The 10th Tartu Conference
on Multivariate Statistics

Abstracts

28 June - 1 July 2016, Tartu, Estonia

Tartu
2016

Tartu Ülikooli Kirjastus
www.tyk.ut.ee
Tellimus nr. ...

Editors: Tõnu Kollo, Kristi Kuljus

Dear participants,

Welcome to Tartu!

The First Tartu Conference on Multivariate Statistics was held 39 years ago, in 1977. We are happy today to have among participants the organizer and invited speaker of the first Tartu conference Professor Ene-Margit Tiit and several other participants. At the end of this volume you can find a short retrospective overview of the series of these conferences.

The talks will be given within four days, June 28-July 1, 2016. They include two keynote lectures delivered by Professors Anthony Atkinson and Thomas Mathew and fourteen invited lectures. The talks cover wide range of areas from probability theory and theoretical developments of mathematical statistics and distribution theory to applications of multivariate analysis in different areas: finance, insurance, economics, genetics, demography etc. Simo Puntanen will share with us memories of Ingram Olkin (1924-2016), the keynote lecturer at the IX Tartu Conference five years ago.

This volume contains the abstracts of the papers to be presented at the conference. The style of the abstracts has been kept unchanged during editing, only some misprints have been corrected. The organizers are grateful to all the authors for their contribution and cooperation.

The programme committee wishes to all of you fruitful ideas and enjoyable time in Tartu.

Tõnu Kollo
Vice-Chair of the programme committee

Level crossing identities under discrete observation with applications in insurance

Hansjoerg Albrecher

University of Lausanne, Switzerland, hansjoerg.albrecher@unil.ch

This talk deals with the effects of randomizing discrete observation periods on exit probabilities and related quantities of continuous-time stochastic processes. It turns out that observation of the process at epochs of an independent Poisson process leads to a number of strikingly simple analogues of classical fluctuation identities. Applications in the context of collective insurance risk models are discussed.

Markov-modulated multivariate linear regression

Alexander Andronov

Transport and Telecommunication Institute, Riga, Latvia, lora@mailbox.riga.lv

Keywords: random environment, Markov chain, estimators

We consider the case, where a process, described by multivariate linear regression, operates in a random environment. The last is presented as a continuous-time homogeneous irreducible Markov chain $J(t), t \geq 0$, with finite state set $N = \{1, 2, ..., k\}$ [1]. Let $\lambda_{i,j}$ be the known transition rate from state i to state $j\big(\lambda_{j,j} = 0\big)$.

The following notations will be used for the η-th observation ($\eta = 1, ..., n$): $x_{(\eta)} = (x_{\eta,1}, ..., x_{\eta,q})$ is the q-row vector of known independent variables; $Y_{(\eta)}(t) = (Y_{\eta,1}(t), ..., Y_{\eta,p}(t))$ is the p-row vector of observed dependent variables; $e_{(\eta)} = (e_{\eta,1}, ..., e_{\eta,p})$ is the p-row vector of random variables, $e_{(\eta)} \in N_p(0, I)$; t_η is the observation time; $T_{\eta,\mu}$ is an unobserved sojourn time in the state $\mu \in N(T_{\eta,1} + ... + T_{\eta,k} = t_\eta)$. Further the $q \times p$-matrix $B(j)$ of regression parameters for the j-th state of the random environment ($j = 1, ..., k$) and the symmetric square root $\sum^{1/2}$ of the positive definite matrix $\sum$ are unknown and identical for all observations.

Thus, if $T_{(\eta)} = (T_{\eta,1}, ..., T_{\eta,p})$ and $\tilde{B} = (B(1)^T, ..., B(k)^T)^T$, then we have the model for the η-th observation:

$$Y_{(\eta)}(t_\eta) = \big(T_{(\eta)} \otimes x_{(\eta)}\big)\tilde{B} + \sqrt{t_\eta} e_{(\eta)}, \quad \eta = 1, ..., n.$$

We consider estimators of $\tilde{B}$ and $\sum$ for the following given data on n observations: the vectors $Y_{(\eta)}$ and $x_{(\eta)}$ of dependent and independent variables; observation time t_η; the initial $i_\eta = J(0)$ and finite $j_\eta = J(t_\eta)$ states of Markov chain. It is supposed that all observations are independent.

Obtained results generalize the previous results of the author for the multiple linear regression [2].

References

[1] Pacheco, A., Tang, L.C., Prabhu, N.U. (2009). *Markov-Modulated Processes & Semiregenerative Phenomena*. World Scientific, New Jersey, London.

[2] Andronov, A. (2012). *Parameter statistical estimates of Markov-modulated linear regression*. In: Statistical Methods of Parameter Estimation and Hypothesis Testing **24**, Perm State University, Perm, Russia, 163–180 (in Russian).

Using k-anonymisation for analysis of registry data: pitfalls and alternatives

Sten Anspal

The Estonian Centre for Applied Research (CENTAR), Estonia,
sten.anspal@centar.ee

Keywords: k-anonymisation, privacy-preserving computation, linked registry data

We describe an applied study of the ICT students employment in Estonia based on linked official registry data. The study offered an opportunity to compare results from both k-anonymised data as well as those from privacy-preserving computations on the Sharemind platform ([1], [2]), which offers a way to use confidential data for research without loss of observations.

The research question was simple: What is the employment rate in general, and specifically in ICT companies, among ICT students during their studies? The question was motivated by the low rate of timely graduation (substantially lower than for non-ICT students), examining the hypothesis that the problems of dropping out and delays in graduation is a worse problem among ICT students than others due to high labour market demand for their skills and, therefore, their higher employment rates.

The study was carried out based on linked data from two registries: the Estonian Education Registry data for all students in higher education from 2006-2012 was used as the source of information on persons studies on various curricula and Tax Board data on social tax declarations in 2006-2013 was used for information on employment.

However, as a condition of using these datasets, the problem of preserving subjects privacy had to be followed. Two approaches were used and compared. The first was k-anonymisation: cases for which there were less than 3 persons with the same unique combination of characteristics were removed from query results. The second was the Sharemind platform for privacy-preserving secure computing, which made it possible to analyse the same dataset without the loss of observations due to k-anonymization, but without access to individual observations. The results of the k-anonymized and lossless analyses indicate substantial differences in employment rates of ICT and non-ICT students. Depending on the level of study (Bachelor's or Master's studies) and institution of higher education, differences in employment rates using the two approaches range from 1 to more than 10 percentage points.

The results illustrate, on the basis of a real-world study, how the effects of k-anonymization can be drastic and unpredictable in terms of inference. While the use of a share computing based privacy-preserving does entail time costs related to unobservability of individual observations and therefore additional efforts to verify the computations, these are offset by greater confidence in the results.

References

[1] Bogdanov, D., Laur, S., Willemson, J. (2008). Sharemind: A framework for fast privacy-preserving computations. *Computer Security-ESORICS 2008*, 192–206.

[2] Bogdanov, D. (2013). *Sharemind: A framework for fast privacy-preserving computations.* University of Tartu, Tartu.

Learning from rank data

Elja Arjas

University of Oslo, Norway, elja.arjas@helsinki.fi

Comparing and ranking of items according to some selected measure of quality, strength or performance is a natural way of collecting and organizing information in many areas of practical interest. The Mallows rank model (1957), a member of the exponential family, would offer an attractive alternative for describing and analysing such rank data; however, apart from situations in which the number of items is very small, the computational complexity of the Mallows model has limited its use to a particular form based on Kendall distance. In this talk I consider computationally tractable methods for Bayesian inference in Mallows models, which apply for any right-invariant metric on the space of permutations. The proposed method performs inference on the consensus ranking of the items, also when based on data on only partial rankings such as top-k items or pairwise comparisons. When the population of assessors is heterogeneous, a mixture model for clustering the assessors into homogeneous subgroups is proposed, with cluster-specific consensus rankings. An approximate stochastic sampling algorithm is introduced, enabling a fully probabilistic analysis, including quantification of the uncertainties involved. In particular, individual preferences can be predicted in situations in which they were missing in the data, and assessors can be assigned to classes after they had ranked only a subset of the items. The talk is based on joint work with Valeria Vitelli, Øystein Sørensen and Arnoldo Frigessi.

Optimum experiments with sets of treatment combinations: univariate or multivariate?

Anthony C. Atkinson

London School of Economics, UK, a.c.atkinson@lse.ac.uk

Keywords: clinical trial, D-optimality, randomization, selection bias, sequential allocation

The motivation is an experiment in deep-brain therapy in which each patient receives a set of eight distinct treatment combinations and provides a response to each. The experimental region contains sixteen different sets of eight treatments. With only six parameters in the linear model, it is unlikely that all sixteen points in the design region need to be included in the experiment. The structure of such experiments is initially elucidated in a response surface setting where each choice of an experimental setting provides a response at each of s distinct settings of the explanatory variables. An extension of the "General Equivalence Theorem" for D-optimum designs with multivariate responses is provided for experiments with sets of treatment combinations. This equivalence theorem is used to elucidate the structure of the optimum design for the experiment in deep-brain therapy. There are many possibilities, all with the same optimum properties.

In practice, patients arrive sequentially, each with an individual set of prognostic factors. Patient allocation should have a random component, to avoid selection bias. However, efficient estimation requires that the allocations be balanced over the distribution of the prognostic factors. These two requirements are in conflict. The talk will describe the application of some of the restricted randomization rules surveyed by [1] that seek a compromise between bias and information. An important measure of loss of information due to imbalances resulting from randomization is that introduced by [2]. Theory and simulation will be used to provide graphical illustration of the loss and bias of the various rules; these comparisons lead to the definition of an admissible rule.

References

[1] Atkinson, A. C. (2014). Selecting a biased-coin design. *Statistical Science* **29**, 144–163.

[2] Burman, C.-F. (1996). *On Sequential Treatment Allocations in Clinical Trials*. Department of Mathematics, Göteborg.

Goodness-of-fit tests based on the empirical characteristic function

Aleksej Bakshaev and Rimantas Rudzkis

Vilnius University, Lithuania, aleksej.bakshaev@gmail.com, rimantas.rudzkis@mii.vu.lt

Keywords: goodness-of-fit, empirical characteristic function

Let $X^n = (X_1, ..., X_n)$ be a sample of observations of a random vector X with unknown cumulative distribution function $F(x)$ and probability density function $f(x)$, $x \in \mathbb{R}^d$. The talk is devoted to the supremum-type multivariate goodness-of-fit tests based on the empirical characteristic function (ecf). Particular attention is devoted to the composite hypothesis of normality and Gaussian distribution mixtures model, which are widely applicable in classification problems. The null hypothesis assumes, that f_0 is a Gaussian density or a mixture of a known number of Gaussian densities. The alternative assumes the existence of an additional small distribution cluster g, that is

$$H_1 : f = (1 - \epsilon)f_0 + \epsilon g, \quad \epsilon \leq 1/2.$$

The problem of analytical approximation of the null distribution of the proposed test statistics, and therefore establishment of the critical region of the test, is briefly discussed. The results are obtained using the theory of high excursions of Gaussian (and, in some sense, close to Gaussian) random fields developed in [1]. Simulation study shows that the precision of the derived approximations is good enough even for small sample sizes and moderate test significance levels.

The comparative Monte Carlo power study shows that the considered tests are powerful competitors to the existing classical criteria, clearly dominating in verification of the goodness-of-fit hypotheses against the specific types of alternatives.

References

[1] Rudzkis, R. and Bakshaev A. (2012). Probabilities of high excursions of Gaussian fields. *Lithuanian Mathematical Journal* **52**, 196–213.

Dose selection in adaptive clinical trials based on pharmacokinetic and pharmacodynamic responses

Iftakhar Alam[1], Barbara Bogacka[2] and D. Stephen Coad[2]

[1] *University of Dhaka, Bangladesh, iftakhar@isrt.ac.bd*
[2] *Queen Mary, University of London, United Kingdom, b.bogacka@qmul.ac.uk, d.s.coad@qmul.ac.uk*

Keywords: area under the drug concentration, D-optimum design, mixed effects PK model, trinomial PD responses

In this talk we present an adaptive design for dose finding in phase I/II clinical trials, where the probabilities of trinomial responses of efficacy, toxicity and 'no-response' as well as pharmacokinetic (PK) information are considered in the dose-selection procedure. The local D-optimal design for estimating population PK parameters is found and applied in each step of the adaptive trial, where the responses to the drug are measured and the models updated accordingly. A new optimum dose for next cohort of patients is then selected, based on the updated information. An ethical approach is considered in this case, where the dose is optimized for efficacy of the response with some constraints on toxicity. We also consider the total exposure to the drug as an additional constraint for dose selection. The area under the drug concentration curve reflects the population variability in the drug absorption and elimination and this fact is included in the procedure. This method gives efficient designs for dose finding from the point of view of population PK parameter estimation and ethical dose selection with maximum probability of efficacy while keeping the chances of toxicity under control.

Discrete approximation theorems for statistics related to Bernoulli variables

Vydas Čekanavičius

Vilnius University, Lithuania, vydas.cekanavicius@mif.vu.lt

Keywords: total variation, m-dependent variables, 2-runs, compound Poisson approximation

We estimate the accuracy of discrete approximations to the distributions of 2-runs and $N(k_1, k_2)$ statistic. Let η_i, $(i = 1, 2, \dots)$ be independent Bernoulli variables, $\xi_j = \eta_j \eta_{j+1}$. The sum $S = \xi_1 + \cdots + \xi_n$ is called 2-runs statistic. Let $Y_j = (1 - \eta_{j-m+1}) \cdots (1 - \eta_{j-k_1}) \eta_{j-k_2} \cdots \eta_{j-1} \eta_j$. Then $Z = Y_m + Y_{m+1} + \cdots + Y_n$ is called $N(k_1, k_2)$ statistics. It is proved that for two-parametric approximations the accuracy is at least of the order $O(n^{-1/2})$. Our results are closely related to the results of [1, 2].

References

[1] Vellaisamy, P. (2004). Poisson approximation for (k_1, k_2) events via the Stein-Chen method. *Adv. Appl. Prob.* **41**, 1081-1092.

[2] Wang, X. and Xia, A. (2008). On negative binomial approximation to k-runs. *J. Appl. Prob.* **45**, 456-471.

Taxicab correspondence analysis of rank data

Vartan Choulakian

Universite de Moncton, Moncton, Canada, vartan.choulakian@umoncton.ca

Keywords: rankings, Borda count, nega coding, global homogeneity index, mixture, taxicab correspondence analysis

We consider the exploratory analysis of ranking data by taxicab correspondence analysis with the nega coding. If the first factor is an affine function of the Borda count, then we say that the ranked data are globally homogenous, and local heterogeneities appear on the consequent factors. Otherwise, the ranked data either are globally homogenous with outliers, or a mixture of globally homogenous groups. The method finds globally homogenous groups in a stepwise manner. Examples are provided.

Econometric modeling of technical provisions in insurance

Tomas Cipra and Radek Hendrych

Charles University in Prague, Czech Republic, cipra@karlin.mff.cuni.cz, hendrych@karlin.mff.cuni.cz

Keywords: econometric models, life insurance, non-life insurance, technical provisions (reserves), Solvency II

Econometric models of cash-flows applying econometric instruments (e.g. simultaneous equation models SEM, vector autoregression VAR) in life and non-life insurance may provide results that are useful both for individual insurance companies (e.g. the internal models in Solvency II) and for global insurance data (e.g. perspectives for insurance industry in a given country or EU).

The first part of the contribution deals with econometric modeling of the Czech life insurance market by means of the dynamic econometric system of linear simultaneous equations. The model enables to describe and explain technical-actuarial relations among important insurance variables including the technical provisions (technical reserves). From the statistical point of view, capabilities of adjusted residual bootstrapping in the connection with the considered econometric model are analyzed since this technique can solve eventual inaccuracies caused by applying the theoretical asymptotic distribution of residuals (more generally, it can be applied in a broad statistical context, e.g. for the significance testing). The contribution deals also with prognoses and scenario generations (optimistic, pessimistic or randomly generated anticipations) which should be taken into account in practice.

The second part of the contribution deals primarily with non-life technical provisions based on the Czech insurance market data (mainly the provision for outstanding claims and the provision for unearned premium). The resulting model can be used for predicting these technical reserves. Both deterministic and randomly generated scenarios are considered. Such an econometric modeling seems to be useful e.g. for stress testing (i.e. for the insurance regulator) or for the internal models in the framework of Solvency II.

