation for diffusions

Wednesday, 16 August 2017 14:00 (30 minutes)

This talk concerns estimation of the diffusion parameter of a diffusion process observed over a fixed time interval. We present conditions on approximate martingale estimating functions under which estimators are consistent, rate optimal, and efficient under high frequency (in-fill) asymptotics. Here, limit distributions of the estimators are non-standard in the sense that they are generally normal variance-mixture distributions. In particular, the mixing distribution depends on the full sample path of the diffusion process over the observation time interval. Making use of stable convergence in distribution, we also present the more easily applicable result that estimators normalized by a suitable data-dependent transformation converge in distribution to a standard normal distribution. The theory is illustrated by a simulation study.

The work presented in this talk is published in:

Jakobsen, N. M. and Sørensen, M. (2017). *Efficient estimation for diffusions sampled at high frequency over a fixed time interval*. Bernoulli, 23(3):1874-1910.

Primary author: JAKOBSEN, Nina Munkholt (University of Copenhagen)

Co-author: Prof. SØRENSEN, Michael (University of Copenhagen)

Presenter: JAKOBSEN, Nina Munkholt (University of Copenhagen)

Contribution ID: 24

Type: **Speaker**

Estimates for distributions of Hölder semi-norms of random processes from spaces $F_\psi(\Omega)$

Wednesday, 16 August 2017 14:30 (30 minutes)

In the following we deal with estimates for distributions of Hölder semi-norms of sample functions of random processes from spaces $F_\psi(\Omega)$, defined on a compact metric space and on an infinite interval $[0, \infty)$, i.e. probabilities

$$P \left\{ \sup_{0 < \rho(t,s) \leq \varepsilon, t,s \in \mathbb{T}} \frac{|X(t) - X(s)|}{f(\rho(t,s))} > x \right\}.$$

Such estimates and assumptions under which semi-norms of sample functions of random processes from spaces $F_\psi(\Omega)$, defined on a compact space, satisfy the Hölder condition were obtained by Kozachenko and Zatul (2015). Similar results were provided for Gaussian processes, defined on a compact space, by Dudley (1973). Kozachenko (1985) generalized Dudley's results for random processes belonging to Orlicz spaces, see also Buldygin and Kozachenko (2000). Marcus and Rosen (2008) obtained L^p moduli of continuity for a wide class of continuous Gaussian processes. Kozachenko et al. (2011) studied the Lipschitz continuity of generalized sub-Gaussian processes and provided estimates for the distribution of Lipschitz norms of such processes. But all these problems were not considered yet for processes, defined on an infinite interval.

Primary author: Mr ZATULA, Dmytro (Taras Shevchenko National University of Kyiv)

Presenter: Mr ZATULA, Dmytro (Taras Shevchenko National University of Kyiv)

Contribution ID: 25

Type: **Speaker**

Modeling of vertical and horizontal variation in multivariate functional data

Tuesday, 15 August 2017 16:00 (30 minutes)

We present a model for multivariate functional data that simultaneously model vertical and horizontal variation.

Horizontal variation is modeled using warping functions represented by latent gaussian variables. Vertical variation is modeled using Gaussian processes using a generally applicable low-parametric covariance structure.

We devise a method for maximum likelihood estimation using a Laplace approximation and apply it to three different data sets.

Primary author: Mr OLSEN, Niels (Københvans Universitet)

Presenter: Mr OLSEN, Niels (Københvans Universitet)

Contribution ID: 26

Type: **Speaker**

Finite Mixture of C-vines for Complex Dependence

Thursday, 17 August 2017 09:00 (30 minutes)

Recently, there has been an increasing interest on the combination of copulas with a finite mixture model. Such a framework is useful to reveal the hidden dependence patterns observed for random variables flexibly in terms of statistical modeling. The combination of vine copulas incorporated into a finite mixture model is also beneficial for capturing hidden structures on a multivariate data set. In this respect, the main goal of this study is extending the study of Kim et al. (2013) with different scenarios. For this reason, finite mixture of C-vine is proposed for multivariate data with different dependence structures. The performance of the proposed model has been tested by different simulated data set including various tail dependence properties.

