The 9th Tartu Conference on Multivariate Statistics

&

The 20th International Workshop on Matrices and Statistics

Abstracts

26 June - 1 July 2011, Tartu, Estonia

under the auspices of the Bernoulli Society

Tartu
2011

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

Editors: Tõnu Kollo, Kelli Sander, Ants Kaasik

Dear Participants,

Welcome to Tartu!

The first Tartu Conference on Multivariate Statistics was held 34 years ago, 1977. We are happy that today we have among active participants of the IX Conference two Invited Speakers of the First Conference, Yuri Belyaev and Ene-Margit Tiit. The IX Tartu Conference on Multivariate Statistics is held jointly with the XX International Workshop on Matrices and Statistics under auspices of the Bernoulli Society for Mathematical Statistics and Probability. In the end of this volume you can find short retrospective overviews of these two conference series.

The talks will be given within four days, June 27-30, 2011. They include two Keynote Lectures delivered by Professor Ingram Olkin and Samuel Kotz Memorial Lecture given by Professor N. Balakrishnan. There will be a Special Section dedicated to the 75-th jubilee of Professor Muni. S. Srivastava. The talks cover wide range of areas from probability theory and theoretical developments of mathematical statististics and distribution theory to applications of multivariate analysis in different areas: finance, insurance, economics, genetics, demography etc.

This volume contains the abstracts of the papers to be presented at the Conference in alphabetic order, following Estonian alphabet. Style of the abstracts has been kept unchanged during editing. Only some misprints have been corrected. Organizers are grateful to all the authors for their cooperation.

Programme Committee wishes all of you fruitful ideas and enjoyable time in Tartu.

Tõnu Kollo
Vice-Chair of the Programme Committee

Block-wise permutation tests for correlated multivariate imaging data

Daniela Adolf and Siegfried Kropf

Department of Biometrics and Medical Informatics,
Otto-von-Guericke University Magdeburg, Germany,
email: daniela.adolf@med.ovgu.de, siegfried.kropf@med.ovgu.de

Keywords: block-wise permutation, correlated sample elements, separated multivariate GLM.

In view of functional magnetic resonance imaging data, that is high-dimensional and correlated in time and space, we consider a multivariate general linear model (GLM) for a fMRI session with one person

$$\boldsymbol{Y} = \boldsymbol{X}\mathbf{B} + \mathbf{E}, \quad \mathbf{E} \sim N_{n\times p}(\mathbf{0}, \boldsymbol{P} \otimes \boldsymbol{\Sigma})$$

The data matrix $\boldsymbol{Y}$ contains n measurements (successive fMRI scans) over p variables whereas $p \gg n$. In general the null hypothesis is $\mathrm{H}_0 : \boldsymbol{C}'\mathbf{B} = \mathbf{0}$ with $\boldsymbol{C}$ being an $s \times m$-dimensional contrast weight matrix. Here contrary to the classical multivariate GLM, the sample vectors are correlated and $\boldsymbol{P}$ is supposed to be a first-order autoregressive process. To analyze these data non-parametrically, we use a block-wise permutation method including a random shift in order to count for the temporal correlation.

Furthermore, we want to be able to test any null hypothesis on the parameter estimates via this special permutation method. This is important because analyzing functional imaging data is particularly based on testing differences of parameter estimates. Therefore, we use a separated multivariate GLM

$$\boldsymbol{Y} = (\boldsymbol{X}_1\boldsymbol{X}_2)\begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix} + \mathbf{E} = \boldsymbol{X}_1\mathbf{B}_1 + \boldsymbol{X}_2\mathbf{B}_2 + \mathbf{E}$$

and the special null hypothesis $\mathrm{H}_0 : \mathbf{B}_2 = \mathbf{0}$ that is only related to $\boldsymbol{X}_2$, that part of the design matrix that contains the information of interest.

We will show that any null hypothesis on the classical multivariate linear model can be transformed into the separated model and can be tested via the block-wise permutation method including a random shift.

References

[1] Friston, K.J., Ashburner, J.T., Kiebel, S., Nichols, T.E., Penny, W.D. (2007). *Statistical Parametric Mapping – The Analysis of Functional Brain Images.* Academic Press, Amsterdam.