Randomly stopped sum of distributions with dominatingly varying tails

Svetlana Danilenko

Vilnius Gediminas Technical University, Lithuania, svetlana.danilenko@vgtu.lt

Keywords: heavy tail, dominatingly varying tail, random sum, closure property

Heavy tailed random variables are useful in the insurance stochastic models. Usually such random variables describe a series of claim amounts. Various subclasses of heavy tailed random variables are considered. The best known subclasses are $\mathcal{L}$, $\mathcal{D}$ and $\mathcal{S}$. It should be recalled that:

- distribution function (d.f.) $F = 1 - \overline{F}$ is said to be heavy-tailed ($F \in \mathcal{H}$) if $\lim_{x\to\infty} \overline{F}\mathrm{e}^{\delta x} = \infty$ for an arbitrary positive δ;
- d.f. F is said to be long-tailed ($F \in \mathcal{L}$) if $\overline{F}(x+y) \sim \overline{F}(x)$ for every positive y;
- d.f. F has dominatingly varying tail ($F \in \mathcal{D}$) if $\limsup_{x\to\infty}(\overline{F}(xy)/\overline{F}(x)) < \infty$ for some $y \in (0,1)$;
- d.f. F is subexponential ($F \in \mathcal{S}$) if $\overline{F_+ * F_+}(x) \sim 2\overline{F}(x)$, where F_+ denotes the positive part of d.f. F.

It is known (see, for instance, [3]) that $\mathcal{L} \cap \mathcal{D} \subset \mathcal{S} \subset \mathcal{L} \subset \mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$.

Various properties of classes $\mathcal{L}$, $\mathcal{D}$ and $\mathcal{S}$ have been considered by many authors. For instance, in [1] the problem of max-sum equivalence and the problem of convolution closure were considered, while in [4] the problem of random convolution closure was investigated.

Particulary, in [4] conditions were obtained under which d.f. of random sum of independent and identically distributed random variables $\xi_1 + \xi_2 + \ldots + \xi_\eta$ belongs to the class $\mathcal{D}$. One can show that the similar results can be obtained in the case when random variables $\xi_1, \xi_2, \ldots$ are independent, but not necessary identically distributed. The exact formulations of the results together with its detailed proofs can be found in [2].

References

[1] Cai, J. and Tang, Q. (2004). On max-sum equivalence and convolution closure of heavy-tailed distributions and their applications. *Journal of Applied Probability* **41**, 117–130.

[2] Danilenko, S. and Šiaulys, J. (2016). Randomly stopped sums of not identically distributed heavy-tailed random variables. *Statistics and Probability Letters* **113**, 84–93.

[3] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer, Berlin.

[4] Leipus, R. and Šiaulys, J. (2012). Closure of some heavy tailed distribution classes under random convolution. *Lithuanian Mathematical Journal* **52**, 249–258.

Closure property and tail probability asymptotics for randomly weighted sums of dependent random variables with heavy tails

Lina Dindienė, Remigijus Leipus and Jonas Šiaulys

Vilnius University, Lithuania, lina.dindiene@gmail.com

Keywords: randomly weighted sum, long-tail distribution, copula, FGM copula

We consider the closure property and probability tail asymptotics for randomly weighted sums $S_n^{\Theta} = \Theta_1 X_1 + \cdots + \Theta_n X_n$ for long-tailed primary random variables $X_1, \ldots, X_n$ and positive random weights $\Theta_1, \ldots, \Theta_n$ under similar dependence structure as in [1]. In particular, we study the case where the distribution of random vector $(X_1, \ldots, X_n)$ is generated by an absolutely continuous copula.

References

[1] Yang, Y., Leipus, R., Šiaulys, J. (2014). Closure property and maximum of randomly weighted sums with heavy tailed increments. *Statistics and Probability Letters* **91**, 162–170.

Kernel based classification for financial failure: A comparative study for Turkish banks

Birsen Eygi Erdogan

Marmara University, Turkey, birsene@marmara.edu.tr

Keywords: bank failures, classification, kernel, panel data, support vector machines

Classification of banks as weak and strong is crucial to the entire economic system. Therefore various classification methods were proposed using either parametric or non-parametric approaches. In this study kernel based support vector machines were applied on longitudinal financial ratios to discriminate between weak and strong banks. The success status of the banks was used as the dependent variable whereas the financial ratios were used as independent variables. For the comparison of the modelling performances the classification measures were used. It was concluded that kernel based support vector machines is a very promising approach for bank classification studies with an appropriate kernel choice.

References

[1] Balcaen, S. and Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review* **38**, 63–93.

[2] Canbas, S., Cabuk, A., Kilic, S.B. (2005). Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. *European Journal of Operational Research* **166**, 528–546.

[3] Demyanyk, Y. and Hasan, I. (2010). Financial crises and bank failures: A review of prediction methods. *Omega: The International Journal of Management Science* **38**, 315–324.

[4] Eygi Erdogan, B. (2013). Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. *Journal of Statistical Computation and Simulation* **83**, 1543–1555.

[5] Ilk, O., Pekkurnaz, D., Cinko, M. (2014). Modeling company failure: a longitudinal study of Turkish banks. *Optimization: A Journal of Mathematical Programming and Operations Research*, **63**, 1837–1849.

[6] Kumar, P.R. and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review. *European Journal of Operational Research* **180**, 1–28.

Sampling finite populations using supersaturated designs

Kent M. Eskridge[1], **Xiaojuan Hao, Juan Diego Hernandez Jarquin and George L. Graef**

[1] *University of Nebraska - Lincoln, USA, keskridge1@unl.edu*

Keywords: fractional factorials, genetic resources, genomic selection

Supersaturated designs are two-level factorial designs useful for screening a large number of factors (p) with a limited number of runs (n) where $n < p$. Substantial work has been done on constructing supersaturated designs for various optimality criteria when there are no restrictions on the treatment combinations to be considered. However, little work has been done on identifying optimal supersaturated designs when candidate design points must be selected from a restricted set of treatment combinations. Constructing supersaturated designs for such finite populations is an important problem in genetics. For example, a large number of p genetic markers may be available on N ($< p$) individuals where the interest is to assess how the markers are related to some end point variable such as disease level or yield. Often the individuals' end point variables are not available and collecting such data on all N individuals is too costly resulting in the need to use the genetic information to identify the n ($< N$) most informative individuals for which the endpoint data should be obtained. The objective of this work is to develop and compare several different criteria in identifying supersaturated designs comprised of the n most informative individuals from a finite population of N individuals where $N < p$ and to compare these criteria in their ability to (1) maximize genetic variance among individuals and (2) identify important markers when predicting end point variables based on the resulting models. The methods are applied to the USDA Germplasm Resource Information System (GRIN) database of ∼20,000 soybean accessions using 50K SNP genetic markers. Multi-environment field trial evidence indicates substantial differences between the methods in terms of genetic information and the predictive ability of the resulting models.

Survival analysis of biobank data: some methodological aspects and examples

Krista Fischer and Kristi Läll

University of Tartu, Estonia, Krista.Fischer@ut.ee, Kristi.Lall@ut.ee

Keywords: survival analysis, nested case-control studies

Recent decades have seen a considerable increase in the availability of data from large prospective biobank cohorts. As the follow-up time increases, analysis of overall survival or disease incidence becomes more feasible and may provide valuable new information on disease and mortality risks in general population. While in conventional epidemiological cohort studies one is mainly interested in the effects of various lifestyle and/or clinical characteristics, the biobank-based studies are often focused on *omics*-based parameters, aiming for discovery of novel predictive biomarkers.

In such studies, several methodological aspects need to be considered. The first issue is the timescale to be used in the analysis. For instance, the DNA-based variables (data on single nucleotide polymorphisms, SNPs) are fixed before birth of the individual and may influence the entire lifespan. As the biobank cohorts have recruited people of different ages, the data is both right-censored and left-truncated. When the DNA-based markers are combined with other potential predictors that have measured at recruitment and, contrary to the DNA, vary in time, the question of proper adjustment arises. We discuss different options and use simple simulations to illustrate how the results would differ when different timescale is used in such analysis.

The second issue is related to the large size of the datasets. The use of Cox proportional hazards model in genome-wide analysis of several million markers in large cohorts ($n > 10000$) is complicated due to the relative slowness of the conventional fitting algorithm. We propose a two-stage algorithm based on martingale residuals to overcome this problem and implement this to identify survival-related genetic variants in the UK Biobank cohort [1].

Finally, we discuss the potential of nested case-control design in cases where molecular data can only be obtained for a limited subset of the cohort, using some examples based on the Estonian Biobank data.

References

[1] Joshi, P.K., Fischer, K., Schraut, K.E., Campbell, H., Esko, T., Wilson, J.F. (2016). Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nature Communications* **7**, 11174.

Segmentation of hidden Markov tree models with hybrid decoders

Mark Gimbutas and Jüri Lember

University of Tartu, Estonia, gimbutas@ut.ee, juri.lember@ut.ee

Keywords: hidden Markov tree models, Viterbi algorithm, posterior decoding, segmentation, Bayesian inference

Hidden Markov models have proven useful in practice for partitioning a sequence of given observations $\bar{\mathbf{X}} = (X_1, X_2, \ldots, X_T)$ into segments according to unobserved discrete variables $\bar{\mathbf{S}} = (S_1, S_2, \ldots, S_T)$. Many different optimality criteria and corresponding computational algorithms have been proposed to recover the hidden $\bar{\mathbf{S}}$, most famous of them are the Viterbi algorithm and the posterior decoding algorithm. A computationally feasible interpolation between the two (a *hybrid decoder*) was presented in [1], which combines favourable aspects of both methods.

We will show that the hybrid decoder can be directly generalised to the case when the observations have a tree structure rather than a sequence structure. Hidden Markov tree models were introduced in [2] and are defined to have similar conditional independence properties to those of hidden Markov (chain) models. For example, the hidden $\bar{\mathbf{S}}$ is assumed to satisfy the global Markov property with respect to the tree with vertex set V. The hybrid decoder works by maximising the product

$$\left(\prod_{v \in V} p(S_v = s_v \mid \bar{\mathbf{X}} = \bar{\mathbf{x}}) \right)^{\alpha} p(\bar{\mathbf{S}} = \bar{\mathbf{s}} \mid \bar{\mathbf{X}} = \bar{\mathbf{x}})^{1-\alpha}$$

over hidden states $\bar{\mathbf{s}} = (s_v)_{v \in V}$ for a fixed interpolation parameter $\alpha \in [0, 1]$ and fixed observations $\bar{\mathbf{x}} = (x_v)_{v \in V}$. The task can be viewed as minimisation of a certain risk function endowed with an interpretation comparable to that of Rabiner's k-blocks in [3]. The computational feasibility is fully retained, owing much to the methods described in [4].

References

[1] Lember, J., Koloydenko, A. A. (2014). Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on hidden Markov models. *Journal of Machine Learning Research* **15**, 1–58.

[2] Crouse, M. S., Nowak, R. D., Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46**, 886–902.

[3] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

[4] Durand, J.-B., Gonçalvès, P., Guédon, Y. (2004). Computational methods for hidden Markov tree models – an application to wavelet trees. *IEEE Transactions on Signal Processing* **52**, 2551–2560.

Beta-hypergeometric probability distribution on symmetric matrices

Abdelhamid Hassairi

Sfax University, Tunisia, abdelhamid.hassairi@fss.rnu.tn

We first give some properties based on independence relations between matrix beta random variables of the first kind and of the second kind which are satisfied under a condition on the parameters of the distributions. We then use results on Jordan algebras and their symmetric cones to introduce a class of matrix-variate beta-hypergeometric distributions containing the beta ones as a particular case. We show that with these distributions, the properties established for the beta distributions are satisfied without any condition on the parameters. The results involve many remarkable properties of the zonal polynomials with matrix arguments and the use of random matrix continued fractions.

On elliptical multivariate quantiles

Daniel Hlubinka[1] **and Miroslav Šiman**[2]

[1] *Charles University in Prague, Czech Republic, hlubinka@karlin.mff.cuni.cz*
[2] *Ústav teorie informace a automatizace AV ČR, Prague, Czech Republic*

Keywords: multivariate quantile, elliptical quantile, quantile regression, multivariate statistical inference

Inspired by nonlinear quantile regression, we propose a concept of elliptical location quantiles in regression setup. For m-variate response $\mathbf{Y}$ and p-variate regressor $\mathbf{Z}$ we define elliptical quantile region as

$$\mathcal{E}_\tau(\mathbf{Y},\mathbf{Z}) = \{(y,z); (y - s_\tau(z))^T A_\tau(z)(y - s_\tau(z)) < c_\tau(z)\},$$

where $s_\tau(z) \in \mathbb{R}^m$ (the centre of the ellipsoid), $A_\tau(z)$ is $m \times m$ (non-singular matrix with unit determinant), and $c_\tau(z)$ are functions of the regressor such that the overall coverage of the elliptical quantile region is τ.

The elliptical quantiles are affine equivariant and may be computed quite efficiently even for large datasets. It is also possible to cover quite general trends and various forms of heteroscedasticity. Therefore the elliptical quantiles may be a good option for testing trend or symmetry of the response variables. It is also possible to incorporate many types of a priori information regarding the model parameters.

References

[1] Hlubinka, D. and Šiman, M. (2013). On elliptical quantiles in the quantile regression setup. *Journal of Multivariate Analysis* **116**, 163–171.

[2] Hlubinka, D. and Šiman, M. (2015). On generalized elliptical quantiles in the nonlinear quantile regression setup. *Test* **24**, 249–264.

A novel modelling approach to increase the explained risk in the proportional hazards regression

Deniz İnan[1], Öyküm Esra Aşkın[2] and Busenur Sarıca[1]

[1]*Marmara University, İstanbul, Turkey, denizlukuslu@marmara.edu.tr, busenur.sarica@marmara.edu.tr*

[2]*Yıldız Technical University, İstanbul, Turkey, oykumesra@gmail.com*

Keywords: proportional hazards regression models, censored data, fuzzy c-means algorithm, multivariate gaussian membership function

In this study, a modelling strategy to increase the information obtained from censored observations is proposed. For this purpose uncensored observations are clustered using fuzzy c-means algorithm and multivariate gaussian membership functions are determined on each cluster. Censored observations are weighted considering membership values and the distances between censoring time and the time component of centers. Simulation studies are performed to investigate the performance of proposed approach according to measure of explained risk.

References

[1] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics* **3**, 32–57.

[2] Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell, MA, USA.

[3] Lemos, A., Caminhas, W., Gomide, F. (2013). Adaptive fault detection and diagnosis using an evolving fuzzy classifier. *Information Sciences* **220**, 64–85.

[4] Heller, G. (2012). A measure of explained risk in the proportional hazards model. *Biostatistics* **13**, 315–325.

About nearly critical branching processes with dependent immigration

Ya. M. Khusanbayev[1], **G. Rakhimov**[2] **and X. Q. Jumaqulov**[1]

[1]*Institute of Mathematics, Uzbekistan, yakubjank@mail.ru, xurshid81@gmail.com*
[2]*Academic Lyceum under Architecture and Building Institute, Uzbekistan, gairat48@gmail.com*

Keywords: branching process, immigration, stationary process

Let $\{\xi_{k,\,j},\ k,j \in N\}$ and $\{\varepsilon_k,\ k \in N\}$ be independent collections of independent, nonnegative, integer-valued, identically distributed random variables. We define a sequence of random variables $\{X_k,\ k \in N_0\}$ by the following recurrence relations:

$$X_0 = 0,\ X_k = \sum_{j=1}^{X_{k-1}} \xi_{k,j} + \varepsilon_k,\quad k \in N.$$

The sequence $\{X_k,\ k \in N_0\}$ is called a branching process with immigration.

Let for each $n \in N$ $\left\{\xi_{k,j}^{(n)},\ k,j \in N\right\}$ be independent, nonnegative integer-valued random variables and $\left\{\tau_k^{(n)},\ k \in N\right\}$ be stationary process in a broad sense, the random variables $\tau_k^{(n)}$ taking nonnegative integer values. Suppose that for each $n \in N$ a set of random variables $\left\{\xi_{k,j}^{(n)},\ k,j \in N\right\}$ and the process $\left\{\tau_k^{(n)},\ k \in N\right\}$ are independent. We consider a sequence of branching processes with immigration $\left\{Z_k^{(n)},\ k \in N_0\right\}$, $n \in N$, following recurrence relations

$$Z_0^{(n)} = 0,\ Z_k^{(n)} = \sum_{j=1}^{Z_{k-1}^{(n)}} \xi_{k,j}^{(n)} + \tau_k^{(n)},\quad k,\ n \in N.$$

We define a random step function $Z_n(t)$, $n \in N$, by setting $Z_n(t) = Z_{[nt]}^{(n)},\ t \geq 0$, where [a] is the integer part of a. Assume that

$$m_n = E\xi_{1,1}^{(n)},\ \sigma_n^2 = var\,\xi_{1,1}^{(n)},\ \gamma_n = E\tau_1^{(n)},\ \delta_n^2 = var\,\tau_1^{(n)},\ \rho_n(k) = cov\left(\tau_1^{(n)},\ \tau_{k+1}^{(n)}\right)$$

are finite for all $n \in N$.