Primary author: EVKAYA, O. Ozan (Atılım University)

Co-authors: KESTEL, A. Sevtap (Middle East Technical University); YOZGATLIGIL, Ceylan (Middle East Technical University)

Presenter: EVKAYA, O. Ozan (Atılım University)

Contribution ID: 27

Type: **Speaker**

Testing independence for multivariate time series by the auto-distance correlation matrix

Tuesday, 15 August 2017 11:30 (30 minutes)

We introduce the notions of multivariate auto-distance covariance and correlation functions for time series analysis. These concepts have been recently discussed in the context of both independent and dependent data but we extend them in a different direction by putting forward their matrix version. Their matrix version allows us to identify possible interrelationships among the components of a multivariate time series. Interpretation and consistent estimators of these new concepts are discussed. Additionally, we develop a test for testing the i.i.d. hypothesis for multivariate time series data. The resulting test statistic performs better than the standard multivariate Ljung-Box test statistic. All the above methodology is included in the R package dCovTS which is briefly introduced in this talk.

Primary author: Dr PITSILLOU, Maria (Department of Mathematics & Statistics, Cyprus)

Co-author: Prof. FOKIANOS, Konstantinos (Department of Mathematics & Statistics, University of Cyprus)

Presenter: Dr PITSILLOU, Maria (Department of Mathematics & Statistics, Cyprus)

Contribution ID: 28

Type: **Speaker**

Best Unbiased Estimators for Doubly Multivariate Data

Tuesday, 15 August 2017 16:30 (30 minutes)

The article addresses the best unbiased estimators of the block compound symmetric covariance structure for m -variate observations with equal mean vector over each level of factor or each time point (model with structured mean vector). Under multivariate normality, the free-coordinate approach is used to obtain unbiased linear and quadratic estimates for the model parameters. Optimality of these estimators follows from sufficiency and completeness of their distributions. Additionally, strong consistency is proven. The properties of the estimators in the proposed model are compared with the ones in the model with unstructured mean vector (the mean vector changes over levels of factor or time points).

Primary authors: Mr KOZIOŁ, Arkadiusz (Faculty of Mathematics, Computer Science and Econometrics University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland); Dr FONSECA, Miguel (Centro de Matemática e Aplicações Universidade Nova de Lisboa Monte da Caparica, 2829-516 Caparica, Portugal); LEIVA, Ricardo (Departamento de Matemática F.C.E., Universidad Nacional de Cuyo, 5500 Mendoza, Argentina); Prof. ZMYŚLONY, Roman (Faculty of Mathematics, Computer Science and Econometrics University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland)

Co-author: Prof. ROY, Anuradha (Department of Management Science and Statistics The University of Texas at San Antonio San Antonio, TX 78249, USA)

Presenter: Mr KOZIOŁ, Arkadiusz (Faculty of Mathematics, Computer Science and Econometrics University of Zielona Góra, Szafrana 4a, 65-516 Zielona Góra, Poland)

Contribution ID: 29

Type: **Speaker**

Stability of the Spectral EnKF under nested covariance estimators

Thursday, 17 August 2017 16:30 (30 minutes)

In the case of traditional Ensemble Kalman Filter (EnKF), it is known that the filter error does not grow faster than exponentially for a fixed ensemble size. The question posted in this contribution is whether the upper bound for the filter error can be improved by using an improved covariance estimator that comes from the right parameter subspace and has smaller asymptotic variance. Its effect on Spectral EnKF is explored by a simulation.

Primary author: TURČICOVÁ, Marie (Charles University, Prague)

Co-authors: Prof. MANDEL, Jan (University of Colorado Denver); Dr EBEN, Krystof (Institute of Computer Science, The Czech Academy of Sciences)

Presenter: TURČICOVÁ, Marie (Charles University, Prague)

Contribution ID: 30

Type: **Speaker**

Methods for bandwidth detection in kernel conditional density estimations

Tuesday, 15 August 2017 13:30 (30 minutes)

This contribution is focused on the kernel conditional density estimations (KCDE). The estimation depends on the smoothing parameters which influence the final density estimation significantly. This is the reason why a requirement of any data-driven method is needed for bandwidth estimation. In this contribution, the cross-validation method, the iterative method and the maximum likelihood approach are conducted for bandwidth selection of the estimator. An application on a real data set is included and the proposed methods are compared.