[2] Kherad-Pajouh, S., Renaud, O. (2010). An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. *Computational Statistics and Data Analysis* **54**, 1881–1893.

[3] Läuter, J., Glimm, E., Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26**, 1972–1988.

[4] Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Applications and Software.* John Wiley & Sons, Chichester.

Some tests for covariance matrices with large dimension

M. Rauf Ahmad[1], Martin Ohlson[1] and D. von Rosen[1,2]

[1] *Linköping University, Sweden, email: {muahm, mohl}@mai.liu.se*
[2] *Swedish University of Agricultural Sciences, Uppsala, Sweden, email: Dietrich.von.rosen@et.slu.se*

Keywords: covariance testing, high-dimensionality, sphericity.

Let $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})'$, $k = 1, \ldots, n$, be n independent and identically distributed random vectors where $\mathbf{X}_k \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We present test statistics for

$$H_0 : \boldsymbol{\Sigma} = \mathbf{I} \quad \text{and} \quad H_0 : \boldsymbol{\Sigma} = \kappa \mathbf{I},$$

when p may be large, and may even exceed n, where $\kappa > 0$ is any constant. The test statistics are constructed using unbiased and consistent estimators composed of quadratic and bilinear forms of the random vectors $\mathbf{X}_k$. Under very general settings, the proposed test statistics are shown to follow an approximate normal distribution, for large n and p, inclusive of the case when $p > n$, or even $p \gg n$. The statistics are based on minimal conditions avoiding the usually adopted stringent assumptions found in the literature for similar high-dimensional inferences, for example assumptions on the traces of powers of the covariance matrix $\boldsymbol{\Sigma}$, or assumptions on the relations between p and n (see, for example, [2], [1]: Chs. 5&8). The performance of the test statistics is shown through simulations. It is demonstrated that the test statistics are accurate for both, size control and power for moderate n and any p, where p can be much large than n. The real life application of the statistics is also illustrated using practical data sets.

References

[1] Fujikoshi, Y., Ulyanov, V. U., Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley, New York.

[2] Ledoit, O., Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, **30**(4), 1081-1102.

Estimating regression parameters: a mosaic of estimation strategies

Ejaz Ahmed

University of Windsor, Canada, email: seahmed@uwindsor.ca

Keywords: linear model, penalty type estimator, regression parameters, shrinkage estimator.

In this talk, I address the problem of estimating a vector of regression parameters in a partially linear model. My main objective is to provide natural adaptive estimators that significantly improve upon the classical procedures in the situation where some of the predictors are inactive that may not affect the association between the response and the main predictors.

In the context of two competing regression models (full and sub-models), we consider shrinkage estimation strategy. The shrinkage estimators are shown to have higher efficiency than the classical estimators for a wide class of models. We develop the properties of these estimators using the notion of asymptotic distributional risk. Further, we proposed absolute penalty type estimator (APE) for the regression parameters which is an extension of the LASSO method for linear models. The relative dominance picture of the estimators are established. Monte Carlo simulation experiments are conducted and the non-parametric component is estimated based on kernel smoothing and B-spline. Further, the performance of each procedure is evaluated in terms of simulated mean squared error. The comparison reveals that the shrinkage strategy performs better than the APE/LASSO strategy when, and only when, there are many nuisance variables in the model. I plan to conclude this talk by applying the suggested estimation strategies on a real data set which illustrates the usefulness of procedures in practice.

Maximum likelihood estimates for Markov-additive processes of arrivals by aggregated data

Alexander Andronov

Transport and Telecommunication Institute, Latvija,
email: lora@mailbox.riga.lv

Keywords: additive components, parameter estimation, time-homogeneous Markov process.