Theorem. Suppose that the following conditions are satisfied:
A. $m_n = 1 + \alpha n^{-1} + o(n^{-1})$ as $n \to \infty$ for any $\alpha \in R$;
B. $\sigma_n^2 \to 0$ as $n \to \infty$;
C. $\gamma_n \to \gamma \geq 0,\ \delta_n^2 \to \delta^2 \geq 0$ as $n \to \infty$;
D. $\frac{1}{n}\sum_{k=1}^{n} |\rho_n(k)| \to 0$ as $n \to \infty$.
Then the following weak convergence takes place in the Skorokhod space $D[0,\ \infty)$ as $n \to \infty$: $n^{-1}Z_n \to \mu$, where μ is defined by the relation $\mu(t) = \gamma \int_0^t e^{\alpha u} du$.

The case of independent, nonnegative, integer-valued, identically distributed random variables $\left\{\tau_k^{(n)},\ k \in N\right\}$ was studied in [1].

References

[1] Ispany, M., Pap, G., Van Zuijlen M.C.A. (2005). Fluctuation limits of branching processes with immigration and estimation of the means. *Adv. Appl. Probab.* **37**, 523–538.

Experimenting with language vitality in a virtual laboratory using data-based evolutionary computation

Andres Karjus

University of Tartu, Estonia, andres.karjus@ut.ee

Keywords: agent-based modelling, language vitality, evolutionary computation, experimental design

The survival of human languages depends on them being used by communities of speakers. It is reasonable to assume that languages are not adopted and discarded on random. Language being an inherently complex domain, researchers working on the growth, survival and death of languages have increasingly turned to models from the field of evolutionary computation, capable of handling complex multivariate relationships. Castelló et al. [1] used a simple agent-based model (ABM), inspired by earlier mathematical work [2], to investigate the problem of language competition, whereby two languages compete to be used by speakers. Similar models have been used by other researchers, and it has been found that a relatively simple model is often capable to explain the rise and decline of minority languages in real-life scenarios.

ABMs in the field of language vitality tend to consist of simple agents and general parameters such as community-wide system volatility and relative prestige. While such models are highly abstract and generalizable, they also produce relatively simple, univariate outcomes, and usually do not allow for experimentation with variation and its causes on the level of the individual agents. This research is based on data on linguistic attitudes collected in Estonia using a questionnaire developed by the Sustainability of Estonian in the Era of Globalisation research group. The agents in the ABM are directly based on the respondents of the questionnaire ($n = 1000$), including their skills in the various languages spoken in Estonia, their attitudes towards speaking other languages and their self-reported social activeness.

An abstract negotiation-based communication model has been implemented, where languages are chosen and rejected by agents probabilistically. Successful interactions incrementally strengthen the links in the network of agents. A single simulation is left to iterate until it converges on stable network. The population mean values (e.g., fractions of the languages spoken) are then validated against (non-included variables in the) questionnaire data. This approach has allowed for the construction of virtual communities of agents representative of the populations of actual Estonian communities. With the validated models at hand, the second stage of the research, currently a work in progress, consists of experimentation with various scenarios, including the influx of speakers of a foreign language.

References

[1] Castelló, X., Loureiro-Porto, L., San Miguel, M. (2013). Agent-based models of language competition. *International Journal of the Sociology of Language* **221**, 21–51.

[2] Abrams, D.M., Strogatz, S.H. (2003). Modelling the dynamics of language death. *Nature* **424**, 900.

Statistical analysis of high-order Markov dependencies

Yuriy Kharin and Michail Maltsew

Belarusian State University, Minsk, Belarus, kharin@bsu.by, maltsew@bsu.by

Keywords: high-order Markov chain, parsimonious model, estimator, test

A universal model for real-world processes with discrete time t, finite state space $A = \{0, 1, \dots, N-1\}$ and high-order dependence $s \gg 1$ (in genetics, computer networks, financial markets, meteorology and other fields) is the order s homogeneous Markov chain ($MC(s)$) x_t on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$ determined by an $(s+1)$-dimensional matrix of one-step transition probabilities $P = (p_{i_1,\dots,i_{s+1}})$, $p_{i_1,\dots,i_{s+1}} = \mathbf{P}\{x_{t+1} = i_{s+1} | x_t = i_s, \dots, x_{t-s+1} = i_1\}$. Unfortunately, the number of independent parameters for the $MC(s)$ increases exponentially w.r.t. the order s: $D_{MC(s)} = N^s(N-1)$, and we need data of huge size to identify this model.

To avoid this "curse of dimensionality" we propose to use parsimonious (or "small-parametric" [1]) models for $MC(s)$ that are determined by small number of parameters $d \ll D_{MC(s)}$. Three known examples of parsimonious models: the Jacobs-Lewis model with $d_{JL} = N + s - 1$ parameters; the MTD-model proposed by A. Raftery with $d_{MTD} = N^2 + s - 1$ parameters; the variable length Markov chain model proposed by P. Buhlmann.

We give short analysis of previous results and propose two new parsimonious models of $MC(s)$.

Markov chain of order s with r partial connections $MC(s, r)$ is determined by the following small-parametric form of P:

$$p_{i_1,\dots,i_{s+1}} = q_{i_{m_1},\dots,i_{m_r},i_{s+1}},\ i_1, \dots, i_{s+1} \in A,$$

where $r \in \{1, \dots, s\}$ is the number of connections; $M_r = (m_1, \dots, m_r)$ is the integer-valued vector with r ordered components $1 = m_1 < m_2 < \dots < m_r \le s$, called the template of connections; $Q = (q_{i_1,\dots,i_r,i_{r+1}})$ is a stochastic $(r+1)$-dimensional matrix. We need $N^r(N-1)$ parameters to completely determine $MC(s, r)$.

Markov chain of conditional order $MCCO(s, L)$ is determined by the equation:

$$p_{i_1,\dots,i_{s+1}} = \sum_{k=0}^{N^L-1} \mathrm{I}\{< i_{s-L+1}, \dots, i_s >= k\} q^{(m_k)}_{i_{s-s_k+1},i_{s+1}},$$

where $\mathrm{I}\{C\}$ is the indicator of event C; $< i_{s-L+1}, \dots, i_s >= \sum_{k=s-L+1}^{s} N^{k-s+L-1} i_k$ is the numeric representation of the sequence $(i_{s-L+1}, \dots, i_s)$, called the base memory fragment of length $L \in \{1, \dots, s-1\}$; $Q^{(1)}, \dots, Q^{(M)}$ are $M \in \{1, \dots, N^L\}$ different stochastic square matrices of the order N: $Q^{(m_k)} = (q^{(m_k)}_{i,j})$, $1 \le m_k \le M$; the value $s_k \in \{L+1, \dots, s\}$ is called the conditional order. The transition matrix for the $MCCO(s, L)$ is determined by $2(N^L + 1) + MN(N-1)$ parameters.

We present theoretical results on probabilistic properties and statistical inferences for these models and results of computer experiments on simulated and real data.

References

[1] Kharin, Yu. (2013). *Robustness in Statistical Forecasting*. Springer, Heidelberg/Dordrecht/New York/London.

Multicomponent stress-strength reliability for a multivariate Weibull distribution

Fatih Kızılaslan

Marmara University, İstanbul, Turkey, fatih.kizilaslan@marmara.edu.tr

Keywords: multivariate Weibull distribution, stress-strength model, s-out-of-k : G system

An s-out-of-k : G system consists of k component functions if and only if at least s components function. The s-out-of-k : G system has wide applications in both industrial and military systems. For example, an automobile with four tires is usually equipped with one additional spare tire. Hence, the vehicle can be driven as long as at least 4-out-of-5 tires are in good condition. For an extensive reviews of s-out-of-k and related systems, see [3].

In this study, we consider the s-out-of-k : G system which has k statistically independent and identically distributed strength components (or subsystems) each consisting of m statistically dependent elements. The system is subjected to a common random stress and works if at least s $(1 \leq s \leq k)$ components simultaneously operate. In that case, the strength component (or subsystem) is alive only if the weakest element is operating, namely it is regarded as a series system.

We assume that each strength component, namely $(X_{i1}, X_{i2}, ..., X_{im})$, $i = 1, ..., k$, follows a multivariate Weibull (MVW) distribution (see [2] and [4]) and a common random stress variable T follows a Weibull distribution. Similar system is considered in [5] when the strength components consist of a pair of statistically dependent elements. Let $Z_i = \min(X_{i1}, X_{i2}, ..., X_{im})$, $i = 1, ..., k$. In terms of these random variables, the system is working if at least s $(1 \leq s \leq k)$ of the Z_i strength variables operate when the common stress variable T is carried out. The reliability of this multicomponent stress-strength model is given by $R_{s,k} = P(\text{at least } s \text{ of the } (Z_1, ..., Z_k) \text{ exceed } T)$ (see [1]). In this study, the estimates of $R_{s,k}$ are investigated by using classical and Bayesian approaches. Then, the derived estimates are compared through Monte Carlo simulations.

References

[1] Bhattacharyya, G. K., Johnson, R. A. (1974). Estimation of reliability in multicomponent stress-strength model. *Journal of the American Statistical Association* **69**, 966–970.

[2] Hanagal, D.D. (1996). A multivariate Weibull distribution. *Economic Quality Control* **11**, 193–200.

[3] Kuo, W., Zuo, M.J. (2003). *Optimal Reliability Modeling, Principles and Applications*. Wiley, New York.

[4] Marshall, A.W., Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association* **62**, 30–44.

[5] Nadar, M., Kızılaslan, F. (2016). Estimation of reliability in a multicomponent stress-strength model based on a Marshall-Olkin bivariate Weibull distribution. *IEEE Transactions on Reliability* **65**, 370–380.

On consecutive alignment with priority in random sequence comparison

Riho Klement and Jüri Lember

University of Tartu, Estonia, riho.klement@ut.ee

Keywords: random sequence comparison, longest common subsequence, suboptimal alignments

One way to measure similarity of two sequences is to calculate the length of the longest common subsequence (LCS). Already with quite short sequences, calculating the length of the longest common subsequence can be too time-consuming. So instead of that we try to find some suboptimal alignments which have similarity scores not far away from the optimal score (length of LCS), but are less resource-demanding to compute.

In this talk we consider an alignment referred to as *consecutive alignment with priority*. Let $X^n = (X_1, \dots, X_n)$ and $Y^n = (Y_1, \dots, Y_n)$ be two mutually independent sequences which both have independent and identically distributed elements taking values from some finite alphabet $\mathcal{A} = \{v_1, v_2, \dots, v_k\}$. Let $P^x = (P_1^x, P_2^x, \dots, P_k^x)$ be the distribution of X_i and $P^y = (P_1^y, P_2^y, \dots, P_k^y)$ the distribution of Y_i. We define $l_i = P_i^x \wedge P_i^y$ and order our letters decreasingly by l_i. Now we take the letters with the highest l_i, let's say v_1. We take the first v_1 in one sequence and align it with the first v_1 in another sequence. Then we align the next pair of v_1-s and so on until we have one sequence with no more v_1-s left. After that we take the letter with the next highest l_i, say v_2, and align them in the same way as first letters, but we keep in mind that we don't ruin the alignment we already have (that is we can align v_2-s only between already aligned pairs of v_1-s). We continue in the same manner until all letters $v_1, \dots, v_k$ are aligned.

Our goal is to calculate the average score of the alignment described above and to prove the large deviation inequality for that score.

Asymptotic normality of estimators for skew-normal distribution

Tõnu Kollo, Meelis Käärik and Anne Selart

University of Tartu, Estonia, tonu.kollo@ut.ee

Keywords: asymptotic normal distribution, multivariate skewness vector, skew-normal distribution

A. Azzalini and A. Dalla Valle introduced multivariate skew-normal distribution in the seminal paper in 1996 (Azzalini & Dalla Valle, (1996)). In Käärik, Selart & Käärik (2015) different parametrizations of skew-normal distribution were considered. In this talk we present asymptotic normal distributions for the shape/skewness vector and the dispersion matrix of the multivariate skew-normal distribution for two parametrizations. Also, an analytic expression and asymptotic normal law are derived for the skewness vector of the skew-normal distribution (Kollo, 2008). Expressions of the first four moments of the skew-normal distribution are used in derivation. Matrix derivative technique is applied for deriving the asymptotic distributions. Convergence to the asymptotic normal laws is examined both computationally and in a simulation experiment.

References

[1] Azzalini A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

[2] Käärik, M., Selart, A., Käärik, E. (2015). On parametrization of multivariate skew-normal distribution. *Communications in Statistics - Theory and Methods* **44**, 1869–1885.

[3] Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ICA. *J. Multivariate Anal.* **99**, 2328–2338.

Comparison of Euclidean and prominent non-Euclidean weighted averages of covariance matrices

Alexey A. Koloydenko[1], Ian L. Dryden[2], Diwei Zhou[3] and Koenraad M.R. Audenaert[1]

[1]*Royal Holloway, University of London, UK, alexey.koloydenko@rhul.ac.uk*
[2]*University of Nottingham, UK,* [3]*Loughborough University, UK*

Keywords: manifold, metric, positive definite, power, Procrustes, Riemannian, symmetric, smoothing, weighted Fréchet mean

As different metrics continue to be considered for measuring distances between symmetric positive semi-definite (SPD) matrices (e.g. covariance matrices), we have compared the more prominent such metrics. Our focus is on the matrix size as measured by the trace and determinant, although we also study matrix shape as measured by fractional anisotropy. For example, Diffusion Tensor Imaging (DTI), using the Euclidean distance to process covariance matrices preserves the trace and subsequenty the mean diffusivity. However, the same Euclidean approach is also often criticised for its "swelling" effect on the determinant, and for possible violation of positive definiteness in extrapolation. The affine invariant and log-Euclidean Riemannian metrics have been subsequently proposed to remedy these deficiencies. However, practitioners have also argued that these geometric approaches might be an overkill in DTI applications. We examine alternatives that in a sense reside between the Euclidean (arithmetic) and affine invariant Riemannian (geometric) extremes. These alternatives are based on the principal square root Euclidean metric and on the Procrustes size-and-shape metric. Unlike the above Riemannian metrics, these root based metrics operate more naturally (in our opinion) with regard to the boundary of the cone of SPD matrices. In particular, we prove that the Procrustes metric, when used to compute weighted averages of two SPD matrices, preserves matrix rank. We also establish and prove a key relationship between these two metrics, as well as inequalities ranking traces and determinants of weighted averages based on the Riemannian, Euclidean, and our alternative metrics. Remarkably, traces and determinants of our alternative interpolants compare differently. Experimental illustrations will also be shown. This discussion is based on [1].

References

[1] Zhou, D., Dryden, I.L., Koloydenko, A.A., Audenaert, K.M.R., Bai, L. (2016). Regularisation, interpolation and visualisation of diffusion tensor images using non-Euclidean statistics. *Journal of Applied Statistics* **43**, 943–978.

Multivariate models connected with random sums and maxima

Marek Arendarczyk[1], **Tomasz J. Kozubowski**[2] **and Anna K. Panorska**[3]

[1] *University of Wroclaw, Poland, marendar@math.uni.wroc.pl*
[2] *University of Nevada, USA, tkozubow@unr.edu*
[3] *University of Nevada, USA, ania@unr.edu*

Keywords: dependence by mixing, extremes, Lomax distribution, stochastic representation

We present recent results concerning a stochastic model for (X, Y, N), where X and Y, respectively, are the sum and the maximum of N dependent, heavy-tail Pareto components. Models with this or similar structure are desirable in many applications, ranging from hydro-climatology to finance and insurance. Our construction is built upon a pivotal model described in [1], involving a deterministic number of i.i.d. exponential variables, where the basic characteristics of the involved multivariate distributions admit explicit forms. In addition to theoretical results, we shall present real data examples, illustrating the applications of the model.