Primary author: Ms KONECNA, Katerina (Masaryk University)

Presenter: Ms KONECNA, Katerina (Masaryk University)

Contribution ID: 31

Type: **Speaker**

Inference on covariance matrices and operators using concentration inequalities

Wednesday, 16 August 2017 09:00 (30 minutes)

In the modern era of high and infinite dimensional data, classical statistical methodology is often rendered inefficient and ineffective when confronted with such big data problems as arise in genomics, medical imaging, speech analysis, and many other areas of research. Many problems manifest when the practitioner is required to take into account the covariance structure of the data during his or her analysis, which takes on the form of either a high dimensional low rank matrix or a finite dimensional representation of an infinite dimensional operator acting on some underlying function space. Thus, we propose using tools from the concentration of measure literature to construct rigorous descriptive and inferential statistical methodology for covariance matrices and operators. A variety of concentration inequalities are considered, which allow for the construction of nonasymptotic dimension-free confidence sets for the unknown matrices and operators. Given such confidence sets a wide range of estimation and inferential procedures can be and are subsequently developed.

Primary author: KASHLAK, Adam (Cambridge Centre for Analysis, University of Cambridge)

Presenter: KASHLAK, Adam (Cambridge Centre for Analysis, University of Cambridge)

Contribution ID: 32

Type: **Speaker**

Delete or Merge Regressors algorithm

Tuesday, 15 August 2017 10:30 (30 minutes)

This paper addresses a problem of linear and logistic model selection in the presence of both continuous and categorical predictors. In the literature two types of algorithms dealing with this problem can be found. The first one well known group lasso (\cite{group}) selects a subset of continuous and a subset of categorical predictors. Hence, it either deletes or not an entire factor. The second one is CAS-ANOVA (\cite{cas}) which selects a subset of continuous predictors and partitions of factors. Therefore, it merges levels within factors. Both these algorithms are based on the lasso regularization.

In the article a new algorithm called DMR (Delete or Merge Regressors) is described. Like CAS-ANOVA it selects a subset of continuous predictors and partitions of factors. However, instead of using regularization, it is based on a stepwise procedure, where in each step either one continuous variable is deleted or two levels of a factor are merged. The order of accepting consecutive hypotheses is based on sorting t-statistics or linear regression and likelihood ratio test statistics for logistic regression. The final model is chosen according to information criterion. Some of the preliminary results for DMR are described in \cite{pro}.

DMR algorithm works only for data sets where $p < n$ (number of columns in the model matrix is smaller than the number of observations). In the paper a modification of DMR called DMRnet is introduced that works also for data sets where $p \gg n$. DMRnet uses regularization in the screening step and DMR after decreasing the model matrix to $p < n$.

Theoretical results are proofs that DMR for linear and logistic regression are consistent model selection methods even when p tends to infinity with n . Furthermore, upper bounds on the error of selection are given.

Practical results are based on an analysis of real data sets and simulation setups. It is shown that DMRnet chooses smaller models with not higher prediction error than the competitive methods. Furthermore, in simulations it gives most often the highest rate of true model selection.

Primary author: Dr PROCHENKA, Agnieszka (Warsaw University)

Presenter: Dr PROCHENKA, Agnieszka (Warsaw University)

Contribution ID: 33

Type: **Invited Speaker**

Invited speaker - Sequential Monte Carlo: basic principles and algorithmic inference

Monday, 14 August 2017 11:00 (1 hour)

Sequential Monte Carlo methods form a class of genetic-type algorithms sampling, on-the-fly and in a very general context, sequences of probability measures. Today these methods constitute a standard device in the statistician's tool box and are successfully applied within a wide range of scientific and engineering disciplines. This talk is split into two parts, where the first provides an introduction to the SMC methodology and the second discusses some novel results concerning the stochastic stability and variance estimation in SMC.