We consider a simplification of Markov-additive process of arrivals. Let $\mathbf{N} = \{0, 1, ...\}$, r be a positive integer, E be a countable set, and $(\mathbf{X}, J) = \{(\mathbf{X}(t), J(t)),\ t \geq 0\}$ be a considered process on state space $N^r \times E$. The increments of $\mathbf{X}$ are associated to arrival events. Different (namely r) classes of arrivals are possible, so $X_i(t)$ = total number of arrivals in $(0, t]$ in the class i, $i = 1, 2, ..., r$. We call $\mathbf{X}$ the *arrival component* of $(\mathbf{X}, J)$, and J - the *Markov component* of $(\mathbf{X}, J)$.
Whenever the Markov component J is in the state j, the following two types of transitions in $(\mathbf{X}, J)$ may occur. 1) The i-arrivals without a change of state in $j \in E$ occur at rate $\lambda_j^i(n),\ \ n > 0$. 2) Changes of state in J without arrivals occur at rate $\lambda_{j,k}$, $k \in E$, $j \neq k$.

We suppose that J is a birth and death process. Let $\overrightarrow{\lambda} = \Big(\lambda_{j,j+1} : j = 1, ..., m-1\Big)$, $\overleftarrow{\lambda} = \Big(\lambda_{j,j-1} : j = 2, ..., m\Big)$. If the state $j \in E$ is fixed, then different arrivals form independent Poisson flows. Further, let $q_i(n)$ be a probability that i-arrival contains n items, $\sum_{n>0} q_i(n) = 1$. These probabilities do not depend on state $j \in J$ and are the known ones. Now, the i-arrival rates have the following structure: $\lambda_j^i(n) = v_j\Big(\alpha^{\langle i\rangle}\Big) q_i(n)$, $j = 1, ..., m$, where v_j is a known function to an approximation of the parameters $\alpha^{\langle i\rangle} = \Big(\alpha_{1,i}, \alpha_{2,i}, ..., \alpha_{k,i}\Big)^T$.
We consider a problem of unknown parameters $\alpha = \Big(\alpha^{\langle 1\rangle}\ \ \alpha^{\langle 2\rangle}\ \ ...\ \ \alpha^{\langle r\rangle}\Big)_{k\times r}$, $\overrightarrow{\lambda}$ and $\overleftarrow{\lambda}$ estimation. It is supposed that we have n independent copies $X^{(1)}(t), ..., X^{(n)}(t)$ of the considered process $X(t) = \Big(X_1(t), ..., X_r(t)\Big)^T$ - total numbers of arrivals of various classes in $(0, t]$. Our initial point is the following: each $X(t)$ has multivariate normal distribution with mean $E(X(t)) = t\mu$ and covariance matrix $Cov(X(t)) = tC$, where μ is r-dimensional column vector and C is $(r \times r)$-matrix. The sample mean μ^* and the sample covariance matrix $\mathbf{C}^*$ are sufficient statistics, therefore we must make statistical inferences on this basis. In the paper maximum likelihood estimates are calculated for unknown parameters.

References

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

[2] Turkington, D. A. (2002). *Matrix Calculus and Zero-One Matrices. Statistical and Econometric Applications.* Cambridge University Press, Cambridge.

On skewed $l_{n,p}$-symmetric distributions

Reinaldo B. Arellano-Valle[1] **and Wolf-Dieter Richter**[2]

[1] *Pontificia Universidad Católica de Chile, Chile, email: reivalle@mat.puc.cl*
[2] *University of Rostock, Germany, email: wolf-dieter.richter@mathematik.uni-rostock.de*

Keywords: $l_{n,p}$-symmetric distributions, skewed distributions.

Skewed elliptically contoured distributions were introduced first in [3]. Many authors extended these consideration under various aspects and in different ways. The book [4] gives an overlook on these efforts.

The authors of [1] bring a certain new structure into the field and unify many different approaches from a selectional point of view. The concept of fundamental skew distributions which unifies all at this time known approaches has been developed in [2].

Based upon a generalized method of indivisibles which makes use of the notion of non-Euclidean surface content, in [7] a geometric measure representation formula for $l_{n,p}$-symmetric distributions is derived. This formula enables one to derive exact distributions of several types of functions of $l_{n,p}$-symmetrically distributed random vectors. This has been demonstrated by generalizing the Fisher distribution in [7] and also for several special cases in [6] and [5].