References

[1] Qeadan, F., Kozubowski, T.J., Panorska, A.K. (2012). The joint distribution of the sum and the maximum of n i.i.d. exponential random variables. *Comm. Statist. Theory Methods* **41**, 544–569.

Parameter estimation of stable laws

Annika Krutto

University of Tartu, Estonia, annika.krutto@ut.ee

Keywords: characteristic function, method of moments, parameter estimation, simulation

Stable laws form a four-parameter class of infinitely divisible distributions that have many mathematically intriguing properties. They allow skewness and heavy tails and are proposed as models for various processes in physics, finance and elsewhere. Alas, explicit representations for the density function of stable laws in terms of elementary functions are unknown. The deficiency of closed form for density complicates the estimation of parameters of stable distributions. A number of techniques are based on the empirical characteristic function. The motivation for this study arises from one of such procedures, known as the method of moments [1]. The method yields explicit point estimators for all four parameters but leaves open the problem that these estimates depend on an arbitrary choice of two pairs of arguments of the characteristic function. In this study an amended version of the method of moments is proposed. To validate the effectiveness of the estimates extensive simulation experiments over the entire parameter space are carried out. The performance of estimation procedure is illustrated with an application to non-life insurance claims.

References

[1] Press, S. J. (1972). Estimation in univariate and multivariate stable distribution. *Journal of the American Statistical Association* **67**, 842–846.

On estimation of insurance claim numbers by combining local regression and distribution fitting ideas

Meelis Käärik, Raul Kangro and Liina Muru

University of Tartu, Estonia, meelis.kaarik@ut.ee

Keywords: premium estimation, local regression, distribution fitting

In this paper we address the well-known problem of premium estimation in non-life insurance. More precisely, we are modelling the claim numbers, which can be seen as one component of the premium. We are looking for certain dynamic regression type model to avoid the "price shock" issue of static classification. We also take into account that it is hard to specify the form of suitable regression functions, and simple choices of such functions usually have undesirable effects by implicitly implying that risk behaviour of clients corresponding to one region of values of regression variables contain information about the risk behaviour of clients corresponding to a very different region of the same variables. Thus we are proposing certain local regression model, where for each new client we first fix a neighborhood of similar clients (depending on the values of certain argument variables) and then apply the local regression model on this neighborhood. Local maximum likelihood estimation is used to determine the parameters of the model and cross-validation techniques are used to determine the optimal size of the neighborhood. As a result, we propose certain semiparametric model for estimating the risk parameters for each new client. A case study with real vehicle casco insurance dataset is included, the results obtained by proposed method are compared by the ones obtained by global regression and the classification and regression trees (C&RT) approach.

On random processes as an implicit solution of equations

Petr Lachout

Charles University in Prague, Czech Republic, Petr.Lachout@mff.cuni.cz

Keywords: econometric models, implicit definition, ARMA process

To describe observed data in economics, several convenient models were introduced. Large family of such models provides an implicit description of a random process. It is required the random process is a solution of a given system of equations. And, an additional requirement is that the considered process must be causal and stationary.

Widely used processes as AR, MA, ARMA, ARCH, GARCH, etc. belong to this class. We will present a discussion on uniqueness of the solution together with a numerical study.

References

[1] Brockwell, P.J. and Davis, R.A. (1987). *Time Series: Theory and Methods*. Springer-Verlag, New York.

[2] Liu, J. and Susko, E. (1992). On strict stationarity and ergodicity of a non-linear ARMA model. *Journal of Applied Probability* **29**, 363–373.

[3] Shumway, R.H. and Stoffer, D.S. (2015). *Time Series Analysis and Its Applications - With R Examples*. EZ Green Edition.

[4] Zheng, Y., Lin, Z., Tay, D.B.H. (2001). State-dependent vector hybrid linear and nonlinear ARMA modeling: theory. *Circuits Systems Signal Processing* **20**, 551–574.

Estimation and testing in the random coefficient dynamic panel data model

Remigijus Leipus

Vilnius University, Lithuania, remigijus.leipus@mif.vu.lt

Keywords: random-coefficient autoregression, empirical process, Kolmogorov-Smirnov statistic, goodness-of-fit testing, kernel density estimator

We discuss nonparametric estimation of the distribution function $G(x)$ of the autoregressive coefficient $a \in (-1, 1)$ from a panel of N random-coefficient AR(1) data, each of length n, by the empirical distribution of lag 1 sample autocorrelations of individual AR(1) processes. Consistency and asymptotic normality of the empirical distribution function and a class of kernel density estimators is established under some regularity conditions on $G(x)$ as N and n increase to infinity. The Kolmogorov-Smirnov goodness-of-fit test for simple and composite hypotheses of Beta distributed a is discussed. A simulation study for goodness-of-fit testing compares the finite-sample performance of our nonparametric estimator to the performance of its parametric analogue discussed in [1]. This research is done jointly with Vytautė Pilipauskaitė, Anne Philippe and Donatas Surgailis.

References

[1] Beran, J., Schützner, M., Ghosh, S. (2010). From short to long memory: Aggregation and estimation. *Computational Statistics and Data Analysis* **54**, 2432–2442.

Semiparametric density estimation for star-shaped distributions

Eckhard Liebscher[1] **and Wolf-Dieter Richter**[2]

[1] *University of Applied Sciences Merseburg, Germany, eckhard.liebscher@hs-merseburg.de*
[2] *University of Rostock, Germany, wolf-dieter.richter@uni-rostock.de*

Keywords: star-shaped distributions, norm contoured distributions, kernel density estimators

In the talk we discuss properties of semiparametric estimators for the density generator function, and for the density in a star-shaped distribution model. The approach under consideration modifies and generalizes that of an earlier paper [1] about elliptical distributions. The semiparametric procedure combines the flexibility of nonparametric estimators and the simple estimation and interpretation of parametric estimators. A parametric model is assumed for the density contours given by the star body. The parameters are estimated using a moment estimation method. The star generalized radius density is estimated nonparametrically by use of a kernel density estimator. Since the star generalized radius density is a univariate function, we avoid the disadvantages of nonparametric estimators in connection with the curse of dimensionality.

The density of the star-shaped distribution of the random vector X is given by

$$\varphi_{g,K,\mu}(x) = C(g,K)\ g\left(h_K(x-\mu)\right) \text{ for } x \in \mathbb{R}^d,$$

where $C(g,K)$ is the normalizing constant, g is the density generator function, K is the star body and h_K the corresponding Minkowski functional. The theory of star-shaped distributions is developed in [2]. We suppose that a formula for h_K is specified. If h_K involves some additional parameters, then estimators of them are to be provided. In the paper moment estimators are used. Let $\psi : [0,\infty) \to \mathbf{R}$ be a strictly increasing function. The semiparametric estimator can be calculated by

$$\hat{\varphi}_n(x) = C(g,K)\ \hat{g}_n\left(h_K(x-\hat{\mu}_n)\right) \text{ for } x \in \mathbb{R}^d,$$

where

$$\hat{g}_n(z) = z^{1-d}\ \psi'(z)\ \hat{\chi}_n(\psi(z)) \text{ for } z \in \mathbb{R},$$

and $\hat{\mu}_n$ is the average of all sample items. In the last formula $\hat{\chi}_n$ is a kernel estimator for the density of $\psi(h_K(X-\mu))$.

In the talk, results on convergence rates of the density estimator are presented. It turns out that, in the case where a neighbourhood of the center μ is excluded, these rates coincide with the rates known from usual one-dimensional kernel density estimators. The behaviour of the estimator in the center of the distribution is discussed, too. We show that the density estimator is asymptotically normally distributed.

References

[1] Liebscher, E. (2005). A semiparametric density estimator based on elliptical distributions. *J. Multivariate Analysis* **92**, 205–225.

[2] Richter, W.-D. (2014). Geometric disintegration and star-shaped distributions. *Journal of Statistical Distributions and Applications* **1**:20.

Using auxiliary information in data collection and in estimation

Kaur Lumiste

University of Tartu, Estonia, kaur.lumiste@ut.ee

Keywords: non-response, responsive design, balanced response

High nonresponse is forcing national statistics offices around the world to find new ways of designing and controlling the data collection in their surveys. Responsive design is a newly emerged view focusing on the possibilities to reduce the effects of nonresponse by monitoring the data collection process. The main survey goodness measure thus far – nonresponse rate – is by itself an insufficient guide, since ordinarily collected sets of respondents tend to be biased towards certain society's groups like elderly or people living in rural areas. So more informative measures of the progress of the data collection have recently been proposed like balance indicators ([1]) and R-indicators ([2]) (where R stands for representativeness). In current work we use the first approach and aspire to representativeness through balance of the response set with respect to a given set of auxiliary variables - means of auxiliary variables have to be approximately the same in the sample and the response set.

The same auxiliary variables can also be used at the estimation stage to improve our estimates, but assume that we have access to more auxiliary variables at the estimation stage than we did in the data collection stage. Is the effect of additional explanation power affected by balancing? Finding an answer to this question brings another one - should we emphasise on acquiring more auxiliary variables for the estimation stage or should we focus more on balancing the response? Which would have a larger effect on the bias and/or accuracy of the final estimates? Applying both strategies can be very costly, so budgets can be optimised by opting for the better strategy.

In this paper we look for evidence to support one of the following strategies:

- put effort into collecting more auxiliary information and focus on post-weighting correction;
- put effort into monitoring response to get a representative set of respondents.

Interactions of response propensities, post-weighting weights, auxiliary variables and estimated variable are studied in the context of these strategies.

References

[1] Särndal, C.-E. (2011). The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics* **27**, 1–21.

[2] Schouten, B., Cobben, F., Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology* **35**, 101–113.

Statistical methods for cost-effectiveness analysis: A selected review

Thomas Mathew

University of Maryland Baltimore County, USA, mathew@umbc.edu

Keywords: average cost-effectiveness ratio (ACER), cost-effectiveness probability (CEP), incremental cost-effectiveness ratio (ICER), incremental net benefit (INB), willingness-to-pay parameter

Identifying treatments or interventions that are cost-effective (more effective at lower cost) is clearly important in health policy decision making, especially in the allocation of health care resources. Various measures of cost-effectiveness that are informative, intuitive and simple to explain have been suggested in the literature, along with statistical inference concerning them. Popular and widely used measures include the incremental cost-effectiveness ratio (ICER), defined as the ratio between the difference of expected costs and the difference of expected effectiveness in two populations receiving two treatments. Although very easy to interpret as the additional cost per unit of effectiveness gained, being a ratio, the ICER presents difficulties regarding interpretation in certain situations, for example, when the difference in effectiveness is close to zero, and it also presents challenges in the statistical inference. Yet another measure proposed in the literature is the incremental net benefit (INB), which is the difference between the incremental cost and the incremental effectiveness after multiplying the latter with a "willingness-to-pay parameter". Both ICER and INB are functions of population means, and inference concerning them has been widely investigated under a bivariate normal distribution, or under a log-normal/normal distribution for the cost and effectiveness measures. In the talk, we will briefly review these, focusing on recent developments. An alternative probability-based approach will also be introduced, referred to as cost-effectiveness probability (CEP), which is the probability that the first treatment will be less costly and more effective compared to the second one. Inference on the CEP will also be discussed. Numerical results and illustrative examples will be given.

The influence of model selection on selected measures of fit

Märt Möls

University of Tartu, Estonia, martm@ut.ee

Keywords: model selection, mixed models

Selecting models based on AIC, covariate statistical significance or by some other model selection criterion can introduce bias into statistics usually used to describe the model fit (and can also lead to biased estimates of effect sizes). A typical remedy of the problem is to use a separate validation dataset to get unbiased estimates of model fit characteristics. But sometimes it is not practical to divide the original dataset into model selection and validation datasets. Therefore some theoretical alternatives to decrease the bias introduced by the model selection process may still be of interest.

How to use available but biased (due to model selection) statistics to derive still meaningful (bias-reduced) estimates of the quantities of interest? A mixed model for truncated response variable is introduced and used to estimate some quantities of interest to reduce the model selection bias. The proposed mixed model behaviour in some practical applications will be evaluated.

Comparing dissimilarity measures: A case of banking ratios

Laurynas Naruševičius and Alfredas Račkauskas

Vilnius University, Lithuania, laurynas.narusevicius@mif.vu.lt, alfredas.rackauskas@mif.vu.lt

Keywords: banking ratios, dissimilarity, time series clustering, functional data clustering, clustering comparison

After the Global financial crisis in 2007-2008 financial sector and especially banks gained much attention. A large number of the macro-level instruments were introduced and those are applied to all banks. However, banking sector is heterogeneous and some tools could be ineffective to some banks. It would be useful to find groups of banks which have similar characteristics and design or calibrate some macroprudential instruments that would become appropriate for that group. Therefore, our first goal is to discuss a clustering of the banks. Our second goal is to consider various dissimilarity measures and apply them to a data under investigation. We exploited distance measures based on time series as well as on functional data properties. In addition to univariate clustering, where banks are grouped into clusters according to one bank-specific ratio, we applied multivariate clustering, where banks are clustered based on their several ratios.

In our study we used six ratios that reflect banks' profitability (return on average assets, return on average equity, net interest margin), efficiency (cost to income), stability (capital adequacy ratio) and portfolio credit risk (loan losses over loan portfolio). We applied twelve different dissimilarity measures. Ten of these measures are commonly used. We proposed two new distance measures, based on functional data properties, that, to our knowledge, were not used in the clustering literature. Furthermore, we extended two univariate distance measures to multivariate case. The results of the univariate clustering show that there is no dissimilarity measure, which would be the best to all ratios. However, in many cases clustering methods based on functional data properties outperfomed distance measures based on time series properties. The results of the multivariate clustering revealed that it is important to take into account not only how close banks' ratios are, but how similarly they change over the years.

Gerber-Shiu discounted penalty function for the bi-seasonal discrete time risk model

Olga Navickienė and Jonas Šiaulys

Vilnius University, Lithuania, olga.navickiene@mif.vu.lt, jonas.siaulys@mif.vu.lt

Keywords: bi-seasonal discrete time risk model, Gerber-Shiu discounted penalty function, ruin probability

The bi-seasonal discrete time risk model for insurer's surplus (property) changing over the time $n \in \mathbb{N}_0 := 0, 1, 2, \ldots$ is described by the following equation:

$$U_u(n) = u + n - \sum_{i=1}^{n} Z_i,$$

where $u = U_u(0) \in \mathbb{N}_0$, is the insurer's (insurance company's) initial surplus; Z_i, $i \in N$ are claim amounts which assumed to be independent nonnegative integer-valued random variables satisfying the following conditions:

$$Z_{2k+1} \stackrel{d}{=} Z_1, \ Z_{2k+2} \stackrel{d}{=} Z_2, \ k = 0, 1, 2, \ldots.$$

If $Z_1 = Z_2$, then the bi-seasonal discrete time risk model reduces to the classical discrete time risk model.

The finite time ruin probability $\psi(u, t)$, the ruin probability $\psi(u)$ and Gerber-Shiu discounted penalty function $\Psi_\delta(u)$ are the main extremal characteristics for both models.

Gerber-Shiu discounted penalty function for the model is defined by equality:

$$\Psi_\delta(u) = \mathbb{E}^{-\delta T_u} \mathbb{1}_{\{T_u < \infty\}},$$

where $\delta \geqslant 0$, $u \in \mathbb{N}_0, T_u$ – the time of ruin, i.e.

$$T_u = \begin{cases} \min \ \{n \geqslant 1 : U_u(n) \leqslant 0\}, \\ \infty, \text{ if } U_u(n) > 0 \text{ for all } n = 1, 2, 3, \ldots. \end{cases}$$

Furthermore

$$\psi(u) = \Psi_0(u) = \mathbb{E}\mathbb{1}_{\{T_u < \infty\}} = \mathbb{P}\{T_u < \infty\}, \psi(u, t) = \mathbb{P}\{T_u \leqslant \infty\}.$$

The formula for calculations of the finite time ruin probability $\psi(u, t)$, the ruin probability $\psi(u)$ and Gerber-Shiu discounted penalty function $\Psi_\delta(u)$ for the classical discrete time risk model are presented in [3].

The formula for calculation of $\psi(u, t)$ is given in the article [1].

The formula for calculation of $\psi(u)$ is presented in [2].

The similar formula for the bi-seasonal discrete time risk model and for Gerber-Shiu discounted penalty function $\Psi_\delta(u)$ can be obtained.