Presenter: OLSSON, Jimmy (Royal Institute of Technology)

Contribution ID: 34

Type: **Invited Speaker**

Invited speaker - Sequential Monte Carlo: basic principles and algorithmic inference

Monday, 14 August 2017 13:00 (1 hour)

Sequential Monte Carlo methods form a class of genetic-type algorithms sampling, on-the-fly and in a very general context, sequences of probability measures. Today these methods constitute a standard device in the statistician's tool box and are successfully applied within a wide range of scientific and engineering disciplines. This talk is split into two parts, where the first provides an introduction to the SMC methodology and the second discusses some novel results concerning the stochastic stability and variance estimation in SMC.

Presenter: OLSSON, Jimmy (Royal Institute of Technology)

Contribution ID: 35

Type: **Invited Speaker**

Invited Speaker - Non-limiting spatial extremes

Tuesday, 15 August 2017 09:00 (1 hour)

Many questions concerning environmental risk can be phrased as spatial extreme value problems. Classical extreme value theory provides limiting models for maxima or threshold exceedances of a wide class of underlying spatial processes. These models can then be fitted to suitably defined extremes of spatial datasets and used, for example, to estimate the probability of events more extreme than we have observed to date. However, a major practical problem is that frequently the data do not appear to follow these limiting models at observable levels, and assuming otherwise leads to bias in estimation of rare event probabilities. To deal with this we require models that allow flexibility in both what the limit should be, and in the mode of convergence towards it. I will present a construction for such a model and discuss its application to some wave height data from the North Sea.

Presenter: WADSWORTH, Jenny (Lancaster University)

Contribution ID: 36

Type: **Invited Speaker**

Invited speaker - Formal languages for stochastic modelling

Wednesday, 16 August 2017 11:00 (1 hour)

Presenter: HILLSTON, Jane (University of Edinburgh)

Contribution ID: 37

Type: **Invited Speaker**

Invited Speaker - Embedding machine learning in stochastic process algebra

Wednesday, 16 August 2017 13:00 (1 hour)

Presenter: HILLSTON, Jane (University of Edinburgh)

Contribution ID: 38

Type: **Invited Speaker**

Invited Speaker - Independent component analysis using third and fourth cumulants

Thursday, 17 August 2017 11:00 (1 hour)

In independent component analysis it is assumed that the observed random variables are linear combinations of latent, mutually independent random variables called the independent components. It is then often thought that only the non-Gaussian independent components are of interest and the Gaussian components simply present noise. The idea is then to make inference on the unknown number of non-Gaussian components and to estimate the transformations back to the non-Gaussian components.

In this talk we show how the classical skewness and kurtosis measures, namely third and fourth cumulants, can be used in the estimation. First, univariate cumulants are used as projection indices in search for independent components (projection pursuit, fastICA). Second, multivariate fourth cumulant matrices are jointly used to solve the problem (FOBI, JADE). The properties of the estimates are considered through corresponding optimization problems, estimating equations, algorithms and asymptotic statistical properties. The theory is illustrated with several examples.

Presenter: OJA, Hannu (University of Turku)

Contribution ID: 39

Type: **Invited Speaker**

Invited Speaker - Independent component analysis using third and fourth cumulants

Thursday, 17 August 2017 13:00 (1 hour)

In independent component analysis it is assumed that the observed random variables are linear combinations of latent, mutually independent random variables called the independent components. It is then often thought that only the non-Gaussian independent components are of interest and the Gaussian components simply present noise. The idea is then to make inference on the unknown number of non-Gaussian components and to estimate the transformations back to the non-Gaussian components.

In this talk we show how the classical skewness and kurtosis measures, namely third and fourth cumulants, can be used in the estimation. First, univariate cumulants are used as projection indices in search for independent components (projection pursuit, fastICA). Second, multivariate fourth cumulant matrices are jointly used to solve the problem (FOBI, JADE). The properties of the estimates are considered through corresponding optimization problems, estimating equations, algorithms and asymptotic statistical properties. The theory is illustrated with several examples.

Presenter: OJA, Hannu (University of Turku)

Contribution ID: 40

Type: **Invited Speaker**

Invited Speaker - Random Networks

Friday, 18 August 2017 09:00 (1 hour)

Presenter: JANSON, Svante (Uppsala University)