Here we extend the class of skewed distributions for cases where the underlying distribution is an $l_{n,p}$-symmetric one. To this end, we first exploit the geometric measure representation formula in [7] to derive marginal and conditional distributions from $l_{n,p}$-symmetric distributions. Then, the general density formula for skewed distributions from [1] applies and finally we follow the general concept in [2].

References

[1] Arellano-Valle, R.B., Branco, M.D., Genton, M.G. (2006). A unified view on skewed distributions arising from selections. *The Canadian Journal of Statistics* **34**(4), 1–21.

[2] Arellano-Valle, R.B., Genton, M.G. (2005). On fundamental skew distributions. *Journal Multiv. Anal.* **96**, 93–116.

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stastist.* **12**, 171–178.

[4] Genton, M.G., ed. (2004). *Skew-elliptical Distributions and Their Applications: A Journey Beyond Normality.* Chapman Hall CRC, Boca Raton.

[5] Kalke, S., Richter, W.-D. (2011). Linear combinations, products and ratios of simplicial or spherical variates, (*submitted*).

[6] Richter, W.-D. (2007). Generalized spherical and simplicial coordinates. *Lithuanian Mathem. J.* **49**(1), 93–108.

[7] Richter, W.-D. (2009). Continuous $l_{n,p}$-symmetric distributions. *Lithuanian Mathem. J.* **49**(1), 93–108.

On estimation problems for multivariate skew-symmetric distributions

Adelchi Azzalini

University of Padua, Italy, email: azzalini@stat.unipd.it

Keywords: anomalies of maximum likelihood estimation, skew-symmetric distributions

A currently active stream of literature deals with continuous multivariate distributions whose density function is the form

$$f(x) = 2\, f_0(x)\, G\{w(x)\}, \qquad\qquad x \in \mathbb{R}^d,$$

where f_0 is a density function such that $f_0(x) = f_0(-x)$, G is a distribution function on the real line such that G' exists and is an even function, and w is odd in the sense that $w(-x) = -w(x) \in \mathbb{R}$. The term 'skew-symmetric' is often used to refer to a density $f(x)$ of this type, although the effect of perturbation of the symmetric density $f_0(x)$ by the factor $G\{w(x)\}$ can be more complex than turning it into an asymmetric distribution.

Two important special cases of this construction are the so-called skew-normal and the skew-t distribution, which are obtained by choosing the ingredients as follows:

f_0	$G(w)$	$w(x)$
$N_d(0, \Omega)$ density	$\Phi(w)$	$\alpha^\top \omega^{-1} x$
$t_d(\nu, \Omega)$ density	$T(w, \nu + d)$	$\alpha^\top \omega^{-1} x \left(\dfrac{\nu + d}{\nu + x^\top \Omega^{-1} x}\right)^{1/2}$

where a standard type of notation is adopted, and ω is a diagonal matrix whose non-null terms are the standard deviations associated to the variance matrix Ω.

While the probability side of this formulation leads to a smooth mathematical development, and several nice properties follow with relatively little effort, its statistics side has shown to be more challenging. More specifically, maximum likelihood estimation (MLE) of the above parameters Ω, α and ν, when this is present, complemented by a location parameter, can exhibit two sort of anomalies:

(i) the observed and the expected information matrices are singular at $\alpha = 0$ for certain families, in particular for the skew-normal family indicated above,

(ii) for finite sample size, the MLE of α may happen to diverge with non-null probability.

We shall first review the state of the art for this estimation problem, and then focus on its case (ii) which so far has not yet been given a satisfactory general solution. A proposal based on a form of penalized likelihood function will be put forward.

On Pearson-Kotz Dirichlet distributions

Narayanaswamy Balakrishnan[1] **and Enkelejd Hashorva**[2]

[1] *McMaster University, Hamilton, Ontario, Canada,*
email: bala@univmail.cis.mcmaster.ca
[2] *University of Lausanne, Lausanne, Switzerland, email: Enkelejd.Hashorva@unil.ch*

Keywords: conditional distribution, Pearson-Kotz Dirichlet distributions, random vectors.

In this talk, I will discuss some basic distributional and asymptotic properties of the Pearson-Kotz Dirichlet multivariate distributions. These distributions, which appear as the limit of conditional Dirichlet random vectors, possess many appealing properties and are interesting from theoretical as well as applied points of view. Finally, I will illustrate an application concerning the approximation of the joint conditional excess distribution of elliptically symmetric random vectors.