References

[1] Blaževičius, K., Bieliauskienė, E., Šiaulys, J. (2010). Finite-time ruin probability in the inhomogeneous claim case. *Lithuanian Mathematical Journal* **50**, 260–270.

[2] Damarackas, J. and Šiaulys, J. (2014). Bi-seasonal discrete time risk model. *Applied Mathematics and Computation* **247**, 930–940.

[3] Dickson, D.C.M. (2010). *Insurance Risk and Ruin*. Cambridge University Press, Cambridge.

Tests based on characterizations, and their efficiencies

Ya. Yu. Nikitin

Saint-Petersburg University, Russia, y.nikitin@spbu.ru

Keywords: goodness-of-fit, asymptotic efficiency, characterizations

Suppose we have a sample $X_1, \dots, X_n$ of i.i.d. observations with distribution function (df) F, and we are testing the composite hypothesis $H_0 : F \in \mathcal{F}$, where $\mathcal{F}$ is some family of distributions, against the alternative $H_1 : F \notin \mathcal{F}$. Often the class $\mathcal{F}$ is characterized by the same distribution of two statistics $g_1(X_1, \dots, X_r)$ and $g_2(X_1, \dots, X_s)$.

Let us introduce two U-empirical df's

$$H_n^1(t) = \binom{n}{r}^{-1} \sum_{1 \le i_1 < \dots < i_r \le n} \mathbf{1}\{g_1(X_{i_1}, \dots, X_{i_r}) < t\}, \quad t \in R^1,$$

$$H_n^2(t) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < \dots < i_s \le n} \mathbf{1}\{g_2(X_{i_1}, \dots, X_{i_s}) < t\}, \quad t \in R^1.$$

According to Glivenko-Cantelli theorem for U-empirical df's, $H_n^1(t)$ and $H_n^1(t)$ become very close under H_0 as $n \to \infty$. Consequently we can build the integral and Kolmogorov type goodness-of-fit tests based on the difference of $H_n^1(t)$ and $H_n^1(t)$.

Consider as an example of this idea the famous Polya's characterization [1]: *If X and Y are i.i.d. centered random variables (rv's), then the relation $X \stackrel{d}{=} (X+Y)/\sqrt{2}$ is valid iff X and Y are normal.* Another example is the Shepp's characterization [2] of normality with zero mean by the equal distribution of rv's X and *nonlinear* statistic $2XY/\sqrt{X^2+Y^2}$. We can construct the $U-$empirical goodness-of-fit statistics, then study their limiting distributions, and calculate their Pitman and Bahadur efficiencies. There are plenty of characterizations for the exponential, Pareto, Cauchy, arcsine, logistic, and other laws, and one can build and study corresponding tests as well. Some of them turn out to be rather efficient.

Goodness-of-fit tests based on characterizations of distributions were proposed by Yu. V. Linnik [3]. Due to technical difficulties, Linnik's idea was implemented only in last 10-15 years when the theory of $U-$empirical measures was sufficiently elaborated. We present a survey of recent results on goodness-of-fit and symmetry testing obtained by two small groups of researchers in Saint-Petersburg and Belgrade.

The author was supported by the grant of RFBR No. 16-01-00258.

References

[1] Polya, G. (1923). Herleitung des Gauss'schen Fehlergesetzes aus einer Funktionalsgleichung. *Math. Zeitschr.* **18**, 96–108.

[2] Shepp, L. (1964). Normal functions of normal random variables. *SIAM Rev.* **6**, 459–460.

[3] Linnik, Yu.V. (1953). Linear forms and statistical criteria I, II. *Ukrainian Math. J.*, **5**:2, 207–243; **5**:3, 247–290.

Third and fourth cumulants in search of independent components

Hannu Oja, Klaus Nordhausen and Joni Virta

University of Turku, Finland, hannu.oja@utu.fi

Keywords: FastICA, independent component analysis, kurtosis, non-gaussian component analysis, skewness

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. In this talk projection pursuit is used to extract the non-Gaussian components and to separate the corresponding signal (non-gaussian) and noise (gaussian) subspaces. For early ideas on projection pursuit, see e.g. [1]. Our choice for the projection index is a convex combination of squared third and fourth cumulants and we estimate the non-Gaussian components either (i) one-by-one (deflation-based approach) or (ii) simultaneously (symmetric approach). The properties of the unmixing matrix estimates are considered in detail through the corresponding optimization problems, estimating equations, algorithms and asymptotic properties. The talk is based on the papers [2, 3].

References

[1] Huber, P.J. (1985). Projection pursuit. *Annals of Statistics* **13**, 435–475.

[2] Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science* **30**, 372–390.

[3] Virta, J., Nordhausen, K., Oja, H. (2016). Projection pursuit for non-Gaussian independent components. *Submitted*.

Triplet Markov models

Wojciech Pieczynski

Telecom Sudparis, France, Wojciech.Pieczynski@telecom-sudparis.eu

Keywords: triplet Markov models, signal segmentation, parameter estimation, optimal filtering, theory of evidence

Hidden Markov models (HMMs) are widely applied in various problems occurring in different areas like biosciences, climatology, communications, ecology, econometrics and finances, or still image or signal processing. In such models, the hidden process of interest X is a Markov chain, which must be estimated from an observable one Y, interpretable as being a noisy version of X. The success of HMC is mainly due to the fact that the conditional probability distribution of the hidden process with respect to the observed process remains Markov, which makes possible different processing strategies such as Bayesian restoration. HMMs have been recently generalized to "Pairwise" Markov models (PMMs) and "Triplet" Markov models (TMMs), which offer similar processing advantages and superior modeling capabilities. In PMMs, one directly assumes the Markovianity of the pair (X, Y) and in TMMs, the distribution of the pair (X, Y) is the marginal distribution of a Markov process (X, U, Y), where U is an auxiliary process, possibly contrived.

For hidden discrete process, TMMs extend known models like hidden semi-Markov models or hidden bivariate models, or still Dempster-Shafer theory of evidence based hidden models. Parameter estimation methods are available leading to unsupervised Bayesian processing.

For hidden continuous process, particular TMMs allow one to consider fast optimal filtering in switching systems. Such systems are of particular interest as they allow approximating any non-linear Markov stationary system.

Applying non-parametric methods on estimation of medical bills pricing limits

Margus Pihlak

Tallinn University of Technology, Estonia, margus.pihlak@ttu.ee

Keywords: non-parametric bootstrap, parametric bootstrap, bias corrected and accelerated method

The aim of this talk is to generalize confidence interval calculations. In classical cases we apply on these calculations central limit theorem, Student t-distribution or χ^2-distribution. Often, however, we meet data where assumptions of classical methods are not met. Such kind of problems exist on data of health care services, for example. In this situation alternative methods of confidence interval calculation have to be found. Basic ideas of these methods are described in the book [2] and in the paper [1], for example.

Firstly we will present Estonian health care pricing calculation methodology. This methodology is based on assumptions of classical statistics. Then new ideas of maximum and minimum prices of health care services calculations will be given. Our idea concerns application of parametric and non-parametric bootstrap methods on confidence interval calculations. Also bias corrected and accelerated method will be demonstrated.

References

[1] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

[2] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Linear sufficiency in linear regression model

Simo Puntanen

University of Tampere, Finland, simo.puntanen@uta.fi

Keywords: best linear unbiased estimator, estimability, generalized inverse, linear model, transformed linear model

A linear statistic Fy, where F is an $f \times n$ matrix, is called linearly sufficient for the Xb under the model $M = \{y, Xb, V\}$, if there exists a matrix A such that AFy is the BLUE for Xb. In this talk we review some specific aspects of the linear sufficiency and consider the relations of BLUE of the estimable parametric functions under the linear model M and its transformed version $\{Fy, FXb, FVF'\}$.

References

[1] Baksalary, J.K., Kala, R. (1981). Linear transformations preserving best linear unbiased estimators in a general Gauss-Markoff model. *Annals of Statistics* **9**, 913–916.

[2] Baksalary, J.K., Kala, R. (1986). Linear sufficiency with respect to a given vector of parametric functions. *Journal of Statistical Planning and Inference* **14**, 331–338.

[3] Drygas, H. (1983). Sufficiency and completeness in the general Gauss–Markov model. *Sankhyā, Ser. A* **45**, 88–98.

[4] Kala, R., Puntanen, S., Tian, Y. (2015). Some notes on linear sufficiency. *Statistical Papers*, available online.

Asymptotically optimal placement of initial centers in k-means clustering

Kalev Pärna

University of Tartu, Estonia, kalev.parna@ut.ee

Keywords: k-means, Lloyd's algorithm, quantization, asymptotic distribution of k-means

We are discussing efficient ways of converting data into a compact discrete form – a problem that arises in many areas. In information theory such a conversion is called quantization and it is used to transmit the data through a discrete channel which allows k different values only. In statistics and data mining, k-means clustering aims at partitioning the data set into k non-overlapping clusters by minimizing the within-sum of squares of deviations from their respective cluster centers (k-means). Efficient calculation of k-means, especially in multivariate setting, is still a problem which needs further research. Lloyd's iterative method [2] – a standard procedure for calculation of k-means – is sensitive with respect to initial centers and, therefore, different methods have been proposed for the choice of initial seed values of k-means [1].

In this paper we focus on how to make use of certain theoretical results about asymptotic behavior of optimal centers if k tends to infinity. It is known that, for large k, the optimal centers are distributed in accordance with the density $f^\star(x)$, which is a power function of the initial data density $f(x)$ [3]. In 1-dimensional case, for example, the asymptotic density of k-means is proportional to $[f(x)]^{1/3}$. In order to benefit from this asymptotics, we propose to use the density $f^\star(x)$ for placement of initial seeds in the Lloyd's iterative algorithm. Our method consists of 1) estimation of $f^\star(x)$ from the data, 2) using $f^\star(x)$ for reweighting initial data (change of measure), 3) sampling k points from reweighted data, and 4) using the points sampled as initial values for further iterations in Lloyd's algorithm.

References

[1] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. *Proceedings of The 18-th Annual ACM-SIAM Symposium on Discrete Algorithms.* SIAM, 1027–1035.

[2] Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* **IT-28**, 129–137.

[3] Zador, P. (1982). Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory* **IT-28**, 139–149.

Tail probabilities of likelihood ratio statistic

Marijus Radavičius

Vilnius University, Lithuania, marijus.radavicius@mii.vu.lt

Keywords: Hoeffding inequality, multinomial distribution, likelihood ratio statistic, probabilities of large deviations

Let $\mathbf{y} = (y_1, \dots, y_n)$ be a random vector having the multinomial distribution

$$\mathbf{y} \sim \text{Multinomial}_n(N, \mathbf{p}),$$

where $\mathbf{p} = (p_1, \dots, p_n)$ is a vector of positive probabilities.

Define (logarithmic) likelihood ratio statistic

$$G_n^2(\mathbf{y}; N, \mathbf{p}) := \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{p_i N}\right).$$

Assuming that $p_{min} := \min_{i=1,\dots,n} p_i \geq \delta_0 > 0$, Hoeffding [1] proved that

$$\mathbf{P}\{G_n^2(\mathbf{y}; N, \mathbf{p}) \geq x\} = O\left(x^{(n-3)/2}\mathrm{e}^{-x}\right), \quad N \to \infty,$$

uniformly in $x \in [c_1, c_2 N]$ for arbitrary positive constants c_1 and c_2. Kallenberg [2] obtained upper and lower bounds for the tail probabilities of $G_n^2(\mathbf{y}; N, \mathbf{p})$ in the case where $p_{min} \to 0$ and $n \to \infty$ not too fast. However, the upper bound exceeds the corresponding lower bound by a factor of order $\sqrt{x}$.

The *problem* is to obtain, for the tail probabilities of $G_n^2(\mathbf{y}; N, \mathbf{p})$, an exact (up to a constant factor) upper bound valid for all positive x.

For $n = 2$, the universal and exact up to a factor of 2 upper bound for the tail probabilities follows from the paper by Zubkov and Serov [3]. The generalization of this bound to the case $n > 2$ is discussed.

References

[1] Hoeffding, W. (1967). On probabilities of large deviations. In: *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **I** (L. LeCam and J. Neyman, eds.) 203–219. Univ. California Press, Berkeley.

[2] Kallenberg, W.C.M. (1985). On moderate and large deviations in multinomial distributions. *Annals of Statistics* **13**, 1554–1580.

[3] Zubkov, A.M. and Serov, A.A. (2012). A complete proof of universal inequalities for distribution function of binomial law. *Teoriya Veroyatnostei i Ee Primeneniya* **57**, 597–602.

Optimal employee behavior and optimal sickness insurance design when employers penalize sickness presenteeism

Colin M. Ramsay[1], Victor I. Oguledo[2] and Annika Krutto[3]

[1]*University of Nebraska-Lincoln, USA, cramsay1@unl.edu*
[2]*Florida A&M University, USA, Victor.Oguledo@famu.edu*
[3]*University of Tartu, Estonia, annika.krutto@ut.ee*

Keywords: presenteeism penalty, absenteeism, shirking, utility, employee strategy/behavior, health shock, multi-state sickness model, Volterra integral equations, renewal equations

In a recent paper [1], Ramsay and Oguledo considered the optimal design of an employer sponsored sickness-disability insurance plan that maximizes the employer's expected discounted profits over each employee's working lifetime. They used a simple multi-state model of the evolution of an employee's health over time to capture the impact of sickness induced absenteeism, presenteeism, and shirking on the employer's profits. They also assumed sick employees were asymptomatic, i.e., they showed no signs of illness at work, and thus were able to work regardless of their level of illness. In this paper, we extend the Ramsay-Oguledo model by introducing a new 'severely ill' sickness state where employees are symptomatic, i.e., they show signs of illness while at work, and presenteeism exists. To combat presenteeism, we also introduce the new concept of a presenteeism penalty whereby employers penalize employees who are found to be very sick while at work. Specifically, penalized employees are sent home and receive a penalized sick-pay that is lower that the normal sick pay. Thus sick employees must decide whether to stay at home and receive a sick pay (that is less than their working pay) or go to work sick and run the risk of being sent home and penalized. Assuming employees get a positive utility from income and a disutility (negative utility) from work, we determine each employee's optimal behavior/strategy in each sickness state (i.e., whether to stay home or to work) that maximizes her discounted expected utility over her working lifetime. As in [1], permissible employee strategies are captured in a set of Volterra integral equations that are solved numerically to determine each employee's optimal strategy/behavior. Assuming employees are expected utility maximizers, we determine the employer's optimal sick pay, presenteeism penalty, and health check probabilities that maximize the employer's discounted expected profits over an employee's working lifetime.

References

[1] Ramsay, C.M., Oguledo, V.I. (2015). Optimal disability insurance with moral hazards: absenteeism, presenteeism, and shirking. *North American Actuarial Journal* **19**, 143–173.

Pattern recognition using hidden Markov models in financial time series

Sara Rebagliati and Emanuela Sasso

Università degli Studi di Genova, Italy, rebagliati@dima.unige.it, sasso@dima.unige.it

Keywords: pattern recognition, hidden Markov models, financial time series, automated trading system

Automated trading systems spread in the last years due to the advances in technologies. Before, traders usually took their investment decisions looking at the prices graphs and recognizing some trading patterns. Since our aim consists in replicating the behavior of a discretionary trader, we have developed a software that can recognize trading patterns in real time using hidden Markov models (HMMs). Financial time series are strongly affected by noise. Moreover, trading patterns are defined by their shape and can occur at different time scales and have different amplitudes. Since the problem is quite similar to speech recognition, we used hidden Markov models to develop our software (see [2] for application of HMMs to speech recognition).
Let us suppose we want to recognize N trading patterns. We trained N HMMs using Baum-Welch Algorithm combined with Genetic Algorithm. Extending the idea shown in [1] to continuous observations HMMs, we trained a threshold model which can recognize all the not predefined patterns. Then, we implemented the classification algorithm to work in real time. Since a trader needs to be fast to enter the market, we have finally modified the algorithm to recognize forecasted scenario before they happen, as our brain does.

References

[1] Lee, H.-K., Kim, J.H. (1999). An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**, 961–973.

[2] Rabiner, R. (1989). A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

Norm and antinorm contoured distributions

Wolf-Dieter Richter

University of Rostock, Germany, wolf-dieter.richter@uni-rostock.de

Keywords: density level set, convex ball, radially concave ball, non-Euclidean surface content measure, non-Euclidean uniform distribution, stochastic vector representation, geometric measure representation, exact statistical distribution

For a long period, exact statistical distribution laws of functions of random vectors were mainly derived in cases when sample vectors follow Gaussian or elliptically contoured distributions, see [1, 2].