Analysis of contingent valuation data with self-selected rounded WTP-intervals collected by two-steps sampling plans

Yuri K. Belyaev

Umeå University, Sweden, email: yuri.belyaev@math.umu.se
Swedish University of Agricultural Sciences, Sweden, email: yuri.belyaev@sekon.slu.se

Keywords: estimable characteristics, interval rounded data, maximisation likelihood, recursion, resampling.

In collecting contingent valuation data on Willingness To Pay (WTP-)points, rather than asking a respondent to state an estimate of his/her WTP-point or select one between given brackets, the respondent may freely self-select any interval of choice that contains the WTP-point. For the collected data, we found that presence of strong rounding is a typical feature. The self-selected intervals can be considered as censoring the true WTP-points. Usually in the Survival Analysis it is assumed that the censoring intervals are independent of such points and cover only some of them. But here these intervals can depend on the unobserved positions of their WTP-points, and all WTP-points are covered. Due to rounding many of the same self-selected intervals will be often stated by different respondents. We suppose that the true WTP-points corresponding different respondents can be considered as values of independent identically distributed random variables. It is useful to find consistent estimates related to the distribution of these WTP-points. We propose statistical models which admit dependency of the self- selected WTP-intervals on the positions of their WTP-points. Note that one has to distinguish between the probability to select an interval containing WTP-point and the probability of the different event that the interval contains the WTP-point.

We suggest a two-step plan of random sampling individuals from a population of interest that it would be possible consistently to estimate (identify) some of important characteristics of the unknown distribution of WTP-points. On the first step freely self-selected WTP-intervals are collected. It is possible to recognize weather the size of the first sample is large enough to guarantee be related to a desired majority of the population of interest. Based on the collected set U of different stated self-select intervals the collection V of division intervals is generated. Each interval in U is a union of related division intervals. Besides that two auxiliary subsets from U and V are calculated. On the second step new random selection of individuals continued. Each selected respondent is asked to state freely a self-selected WTP-interval containing true WTP-point. If the stated interval has already been registered in U then as soon as possible the respondent should be suggested to select, from the related division intervals, the interval containing the true WTP-point. In this case the pair of both, the initially stated WTP-interval and the more exact selected division interval has to be added to the second step sample. If the respondent was not able to select such division interval then the only single self-selected interval has to be added to the second step sample. The subset of pairs is used for estimation of conditional probabilities to state a self-selected interval given the division interval containing the true WTP-point.

The log likelihood function, which parameters are probabilities of divisions intervals containing the true WTP-points, given the list of all selected division intervals in the pairs, and the all single self-selected intervals, can be written. The maximum likelihood (ML-)estimates of the projection of WTP-distribution on the set of all division intervals is obtained based on special recursion. The maximizing likelihood recursion is obtained by the method of Lagrange multipliers. The consistent lower and upper bounds of the mean WTP-value and the consistent estimate of medium mean WTP-distribution are calculated. Accuracy of these estimators can be characterized by the distributions of their deviations from the true unknown values. The distributions of deviations can be found by applying related resampling method. The detailed description of this research work, joint with Bengt Kriström, is given in [1].

References

[1] Belyaev, Yu. K., Kriström, B. (2011). Two-Step Approach to Self-Selected Interval Data in Elicitation Surveys, (in preparation).

Error Orthogonal Models and Commutative Orthogonal Block Structure: equivalence

Francisco Carvalho[1] **and João T. Mexia**[2]

[1] *Instituto Politécnico de Tomar, email: fpcarvalho@ipt.pt*
[2] *Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa*
CMA - Centro de Matemática e Aplicações

Keywords: COBS, error orthogonal, linear models, OBS.

We establish the equivalence of two important classes of models with Orthogonal Block Structure (OBS), namely:

- Error orthogonal models, whose least squares estimators are UBLUE, having the family of variance-covariance matrices given by $\boldsymbol{V} = \left\{ \sum_{j=1}^{m} \gamma_j \boldsymbol{Q}_j \right\}$;
- COBS, these are the models whose orthogonal projection matrix on the space spanned by the mean vector commutes with the matrices $\boldsymbol{Q}_1, \dots, \boldsymbol{Q}_m$.