A new geometric method for deriving exact distributions of statistics if the sample vector follows an $l_{n,p}$-symmetric distribution law was discussed in [4] and applied later on to several statistical distributions such as that of extreme values, general order statistics, ratios and products as just to mention a few of them.

An extension of this disintegration method, which one can also consider as a generalization of the method of indivisibles, to norm and antinorm contoured sample distributions, is proved in [5]. For the notion of antinorm, see [3].

Corresponding stochastic vector representations generalize the well known one from [2].

In this talk, we overview recent developments of geometrically representing multivariate random vectors and their distributions, present several particular cases and applications, and outline further perspectives.

References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Analysis*. Wiley, New York.

[2] Fang, K.-T., Kotz, S., Ng, K.-W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

[3] Moszyńska, M., Richter, W.-D. (2012). Reverse triangle inequality. Antinorms and semi-antinorms. *Studia Scient. Mathem. Hungar.* **49**, 120–138.

[4] Richter, W.-D. (2012). Exact distributions under non-standard model assumptions. *AIP Conf. Proc.* **1479**, 442; doi: 10.1063/1.4756160.

[5] Richter, W.-D. (2015). Convex and radially concave contoured distributions, *Journal of Probability and Statistics*, Article ID 165468, 12 pages.

Influential observations in multivariate linear models

Dietrich von Rosen[1,2]

[1]*Swedish University of Agricultural Sciences, Sweden, Dietrich.von.Rosen@slu.se*
[2]*Linköping University, Sweden*

Keywords: growth curve model, maximum likelihood estimator, perturbation scheme

A general approach to identify influential observations in multivariate linear models is presented. The main idea is to perturb models which then are evaluated via Taylor expansions. In particular bilinear models such as the growth curve model and its extensions are discussed.

On reduced rank regression analysis in GMANOVA-MANOVA models

Tatjana von Rosen[1] **and Dietrich von Rosen**[2,3]

[1]*Stockholm University, Sweden, tatjana.vonrosen@stat.su.se*
[2]*Swedish University of Agricultural Sciences, Sweden, Dietrich.von.Rosen@slu.se*
[3]*Linköping University, Sweden*

Keywords: growth curve model, maximum likelihood estimator, multivariate analysis of variance, reduced rank model

We discuss the maximum likelihood estimation in general multivariate linear models where the rank restrictions are imposed on the matrix of regression coefficients in order to enable parsimonious modeling. In particular, the following GMANOVA-MANOVA model is of interest:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}_1\boldsymbol{C}_1 + \boldsymbol{B}_2\boldsymbol{C}_2 + \boldsymbol{E},$$

where the columns of $\boldsymbol{X}$: $p \times n$ are p-dimensional independent normally distributed response vectors, $\boldsymbol{A}$:$p \times q_1$, is a known within-individuals design matrix, $\boldsymbol{C}_1$: $k_1 \times n$ is a known between-individuals design matrix, $\boldsymbol{B}_1$:$q_1 \times k_1$ is an unknown parameter matrix that summarizes the mean structure of the response variables over time, $\boldsymbol{B}_2$: $p \times k_2$ describes the mean structure, $\boldsymbol{C}_2$: $k_2 \times n$ is a known between-individuals design matrix, the columns of $\boldsymbol{E}$: $p \times n$ are independent normally distributed random error vectors, $\boldsymbol{E}_k \sim N_p(0, \Sigma)$, where $k = 1, \ldots, n$, and Σ is an unknown positive definite matrix.

This model being the mixture of the growth curve (GMANOVA) and the multivariate analysis of variance (MANOVA) models extends the classical growth curve model to include the covariates. The GMANOVA-MANOVA model was proposed by Chinchilli & Elswick (1985) and von Rosen (1989), and has been applied in a variety of disciplines, including economics, biology, medicine, engineering and life sciences.

Several modifications of the GMANOVA-MANOVA model will be studied depending on the rank restrictions which are characterized by the following conditions:
(i) $r(\boldsymbol{B}_2) = q_2 < \min(p, k_2)$, $q_1 < p$, $C(\boldsymbol{C}_2')C(\boldsymbol{C}_1')$;
(ii) $r(\boldsymbol{B}_2) = q_2 < \min(p, k_2)$, $\boldsymbol{A} = \boldsymbol{I}_p$;
(iii) $r(\boldsymbol{B}_2) = q_2 < \min(p, k_2)$, $q_1 < p$, $C(\boldsymbol{C}_1') \subseteq C(\boldsymbol{C}_2')$;
(iv) $r(\boldsymbol{B}_1) = f < \min(q_1, k_1)$, $r(\boldsymbol{B}_2) = q_2 < \min(p, k_2)$, $C(\boldsymbol{C}_2') \subseteq C(\boldsymbol{C}_1')$,
where $C(\bullet)$ denotes the column vector space.

References

[1] Chinchilli, V.M. and Elswick, R.K. (1985). A mixture of the MANOVA and GMANOVA models. *Comm. Statist. Theory Methods* **14**, 3075–3089.

[2] von Rosen, D. (1989). Maximum likelihood estimators in multivariate linear normal models. *Journal of Multivariate Analysis* **31**, 187–200.

Modeling of replicated response measures by using interval arithmetic

Busenur Sarıca[1], Özlem Türkşen[2] and Cengiz Kahraman[3]

[1]*Marmara University, Turkey, busenur.sarica@marmara.edu.tr*
[2]*Ankara University, Turkey, turksen@ankara.edu.tr*
[3]*Istanbul Technical University, Turkey, kahramanc@itu.edu.tr*

Keywords: replicated response measures, fuzzy numbers, α-cuts, interval arithmetic, least squares, NSGA–II

In experimental design, some data sets are composed with the replicated response measures. The replicated response measures are assumed to have uncertainty in the quantification of their values. In this case, it may be suitable to represent these values by using fuzzy numbers. The fuzzy numbers are an extension of real numbers which allow the incorporation of uncertainty [1].

In this study, the replicated measures of the responses are represented as triangular type-1 fuzzy numbers similar to [2]. In addition, trapezoidal and Gaussian type-1 fuzzy numbers are used for comparison purposes. The novelty of this study is the formalization of these fuzzy numbers as intervals. In order to convert the fuzzy numbers to intervals, α-cut values of fuzzy numbers are used. Here, α is chosen equal to 0 ($\alpha = 0$) for the simplicity. It should be noted here that the responses and model parameters are assumed to be interval values whereas inputs are crisp. Interval arithmetic is used to estimate the unknown model parameters in which the least squares criterion is applied. Since the interval matrix multiplication is a major time consuming process, midpoint and radius values are used for calculations instead of min-max operators as in [3], [4]. Then the objective function is expressed as an interval where the upper and lower bounds need to be minimized simultaneously. By considering each bound as an objective function, the estimation problem is solved by using Nondominated-Sorting Genetic Algorithm-II (NSGA–II) which is a well-known multi objective optimization method [5]. In the application part, a well-known data set, the wheel cover component data set [6] is used. Finally, the conclusions and future research directions are given in the conclusion section.

References

[1] Ban, A., Coroianu, L., Khastan, A. (2016). Conditioned weighted L-R approximations of fuzzy numbers. *Fuzzy Sets and Systems* **283**, 56–82.

[2] Türkşen, Ö. and Güler, N. (2015). Comparison of fuzzy logic based models for the multi-response surface problems with replicated response measures. *Applied Soft Computing* **37**, 887–896.

[3] Rump, S. (2012). Fast interval matrix multiplication. *Numerical Algorithms* **61**, 1–34.

[4] Ozaki, K., Ogita, T., Rump, S., Oishi, S. (2012). Fast algorithms for floating-point interval matrix multiplication. *Journal of Computational and Applied Mathematics* **236**, 1795–1814.

[5] Deb, K. (2004). *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley and Sons, New York.

[6] Harper, D., Kosbe, M., Peyton, L. (1987). Optimization of Ford Taurus wheel cover balance. *5th Symposium on Taguchi Methods*, 527-539.

Comparison of different approaches for micro panel data clustering

Lukas Sobisek[1] **and Maria Stachova**[2]

[1] *University of Economics, Prague, Czech Republic, lukas.sobisek@vse.cz*
[2] *Matej Bel Universtity, Banska Bystrica, Slovakia, maria.stachova@umb.sk*

Keywords: micro panel data, *kml* package, characteristics-based clustering

Nowadays, the micro panel data analysis is being often used to recognize a pattern of an individual change in time in many research areas such as finance or medicine. The clustering methods are used to identify groups so that individuals who belong to the same cluster are the most similar in shape to their time trajectory. The main goal of our contribution is to compare two different approaches that can be used for micro panel data clustering, particularly a characteristics-based clustering and a direct clustering of the raw-data. The former approach doesn't work directly with raw data but it rather clusters individuals using extracted characteristics from point values. The latter approach is applied in the R system package *kml*. In this paper a performance of selected clustering approaches is compared using a simulation study where the balanced clusters' sizes are expected.

The existence of infinite Viterbi alignment for PMC models

Joonas Sova and Jüri Lember

University of Tartu, Estonia, joonas.sova@ut.ee, juri.lember@ut.ee

Keywords: pairwise Markov chain, hidden Markov chain, Viterbi alignment, Viterbi algorithm, Viterbi training, maximum a posteriori path

We consider a two-dimensional homogeneous Markov chain $((X_1, Y_1), (X_2, Y_2), ...)$, where random variables X_i (observations) are taking values from some set $\mathcal{X}$ and random variables Y_i (unobserved or "hidden" states) are taking values from state space $\mathcal{Y} = \{1, ..., |\mathcal{Y}|\}$. Following Pieczynski [1] we call this model a pairwise Markov chain (PMC). The name reflects the fact that conditionally, given the marginal process $X = (X_1, X_2, ...)$, the process $Y = (Y_1, Y_2, ...)$ is a Markov chain, and conditionally, given Y, X is a Markov chain. In general though, neither X nor Y are necessarily Markov chains. PMC is a natural generalization of hidden Markov model (HMM). In particular, just like in case of HMM, given a realization $x_{1:n} = (x_1, \ldots, x_n)$ of $X_{1:n} = (X_1, \ldots, X_n)$, the Viterbi algorithm can be employed to find the maximum a posteriori (MAP) estimate $(v_1(x_{1:n}), ..., v_n(x_{1:n}))$ of $Y_{1:n}$. This estimate is also called the Viterbi alignment or Viterbi path. We prove a theorem which gives sufficient conditions for extending the Viterbi alignment to infinity. We further provide examples where we apply this theorem to specific models.

References

[1] Pieczynski, W. (2003). Pairwise Markov chains. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 634–639.

Estimation of causal effects with unobserved confounding: an alternative to instrumental variables

Elena Stanghellini[1], Marco Doretti[1] and Sara Geneletti[2]

[1] *University of Perugia, Italy, elena.stanghellini@stat.unipg.it*
[2] *London School of Economics and Political Science, United Kingdom, s.geneletti@lse.ac.uk*

Keywords: causal effect, confounder, directed acyclic graph, identification, latent variable, regression graph, structural equation model

We consider the problem of identifying causal effects in the presence of unmeasured confounding and two auxiliary instruments that do not qualify as instrumental variables in the usual notion, but nevertheless allow to achieve identification of the effect of interest. Specifically, we compare the classical instrumental variables (IV) estimator to the estimator presented in [2] (KP); see also [3]. The assumptions underlying the KP estimator differ from those typically required by the IV method and can be encoded by directed acyclic graphs (DAGs). It can be shown that the KP estimator reduces to the IV when the two instruments are marginally independent. We evaluate the performances of IV and KP under a number of scenarios. Our simulations show that the KP estimator is generally less prone to misspecification bias than the IV estimator. By exploiting the result in [1], we extend the derivations to all regression graphs that are Markov equivalent to a DAG that satisfies the KP assumption, thereby enlarging the class of models to which the results are of use. The two estimators are applied to the Counterweight Programme Data to evaluate the effect of a binary (hard/soft) treatment on BMI reduction.

References

[1] Wermuth, N. and Sadeghi, K. (2011). Sequences of regressions and their independences. *TEST* **21**, 215–252.

[2] Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101**, 423–437.

[3] Stanghellini, E. and Pakpahan, E. (2015). Identification of causal effects in linear models: beyond instrumental variables. *TEST* **24**, 489–509.

Consumption, insurance and health decisions under life-time uncertainty

Mogens Steffensen

University of Copenhagen, Denmark, mogens@math.ku.dk

We formalize a global objective under separation of preferences for risk and intertemporal substitution. We discuss its connection with stochastic differential utility (time-continuous recursive utility) which is based on local separation. For a Merton market the optimal decisions with respect to consumption and investment coincide. We consider two more general markets and characterize the solutions for these markets. In one case we study an incomplete market by adding an extra state process. In another case, we study the effects from an uncertain lifetime and access to life insurance. The latter gives new insight in how, possibly, an endogenous demand for hump-shaped consumption can arise even with 'fair' pricing of insurance. Finally, we discuss briefly how frictions in the insurance market may, or may not, alter the conclusions.

Estimating compression of polar ice

Tuomas Rajala[1], Aila Särkkä[2], Claudia Redenbach[3] and Martina Sormani[3,4]

[1] *University College London, UK, t.rajala@ucl.ac.uk*
[2] *Chalmers University of Technology and University of Gothenburg, Sweden, aila@chalmers.se*
[3] *University of Kaiserslautern, Germany, redenbach@mathematik.uni-kl.de*
[4] *Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Germany, martina.sormani@itwm.fraunhofer.de*

Keywords: anisotropy, ellipsoid, non-parametric statistics, polar ice, spatial point process

Analysis of deep polar ice cores has become an important tool for deriving climate information from the past. Interpretation of ice core records requires an accurate dating of the ice. The recent dating relies on models where the key element is the simulation of the individual history of ice deformation for each specific core site. We present a two-stage non-parametric method for the estimation of the deformation history in polar ice using the measured anisotropy of air inclusions from deep ice cores. First, we fit ellipsoids to the pattern of point-to-point distance vectors to estimate the direction of anisotropy. Then, we estimate the scale of anisotropy by identifying the back-transformation resulting in the most isotropic pattern. Finally, the method is applied to estimate the compression in polar ice air bubble patterns.

A Lundberg-type inequality for an inhomogeneous renewal risk model

Jonas Šiaulys

Vilnius University, Lithuania, jonas.siaulys@mif.vu.lt

Keywords: renewal model, inhomogeneous model, Lundberg-type inequality, exponential bound, ruin probability

We say that the insurer's surplus $U(t)$ varies according to the inhomogeneous renewal risk model if

$$U(t) = x + ct - \sum_{i=1}^{\Theta(t)} Z_i, t \geqslant 0,$$

where $x \geqslant 0$ is the initial risk reserve; $c > 0$ is the constant premium rate; $\{Z_1, Z_2, ...\}$ are independent non-negative claim sizes; $\Theta(t)$ is the number of accidents in the interval $[0, t]$ given by formula

$$\Theta(t) = \sum_{n=1}^{\infty} \mathbf{I}_{\{\theta_1+\theta_2+...+\theta_n \leqslant t\}}$$

and $\{\theta_1, \theta_2, ...\}$ are non-negative and non-degenerate at zero random variables, standing for the inter-arrival times. In addition, we suppose that sequences $\{Z_1, Z_2, ...\}$ and $\{\theta_1, \theta_2, ...\}$ are mutually independent.

If all claim sizes $\{Z_1, Z_2, ...\}$ are identically distributed and all inter-arrival times $\{\theta_1, \theta_2, ...\}$ are also identically distributed, then the inhomogeneous renewal risk model becomes the homogeneous renewal risk model.

The time of ruin and the ruin probability are the main critical characteristics of the homogeneous and the inhomogeneous risk models. The first time τ when the surplus $U(t)$ drops to a level less than zero is called the time of ruin. In such a case, the ruin probability ψ is defined by equality $\psi(x) = \mathbb{P}(\tau = \infty)$, where x is the initial reserve which is supposed to be the main model parameter.

The Lundberg inequality is well known for the homogeneous renewal risk model. This inequality states that $\psi(x) \leqslant \mathrm{e}^{-Hx}$ for some positive H in the case when $\mathbb{E}Z_1 - c\mathbb{E}\theta_1 < 0$ and $\mathbb{E}\mathrm{e}^{hZ_1} < \infty$ for some positive h. The proofs of this statement can be found for instance in [2] or [3].

One can show that the similar exponential estimate of the ruin probability holds for an inhomogeneous risk renewal model. Naturally, we need for such estimate to use more complex requirements for random variables $\{Z_1, Z_2, ...\}$ and $\{\theta_1, \theta_2, ...\}$. The exact formulation of the Lundberg-type inequality for an inhomogeneous renewal risk model together with the detailed proof can be found in [1].

References

[1] Andrulytė, I.M., Bernackaitė, E., Kievinaitė, D., Šiaulys, J. (2015). A Lundberg-type inequality for an inhomogeneous renewal risk model. *Modern Stochastics: Theory and Applications* **2**, 173–184.

[2] Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*. World Scientific Publishing, New Jersey.

[3] Teugels, J. and Sund, B. (eds.) (2004). *Encyclopedia of Actuarial Science*. Wiley.

On comparison of stochastic reserving methods

Liivika Tee, Meelis Käärik and Rauno Viin

University of Tartu, Estonia, liivika.tee@ut.ee

Keywords: Chain Ladder, claims reserves, bootstrap

A non-life insurance company has to set up a fund to enable the company to meet and administer its contractual obligations to policyholders. There are several methods to predict or estimate the future reserves. Chain Ladder method is the most widely applied claim reserving method, but the method gives us no information about the variability of the estimation. We consider certain stochastic reserve estimation methods in the basis of generalized linear models to estimate the likely variability in the outcome. In addition, as the models are usually based on different assumptions, then in case the assumptions are not fulfilled, we also use the bootstrap method to approach the problem.

In claims reserving, the data is usually assumed to be independent, but not identically distributed since the means and also the variances depend on covariates. Therefore it is common to bootstrap residuals, rather than the data themselves, since the residuals are approximately independent and identically distributed or can be made so. Different possibilities that can be used before applying the bootstrap method to the data will be discussed. We consider several stochastic reserving models along with the bootstrap method with different types of residuals to carry out a practical implementation using real-life insurance data to estimate reserves and their prediction errors.

References

[1] England, P.D. and Verrall R.J. (1998). Standard errors of prediction in claims reserving: A comparison of methods. *General Insurance Convention and ASTIN Colloquium*, 460–478.

[2] Pinheiro, P.J.R., Andrade e Silva, J.M., Centeno, M.L. (2003). Bootstrap methodology in claim reserving. *The Journal of Risk and Insurance* **70**, 701–714.

Estimation and calibration of response probabilities

Natalja Lepik and Imbi Traat

University of Tartu, Estonia, natalja.lepik@ut.ee, imbi.traat@ut.ee

Keywords: sampling, nonresponse, response probability, backward calibration

Common feature of nowadays sample surveys is low response rate. As a result, adjustments have to be made in estimators to guarantee (nearly) unbiasedness. If the response probabilities θ_k were known then unbiased estimator for the population total $t = \sum_U y_k$ is

$$\hat{t} = \sum_r \frac{y_k}{\pi_k \theta_k},$$

where r is the response set and π_k the inclusion probability of unit k. A review on nonresponse weighting adjustments is given in [1].

We concentrate on the backward calibration property of estimates $\hat{\theta}_k$. Broadly speaking, this property forces $\hat{\theta}_k$ to be close to the sample-based response proportions, e.g. in the groups of units with given auxiliary vector values. If the true response probability is constant in that group then the response proportion is its unbiased estimate. We show that the estimates $\hat{\theta}_k$ resulting from the regression and logistic regression modeling obey the backward calibration property.

Nowadays, there are many new statistical learning methods for classification that could be used for prediction of response probabilities [2]. We compare $\hat{\theta}_k$ of different estimation methods and study the backward calibration property of these methods.

References

[1] Brick, J.M. (2013). Unit nonresponse and weighting adjustments: a critical review. *Journal of Official Statistics* **29**, 329–353.

[2] Valliant, R., Dever, J.A., Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

Estimation of compartment model parameters by combining genetic algorithm and jackknife method

Özlem Türkşen[1] **and Müjgan Tez**[2]

[1]*Ankara University, Turkey, turksen@ankara.edu.tr*
[2]*Marmara University, Istanbul, Turkey, mtez@marmara.edu.tr*

Keywords: compartment models, generalized nonlinear least squares, genetic algorithm, jackknife method

Modeling the pharmacokinetic behavior of a particular drug is a valuable tool in the drug development process. A well-known and commonly used model is two-compartment model which provides good insight into the underlying behavior of most drugs [1]. The model can be described analytically in the form of a system of ordinary differential equations. The solution of equation system is nonlinear form of the model parameters. Furthermore, compartments are correlated across the equations. In this case, generalized nonlinear least squares (GNLS) estimator is more efficient than nonlinear least squares (NLS) estimator [2]. The GNLS approach minimizes the Minkowski metric with respect to model parameters in which the covariance structure is not ignored.

In this study, estimation of two-compartment model parameters is considered in case of correlated equations. It is aimed to estimate the unknown model parameters based on GNLS minimization. For this purpose, genetic algorithm (GA), a well-known population based search algorithm [3], is used as optimization tool. In order to reduce the bias of the estimators, Jackknife delete-one algorithm [4] is used. The suggested approach is applied on simulated data set. It is seen from the results that bias of parameter estimates is reduced by using Jackknife method which helps to get statistical inference about the parameters.

References

[1] Bonate, P.L. (2011). *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*. Springer, New York.

[2] Seber, G.A.F. and Wild, C.J. (2003). *Nonlinear Regression*. Wiley, New York.

[3] Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, New York.

[4] Obiora-Ilouno H.O. and Mbegbu J.I. (2013). A jackknife approach to error-reduction in nonlinear regression estimation. *American Journal of Mathematics and Statistics* **3**, 32–39.

Independent component analysis for tensor-valued data

Joni Virta[1], Bing Li[2], Klaus Nordhausen[1] and Hannu Oja[1]

[1]*University of Turku, Finland, joni.virta@utu.fi, klaus.nordhausen@utu.fi, hannu.oja@utu.fi*
[2]*Pennsylvania State University, US, bing@stat.psu.edu*

Keywords: FOBI, Kronecker structure, matrix-valued data, multilinear algebra

In preprocessing high-dimensional tensor data, e.g. images or videos, a common procedure is to vectorize the observed tensors and subject the resulting vectors to one of the many methods used for independent component analysis (ICA). However, the structure of the original tensor is lost in the vectorization along with any meaningful interpretations of its modes. To provide a more suitable alternative, we propose the Tensor fourth order blind identification (TFOBI), a tensor-valued analogy of the classic Fourth order blind identification (FOBI), to be used with the semiparametric tensor independent component model. In TFOBI, instead of vectorizing, we stay in the tensor form and in a sense perform FOBI simultaneously on all the modes of the observed tensors. Furthermore, being an extension of FOBI, TFOBI shares with it its computational simplicity. Simulated and real-world examples are used to showcase the method's usefulness and superiority over the combination of vectorizing and FOBI.

On a family of robust estimators for autocorrelation coefficients under outliers

Valeriy Voloshko[1] **and Yuriy Kharin**[2]

[1]*Research Institute for Applied Problems of Mathematics and Informatics, Belarus, valeravoloshko@yandex.ru*
[2]*Belarusian State University, Minsk, Belarus, kharin@bsu.by*

Keywords: robustness, outlier, autocorrelation coefficient, time series

We consider observations $\{y_t\}$ derived from Gaussian stationary time series $\{x_t\}$, distorted by the so-called replacement outliers [3]. The robust ψ-estimate $\hat{\theta}_\tau$ for autocorrelation coefficient $\theta_\tau = \mathbb{COR}\{x_t, x_{t+\tau}\}$ of undistorted (hidden) time series $\{x_t\}$ is computed from observed distorted time series $\{y_t\}$ by the formula [1]:

$$\hat{\theta}_\tau ::= f_\psi^{-1}\left(\frac{(1-\varepsilon)^{-2}}{T-\tau}\sum_{t=1}^{T-\tau}\psi\left(\frac{y_t}{y_{t+\tau}}\right)\right),\ 0<\tau<T, \tag{1}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an odd bounded function, $f_\psi(\theta) ::= \mathbb{E}\psi(\zeta)$, ζ is the Cauchy distributed random variable with law $\mathcal{C}(\theta, \sqrt{1-\theta^2})$, $0 < \varepsilon \ll 1$ is the probability of a replacement outlier presence in $\{y_t\}_{t=1}^T$.

Under several assumptions on function ψ and asymptotical behavior of autocorrelation θ_τ at $\tau \to \infty$ the ψ-estimator (1) is shown to be consistent and asymptotically Gaussian [1, 2]. Some examples of $\psi(\cdot)$ generated ψ-estimators in the family (1) are given. Optimal function $\psi_*(\cdot)$ that minimizes the functional (an approximation for the mean squared error for the ψ-estimator) is found. Numerical comparison of ψ-estimator (1) w.r.t. the robust Huber estimator [4] is made based on real and simulated data.

References

[1] Kharin, Yu. and Voloshko, V. (2011). Robust estimation of AR coefficients under simultaneously influencing outliers and missing values. *Journal of Statistical Planning and Inference* **141**, 3276–3288.

[2] Kharin, Yu. (2013). *Robustness in Statistical Forecasting*. Springer, Heidelberg-Dordrecht-New York-London.

[3] Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

[4] Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Defining the population size using the residency index. Case of Estonia

Mare Vähi[1], Ethel Maasing[2] and Ene-Margit Tiit[3]

[1] *University of Tartu, Estonia, mare.vahi@ut.ee*
[2] *Statistics Estonia, Estonia, ethel.maasing@stat.ee*
[3] *University of Tartu and Statistics Estonia, Estonia, enemargit.tiit@stat.ee*

Keywords: residency index, discriminant analysis

Census data have been the most important and valuable data in population statistics from the very beginning of scientific thinking. But nowadays, when there exist different information sources, the reliability and exactness of census data does not satisfy. Here are two reasons: probably, the coverage of census has fallen due to mobility of people. Also, needs of researchers are higher today. After the Estonian census 2011 the census team found that there was some under-coverage present. To determine the amount of non-enumerated people the following procedure was used. The set of people belonging to Estonian population register as residents, but not enumerated in census 2011 were regarded as potential residents. All existing administrative registers were used to define the signs of life for these people: activity in a register during 2011 gave to a person a sign of life. The signs of life were used as binary variables to discriminant the residents and non-residents using several multivariate technics (linear and logistic discriminant analysis). Training groups consisted of the "true residents" (who belonged to PR as Estonian residents and were enumerated) and "true non-residents" (who were in PR as residents of foreign countries and also not enumerated as people living currently in Estonia).

As a result, the under-coverage of about 2,3% was found. The error of decisions was not more than 5%. The following task was to use the methodology for following years and to cover the whole population. Here the following problems arose: 1) No training groups exist for subsequent years; 2) Making decision by years the continuity and stability of population might disappear. Hence we decided to define for each person from the population a residency index between 0 and 1 that will be recalculated annually using the signs of life. If the value of index of a person is higher than threshold, then he/she is resident, if it drops below the threshold, the person is excluded from the set of residents. The recalculation formula of residency index $R(i, j)$ for person i in year j is the following:

$$R(i, j) = d(Ri, j - 1) + g \sum a_k X(i, j - 1, k),$$

where $R(i, j-1)$ is the index of the person in last year, $X(i, j, k)$ is the value of k-th sign of life of the person i in year j and a_k is the weight of the sign of life k. The parameters d and g have the value between 0 and 1. If the value of $R(i, j)$ is bigger than 1, it will be truncated to 1. All persons having the value of R higher than threshold b are considered as residents in year j, the others are non-residents, but will stay in the population and will be able to get the status of resident in future. The values of parameters d, g and b are defined from some theoretical concepts connected with the acceptable time of change the status. The weights a_k have been defined in different ways: all equal to 1 (simple sum of signs of life), proportional to their description value and also using some logarithmic scale. The efficiency of the index has been tested using a series of years: 2012, 2013, 2014 and 2015. The comparison of estimated number of residents and the official number of residents calculated by traditional way in Statistics Estonia has been made.

Recent advances in multivariate filter methods of variable selection for discrimination

Pädraic Walsh[1], Marta Garcia-Finana[2] and Gabriela Czanner[3]

[1] *University of Liverpool, United Kingdom, Padraic.Walsh@liverpool.ac.uk*
[2] *University of Liverpool, United Kingdom*
[3] *University of Liverpool, United Kingdom*

Keywords: variable selection for discrimination, probability of correct classification, filter, embedded

Variable selection is a common problem in discrimination when many potential predictors are considered. Filter methods are computationally fast and classifier-independent utilising a metric of discriminating ability to select variables however they may impose invalid assumptions on data. Embedded methods have high computational requirements (e.g. Random Forest). Hotelling's T^2 statistic is a multivariate index of the discriminating potential used by filter methods to select variables. It assumes equality of variance-covariance matrices across groups, which is often violated. We generalised Hotelling's T^2 statistic into the SNR ratio allowing heterogeneity of variance-covariance matrices across groups. We implemented SNR into a forward selection algorithm producing a novel method for variable selection. Using simulated data we demonstrated that SNR is better than T^2 at choosing the relevant discriminating variables (100 % vs. 46 %). In a comparison study with existing filter and embedded methods our algorithm demonstrated superior performance to filter methods and comparable performance to embedded methods but with reduced computational requirements. We investigated our methods in two clinical datasets: diabetic retinopathy (27 variables, and 103 patients) and a screening programme for sight-threatening diabetic retinopathy (17 variables, 5272 patients). We found that the variables chosen by SNR lead to better or equivalent classification accuracy compared to T^2 in terms of probability of correct classification (83 % vs. 76 %, and 76.5 % vs 76.5 %, respectively). We studied the performance of our methods for non-normal data; and demonstrated that our method is either superior or at least as good as alternative methods with high computational requirements. In our talk we will i) summarise recent advances in filter methods of variable selection for discrimination, ii) present our novel SNR describing its methodological properties in simulations, iii) present the forward selection algorithm discussing possible stopping criteria and iv) outline the challenges in applying SNR to real datasets.

Is non-parametric Bayesian MCMC a good model of evolutionary computation?

Chris Watkins

Royal Holloway, University of London, United Kingdom, C.Watkins@cs.rhul.ac.uk

Keywords: evolutionary computation, detailed balance, non-parametric Bayesian MCMC

Nature gives us two remarkable adaptive algorithms in asexual and sexual evolution. These algorithms produce radically different types of organisms. Simple considerations about fossil ancestry show that sexual evolution can be impressively fast.

What is the source of these two algorithms' computational effectiveness, and how should these two natural algorithms be computationally modelled? Many people have approached these questions – yet it is curious that there is so little communication between the machine learning and evolutionary computation communities, and it is also curious that evolution has been so little considered using the tools of machine learning.

Starting from first principles, I will propose models of both sexual and asexual evolution that are, at the same time, plausible computational abstractions of natural evolution, and also non-parametric Bayesian MCMC algorithms. They are genetic algorithms that satisfy detailed balance, and for which the stationary distribution over populations (also known as mutation-selection equilibrium) can be written in closed form for a general class of fitness functions.

Basic properties of the two models will be considered. The asexual and sexual models have structurally different 'prior' distributions, with radically different properties. There are also natural scaling relationships among parameters.

The relationship between individual learning and evolution can be modelled by placing both individual and evolutionary learning within one probability model in this framework.

Lastly, the "Bayesian" interpretation of evolution applies only when individual fitnesses do not depend on other individuals in the population. A natural question is whether group-dependent fitness, or coevolution, are essentially more powerful processes.

On palindromic Ising models with graph structure

Nanny Wermuth[1,2]

[1]*Chalmers University of Technology, Sweden, wermuth@chalmers.se,*
[2]*Gutenberg University, Mainz, Germany*

An example of a palindromic sentence which respects the spacings between words is 'step on no pets': it gives the same sentence when read in reverse order. This notion has now been applied to Bernoulli distributions, which are then characterised by having no odd-order interactions, no matter whether these are of the linear, log-linear or multivariate logistic type.

For Ising models with this structure, there are no main effects and at most two factor log-linear interactions so that the vanishing of such a term shows just as in joint Gaussian distributions in the concentration matrix, that is in their inverse covariance matrix. In this lecture, I concentrate on additional features which arise especially when their concentration graphs have simplified structure.

The results are based on [1] and [2] which are to appear.