This equivalence is fruitful since it enables us to use the model structure to estimate variance components.

Combined method which includes statistical and heuristic approaches in object recognition

Nikolay Chichvarin and Ivan Chichvarin

Bauman Moscow State Technical University, Russia, email: genrih.gerz@bmstu.ru

Suggested method is based on classical formulation of the problem of recognition of an object in complex environment, composed from components listed below:

- the deterministic background;
- random background;
- the object that is recognized.

The formulation of the problem of recognition is devided into three parts:

- Problem of preprocessing the image, that can contain or doesn't contain the desired object;
- Problem of recognition;
- Problem of identification.

This article describes solving of the problem of preprocessing in two aspects:

- Filtration of the incoming signal by means of statistical methods;
- Partly restoring of defocused images by solving the inverse problem.

Object's detection in distorted image is defined as a procedure of comparing the result of transformation of analysed image with some threshold value:

$$L \mid A(x,y) \mid \geq \prod \mid A(x,y) \mid$$

where $L|\cdot|$ is a transformation operator, $\prod|\cdot|$ is a threshold value operator. The object is detected, if image meets the condition, described above. The quality of recognition is characterized by the probability that the condition is fullfilled in the case when the image contains the object.

It is also well-known the exact form of the operators $L|\cdot|$ and $\prod|\cdot|$ and quality of the recognition depend on the existance of apriori data about desired objects, noises, interferences and distortions. Therefore, as a basis for determining optimal parameters of operators and criteria, in this article the fundamentals of statistical decision theory are used, and corresponding criterion is proposed.

In this article we define identification problem as comparison of image meant to be the desired object with etalons from some defined class. So the identification problem is reduced to a classification problem. We also take into account that identification problem is commonly solved with the following methods:

- Method of direct comparison of the object with an etalon image;
- Correlation method;

- Identification method based on the system of features.

Identification method based on the system of features also uses etalons, but compares object's features instead of the whole etalon. It helps to reduce the volume of needed memory and the processing time. One has to remember that extraction can introduce errors, so it is better to use histograms for features values.

When there are many different objects, the hierarchical algoritms are commonly used, so that on lower levels we deal with features which do not require big amount of computation, and on the higher levels, where the amount of objects is less, one can use more informative features.

The first two methods have high computational complexity.

The increase of the processing speed in solving of the problem of recognition is an actual task. We propose a method and implementation for its algorithm, which allows to increase the speed of recognition of the object against the background of a complex scene in terms of interference.

At the identification stage we propose to use a heuristic learning algorithm. The article illustrates the results of the programs that implement the algorithm based on the proposed method.

Testing goodness-of-fit with parametric AFT model

Ekaterina Chimitova and Natalia Galanova

Novosibirsk State Technical University, Novosibirsk, Russia, email: chim@mail.ru, natalia-galanova@yandex.ru

Keywords: AFT model, Anderson-Darling test, censored samples, Cramer-von Mises-Smirnov test, goodness-of-fit, Kolmogorov test, χ^2 RRN test for AFT model.

Let the nonnegative random variable ξ denote the time-to-event or failure time of an individual. The probability of an item surviving up to the time t is given by the survival function:

$$S(t) = P\{\xi > t\} = 1 - F(t),$$

where $F(t)$ is the cumulative distribution function of random variable ξ.

One of the well-known regression models in reliability and survival analysis is the Accelerated Failure Time model (AFT model). Usually in accelerated life testing all items are divided into several groups and tested under different accelerated stress conditions. Following [1], the survival function for parametric AFT model under constant over time stress x can be calculated as:

$$S(t, \beta) = S_0\left(\frac{t}{\rho(x, \beta)}\right),$$

where $\rho(x, \beta)$ is the stress function and S_0 is the baseline survival function, which usually belongs to some parametric family of distributions, such as Exponential, Weibull, Gamma, Generalized Weibull and others.