References

[1] Marchetti, G.M. and Wermuth, N. (2016). Palindromic Bernoulli distributions. *Electronic Journal of Statistics (to appear).*

[2] Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., Zwiernik, P. (2016). Total positivity in Markov structures. *The Annals of Statistics (to appear).*

On multivariate geometric random sums

Igor V. Zolotukhin

P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences, Russia,
Igor.Zolotukhin@gmail.com

Keywords: multivariate geometric distribution, multivariate exponential distribution Marshall-Olkin type, generalized multivariate Laplace distribution

We can use multivariate geometric distribution to generalize the notion of geometric random sum to the multidimensional case.

To date have been studied limit distributions, which approximate the geometric sums in the form $\sum\limits_{j=1}^{L} W^{(j)}$, where $W^{(j)} = \left(W_1^{(j)}, \ldots, W_k^{(j)}\right)$ are independent identically distributed k-dimensional random vectors, L is a random variable having geometric distribution; L and $W^{(j)}$ ($j = 1, 2, \ldots$) are independent. Note that the number of terms will be the same for each component.

Let us consider the more general case. The number of random variables L_j ($j = 1, \ldots, k$) will be different for each component, while values of L_j could be **dependent**.

Multivariate geometric random sum is called a random vector sum of the form $S = (S_1, \ldots, S_k) = \left(\sum\limits_{j=1}^{L_1} W_1^{(j)}, \ldots, \sum\limits_{j=1}^{L_k} W_k^{(j)}\right)$, where L_m ($m = 1, \ldots, k$) will be defined below, $W_m^{(j)}$ are independent random variables identically distributed for each m having known characteristic function $E\,exp(i\,t_m\,W_m) = \varphi_m(t_m)$; lastly L_m and $W_m^{(j)}$ are independent.

The vector $L = (L_1, \ldots, L_k)$ is introduced by the following way. Let $\mathcal{E} = \{\epsilon\}$ be a set of $k-$dimensional indices $\epsilon = (\varepsilon_1, \ldots, \varepsilon_k)$ and each component of ε_m is 0 or 1; $\mathcal{E}_m$ is a set of $k-$dimensional indices for which $\varepsilon_m = 1$; N_ε are independent geometrically distributed random variables with parameters p_ε. By definition, put the value $L_m = \min\limits_{\varepsilon \in \mathcal{E}_m} \{N_\varepsilon\}$.

We show that the limit distributions of such sums by the corresponding normalization can be:

- multivariate exponential distribution introduced by Marshall and Olkin;
- multivariate generalized Laplace distribution introduced earlier by author.

Let us define the *marginally strictly geometric stable distribution* as the distribution of vector $R = (Z_1^{1/\alpha_1} Y_1, Z_2^{1/\alpha_2} Y_2, \ldots, Z_k^{1/\alpha_k} Y_k)$, where Y_m are independent random variables with strictly stable distributions with characteristic functions $g_m(\theta_m)$ and parameters $\alpha_m, \eta_m, \beta_m$; $Z = (Z_1, \ldots, Z_k)$ is independent from $Y_1, ..., Y_k$, random vector having the Marshall-Olkin multivariate exponential distribution.

Now let $p_\varepsilon = \lambda_\varepsilon p$. Assume that $\varphi_m(p^{1/\alpha_m}\theta_m) = 1 + p \ln g_m(\theta_m) + o(p)$ as $p \to 0$. We have proved that the normalized vector $\left(p^{1/\alpha_1} \sum\limits_{j=1}^{L_1} W_j^{(1)}, \ldots, p^{1/\alpha_k} \sum\limits_{j=1}^{L_k} W_j^{(k)}\right)$ converges weakly to R as $p \to 0$, where R has the *marginally strictly geometric stable distribution.*

On high dimensional data analysis in case of no sparsity

Silvelyn Zwanzig

Uppsala University, Sweden, silvelyn.zwanzig@math.uu.se

Keywords: high-dimensional, p larger than n, lasso, ridge

Consider a data set

$$\left(\mathbf{Y}, \mathbf{X}\right) = \begin{pmatrix} Y_1 & X_{1,1} & \cdots & X_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{1,n} & \cdots & X_{p,n} \end{pmatrix}. \tag{1}$$

When

$$p > n \quad \text{or even} \quad p \gg n,$$

the data are high-dimensional. A formal relationship between $\mathbf{Y}$ and $\mathbf{X}$ to be considered is through the linear model

$$\mathbf{Y} = \mathbf{X}\beta_0 + \epsilon, \tag{2}$$

where $\beta \in \mathbb{R}^p$ is the parameter vector, $\epsilon \in \mathbb{R}^n$ is the error vector of i.i.d. elements assumed to follow certain distribution with $\mathrm{E}(\epsilon) = \mathbf{0}$. The most frequently adopted route to tackle the problem of high-dimensionality is regularization by which a penalty term, $\mathrm{pen}(\beta)$, is added to the least-squares criterion. The regularized least-squares objective function is

$$\min_{\beta \in \mathbb{R}^p} \left[\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \mathrm{pen}(\beta) \right], \tag{3}$$

where the *tuning parameter* λ controls the intensity of penalization. The ridge estimator uses $\mathbb{L}_2$ norm as the penalty. LASSO combines least-squares L_2 loss with L_1 penalty. There now exist many variants of original LASSO with different penalty terms. In literature the basic idea is to set a condition by which not all covariates are needed, although it is unknown which of them can be deleted. The true parameter $\beta_0 = (\beta_1, \ldots, \beta_p)^T$ satisfies sparsity condition when

$$\|\beta_0\| = \sum_{j=1}^{p} |\beta_j| = o\left(\sqrt{\frac{n}{\log(p)}}\right). \tag{4}$$

In the talk the least squares estimator and different penalized estimators are compared under non-sparsity.

Generalized uniform correlation structure in the growth curve model

Ivan Žežula[1] **and Rastislav Rusnačko**

[1]*P. J. Šafárik University, Slovakia, ivan.zezula@upjs.sk*

Keywords: growth curve model, generalized uniform correlation structure, intra-class correlation structure

We consider classical growth curve model in the form

$$Y_{n\times p} = X_{n\times m}B_{m\times r}Z_{r\times p} + \varepsilon_{n\times p}, \quad E\varepsilon = 0, \quad \operatorname{var}\operatorname{vec}(\varepsilon) = \Sigma_{p\times p}\otimes I_n,$$

X being an ANOVA design matrix and Z being a regression design matrix, where $\Sigma = \theta_1 G + \theta_2 ww'$, $G \geq 0$, $w \in \mathcal{R}(G)$.

We study and compare properties of three different proposed estimators of the variance parameters θ_1 and θ_2 (see [1], [3], [2]).

Acknowledgment
This research was supported by grants VEGA MŠ SR 1/0344/14 and VEGA MŠ SR 1/0073/15.

References

[1] Khatri, C.G. (1973). Testing some covariance structures under a growth curve model. *Journal of Multivariate Analysis* **3**, 102 – 116.

[2] Ye, R.-D. and Wang, S.-G. (2009). Estimating parameters in extended growth curve model with special covariance structures. *Journal of Statistical Planning and Inference* **139**, 2746 – 2756.

[3] Žežula, I. (2006). Special variance structure in the growth curve model. *Journal of Multivariate Analysis* **97**, 606–618.

Tartu Conferences on Multivariate Statistics: A short retrospective view

Tõnu Kollo

University of Tartu, Estonia, tonu.kollo@ut.ee

The First Tartu Conference on Multivariate Statistics was held 34 years ago, 28-30 September 1977, as a Soviet Union wide conference with participants from various parts of the USSR. The principal speakers at the conference were Sergei A. Aivazjan (Moscow), Yuri K. Belyaev (Moscow), Ene-Margit Tiit and Liina-Mai Tooding (Tartu), who are still active researchers today. The conference was held 56 km south from Tartu at Kääriku recreation centre of Tartu University. Tartu conferences became the only regular event on multivariate statistics and data analysis in the Soviet Union.

The second conference was organized four years later, in 1981, at Sangaste manor house. The programme committee was chaired by Academician Yuri V. Prohorov and the keynote lecture was again delivered by Professor Sergei A. Aivazjan. On plenary sessions ten invited lectures were presented. Among the invited lecturers were Yuri K. Belyaev, Vladimir N. Vapnik, Ene-Margit Tiit, Liina-Mai Tooding and Vasili V. Nalimov.

The third conference was held in 1985, again at Kääriku. Fifteen invited lectures were delivered. The list of invited speakers included Sergei A. Aivazjan, Yuri K. Beljajev, Vyacheslav L. Girko, Igor G. Žurbenko, Yuri N. Blagoveschenski, Lev D. Meshalkin, Šarunas Raudis, Dmitri S. Silvestrov, Ene-Margit Tiit and Boris V. Gnedenko.

From left: J. Reiljan, Y.N. Blagoveschenski, B.V. Gnedenko, L.G. Afanasyeva, L.D. Meschalkin (1985)

The fourth conference, in 1989, was the last in the series of the Soviet Union wide conferences. It was again organized at Kääriku. Among invited speakers were Alexander V. Nagaev, V. V. Feodorov, Vladimir V. Anissimov, Vyacheslav L. Girko,

E.-M. Tiit at the Opening Section (1994)

Taivo Arak, Donatas Surgailis, Šarunas Raudis, Boris G. Mirkin. All the Soviet Union wide conferences were attended by more than one hundred participants, and there was always a tight competition to have your talk included in the program. For the first four conferences Professor Sergei A. Aivazjan was the main organizer in Moscow, while the local organization in Tartu was led by Ene-Margit Tiit.

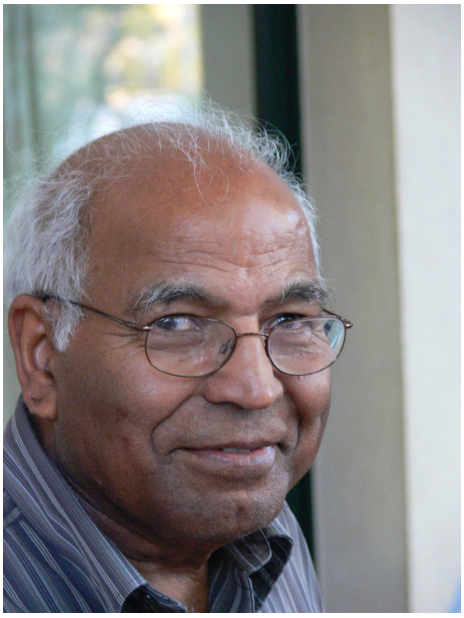
The V Tartu Conference on Multivariate Statistics was the first international conference in the series. It had taken longer than four years to organise this conference, now at international level. It was held 23-28 May 1994 jointly with the 3rd International Workshop on Matrices in Statistics. About 70 participants from 18 countries travelled to Tartu where the conference was opened. The following days were spent at the picturesque village of Pühajärve. The keynote speaker, Professor C. Radhakrishna Rao, found the atmosphere "friendly and stimulative". The creative atmosphere was enhanced by invited speakers Kai-Tai Fang, Yasunori Fujikoshi, Ingram Olkin and George P. H. Styan. The conference was followed by The 3rd International Workshop on Matrices in Statistics, the general organizer of which was Professor George P. H. Styan.

I. Olkin giving a talk (1994)

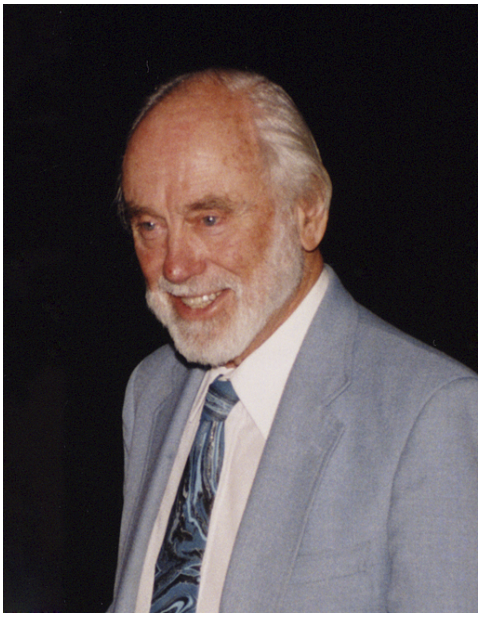
From left: Y. Fujikoshi, Mrs. C.R. Rao, C.R. Rao (1994)

From left: T. Kollo, K.-T. Fang, D. v. Rosen (1994)

From left: G.P.H. Styan, H.J. Werner, E.I. Im, H. Neudecker, S. Liu (1994)



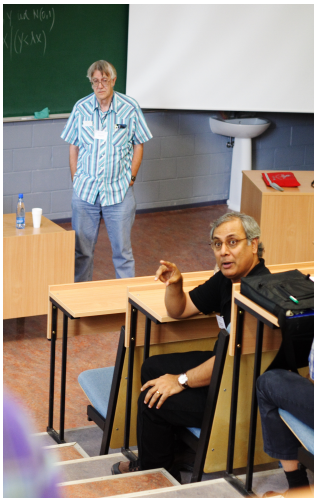
M.S. Srivastava (2003)



T.W. Anderson (1999)

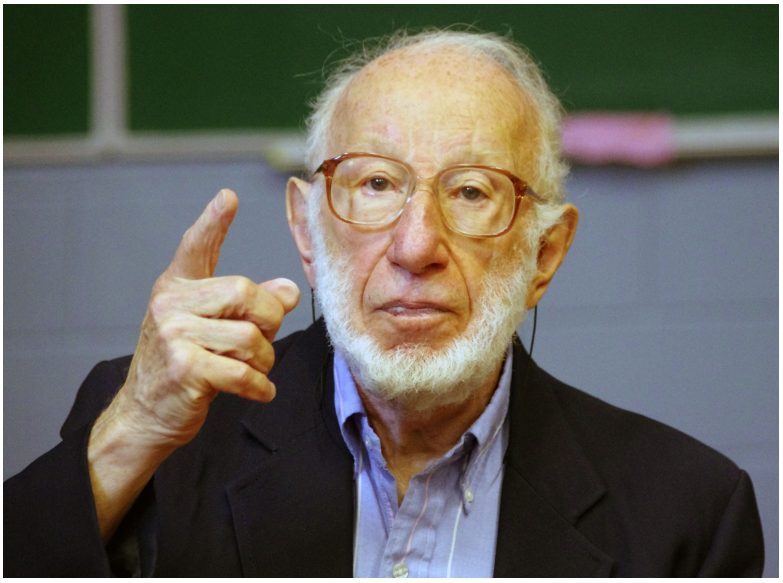
The VI Tartu Conference on Multivariate Statistics was held in 1999 in Tartu, as a satellite meeting of the 52nd Session of the International Statistical Institute in Helsinki. The stimulating working atmosphere at the conference was created by the honourable keynote lecturer Theodore W. Anderson and the distinguished invited speakers James Durbin, Kai-Tai Fang, Søren Johansen, Jürgen Läuter, Heinz Neudecker, Muni S. Srivastava and Helmut Strasser.

The VII conference was held in Tartu, 7-12 August 2003, as a Satellite Meeting of ISI 54th Session in Berlin. This time the keynote speakers were Professors Narayanaswamy Balakrishnan and Barry Arnold. Excellent invited lectures were given by Boris Mirkin, Akimishi Takemura, Steen Andersson, Muni S. Srivastava, Hannu Oja and Gad Nathan.



B. Arnold (standing) and
N. Balakrishnan (2007)

The VIII Tartu Conference was held jointly with The VI Conference on Multivariate Distributions with Fixed Marginals, 26-29 June 2007, under the auspices of the Bernoulli Society. The keynote speakers were Professors Muni S. Srivastava and Narayanaswamy Balakrishnan. The list of invited speakers included Michael Perlman, Nikolai Kolev, Peter E. Jupp, Steen Andersson, Christian Genest, Lennart Bondesson and Ludger Rüschendorf.



I. Olkin (2011)

The previous IX conference was held jointly with the XX International Workshop on Matrices and Statistics, 26 June - 1 July 2011, in Tartu under the auspices of the Bernoulli Society. The keynote speaker was Professor Ingram Olkin. Samuel Kotz Memorial Lecture was delivered by Professor Narayanaswamy Balakrishnan. A special session was organized dedicated to the 75th jubilee of Muni S. Srivastava. Invited lectures were given by Adelchi Azzalini, Michael Greenacre, Bimal Sinha, Bent Jørgensen, Augustyn Markiewicz, Muni S. Srivastava and Julia Volaufova.

Tõnu Kollo
Vice-Chair of the Programme Committee

Index