In this paper we consider the problem of testing goodness-of-fit with parametric AFT-model. One approach to this problem is based on using residuals. If the choice of baseline survival function is appropriate, then the sample of residuals belongs to the baseline distribution, standardized by the scale parameter. For testing this hypothesis classical nonparametric goodness-of-fit tests can be used: Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test ([2]).

The second approach considered in this paper is the χ^2 RRN goodness-of-fit test for parametric AFT model [1]. This test is based on division of the interval $[0, T]$ into smaller intervals and comparing observed and expected numbers of failures.

In this paper statistical distributions of these considered goodness-of-fit tests are investigated with computer simulation technique for complete and censored data. Statistical distributions under the valid null hypothesis are considered in dependance of baseline distribution, size of failure sample and censoring degree. The considered goodness-of-fit tests are compared by power for close competing hypotheses.

References

[1] Bagdonavicius, V., Kruopis, J., Nikulin, M. (2010). *Nonparametric Tests for Censored Data.* Wiley-ISTE.

[2] Lemeshko, B.Yu., Lemeshko, S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *The Annals of Statistics* **52**(6), 555–565.

Another generalization of bivariate FGM distributions

Carles M. Cuadras and Walter Diaz

University of Barcelona, Spain, email: ccuadras@ub.edu, wdiaz0@hotmail.com

Keywords: copulas, Farlie-Gumbel-Morgenstern distribution, given marginals, Pearson's contingency coefficient.

Let $H(x,y)$ be the bivariate cdf of (X,Y), with univariate marginals $F(x), G(y)$ and supports $[a,b], [c,d]$, respectively. Throughout this abstract, x and y in $H(x,y)$, $F(x)$, $G(y)$, as well as u and v in $C(u,v)$, where $0 \leq u, v \leq 1$, will be suppressed. We write $H \in \mathcal{F}(F,G)$, where $\mathcal{F}(F,G)$ is the family of cdf's with marginals F, G.

The Farlie-Gumbel-Morgenstern (FGM) family is $H_\theta = FG[1+\theta(1-F)(1-G)]$, $-1 \leq \theta \leq 1$, and the corresponding copula is $C_\theta = uv[1+\theta(1-u)(1-v)], -1 \leq \theta \leq 1$. This family is frequently used in theory and applications. This motivated to study proper extensions in [2] and [1].

Let Φ, Ψ be two univariate cdf's with the same supports $[a,b], [c,d]$. Suppose that the Radon-Nykodim derivatives $d\Phi/dG$, $d\Psi/dG$ exist. We define the bivariate cdf

$$H = FG + \lambda(F - \Phi)(G - \Psi).$$

This cdf reduces to the classic FGM for $\Phi = F^2, \Psi = G^2$, and has interesting properties:

1. $H \in \mathcal{F}(F,G)$ for λ belonging to an interval depending on $d\Phi/dG$, $d\Psi/dG$.
2. H suggests the congugate family $H_* \in \mathcal{F}(\Phi, \Psi)$.
3. Define $a_1 = 1 - d\Phi/dF, b_1 = 1 - d\Psi/dG$. Then $E[a_1(X)] = E[b_1(Y)] = 0$ and $E[a_1^2(X)] = \alpha - 1$, $E[b_1^2(Y)] = \beta - 1$, where $\alpha = \int_a^b (\frac{d\Phi}{dF})^2 dF, \quad \beta = \int_c^d (\frac{d\Psi}{dG})^2 dG$.
4. The first canonical correlation is $\rho_1 = \lambda\sqrt{(\alpha-1)(\beta-1)}$ and Pearson contingency coefficient is $\phi^2 = \rho_1^2$.
5. Spearman's rho and Kendall's tau are $\rho_S = 12\lambda(\frac{1}{2} - F_\Phi)(\frac{1}{2} - G_\Psi)$ and $\tau = 8\lambda(\frac{1}{2} - F_\Phi)(\frac{1}{2} - G_\Psi)$, where $F_\Phi = \int_a^b \Phi dF$, $\Phi_F = \int_c^d F d\Phi$.

The geometric dimensionality of a bivariate cdf is defined and discussed. Then we introduce the following generalized FGM

$$\begin{aligned} H &= FG + \lambda_1(F - \Phi)(G - \Psi) \\ &\quad + \lambda_2[(\tfrac{1}{2}F^2 + (F_\Phi - \tfrac{1}{2})F - F_\Phi(x)][(\tfrac{1}{2}G^2 + (G_\Psi - \tfrac{1}{2})G - G_\Psi(y)], \end{aligned}$$

where $F_\Phi(x) = \int_a^x \Phi(t)dF(t)$, $G_\Psi(y) = \int_c^y \Psi(t)dG(t)$. This $H \in \mathcal{F}(F,G)$ is diagonal and two-dimensional. Finally we study how to approximate any cdf by a member of this family.

References

[1] Cuadras, C. M. (2008). Constructing copula functions with weighted geometric means. *Journal of Statistical Planning and Inference* **139**, 3766–3772.

[2] Rodríguez-Lallena, J. A., Úbeda-Flores, M. (2004). A new class of bivariate copulas. *Statistics & Probability Letters* **66**, 315–325.

Measures of conditional asymmetry

Ali Dolati

Department of Statistics, College of Mathematics, Yazd University, Yazd, Iran,
adolati@yazduni.ac.ir

Keywords: conditional symmetry, test.

Let (X, Y) be a pair of continuous random variables with the joint distribution function H and univariate marginal distribution functions F and G. Let $H(y|x) = P(Y \leq y|X = x)$ denote the conditional distribution of Y given $X = x$. Formally, the random variable Y is conditionally symmetric given $X = x$ if $Y|X = x$ is symmetric; i.e., $H(y|x) = 1 - H(-y|x)$. Consequently, conditionally symmetric random variable Y given $X = x$ must be symmetric, i.e., $G(y) = 1 - G(-x)$. Of course, the converse is false. When Y is symmetric but not conditionally symmetric, then $H(y|x) \neq 1 - H(-y|x)$, for some x and y. Conditional symmetry is of interest in modelling time series data in business and finance; see e.g., [2, 3, 5]. There are some tests for identifying conditional symmetry in the statistical literature, see for example [1, 4, 6]. However, little effort was made in proposing measures for evaluating the degree of this kind of asymmetry present in data. This talk discusses some indices to measure conditional asymmetry for continuous random variables.

References

[1] Bai, J., Ng, S. (2001). A consistent test for conditional symmetry in time series models. *Journal of Econometrics* **103**, 225–258.

[2] Brännäs, K., De Gooijer, J.G. (1992). Modelling business cycle data using autoregressive-asymmetric moving average models. *ASA Proceedings of Business and Economic Statistics Section*, 331–336.

[3] De Gooijer, J.G., Grannoun, A. (2000). Nonparametric conditional predictive regions for time series. *Computational Statistics and Data Analysis* **33**, 259–275.

[4] Hyndman, R. J., Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics* **14**, 259–278.

[5] Polonik, W., Yao, Q. (2000). Conditional minimum volume predictive regions for stochastic processes. *Journal of American Statistical Association* **95**, 509–519.

[6] Zheng, J. X. (1998). Consistent specification testing for conditional symmetry. *Econometric Theory* **14**, 139–149.

Optimal classification of the multivariate GRF observations

Kęstutis Dučinskas and Lina Dreižienė

Department of Statistics, Klaipėda University, H. Manto 84, Klaipėda LT 92294, Lithuania, email: kestutis.ducinskas@ku.lt, l.dreiziene@gmail.com

Keywords: actual risk, Bayes discriminant function, covariogram, Gaussian random field, training labels configuration.

The problem of classifying a single observation from a multivariate Gaussian field into one of the two populations specified by different parametric mean models and common intrinsic covariogram is considered. This paper concerns with classification procedures associated with Bayes Discriminant Function (BDF) under the deterministic spatial sampling design. In the case of parametric uncertainty, the maximum likelihood estimators of unknown parameters are plugged in the BDF. The actual risk and the Approximation of the Expected Risk (AER) associated with aforementioned plug-in BDF are derived. This is an extension of the results in the papers [1], [2] to the multivariate case with general loss function and for complete parametric uncertainty, i.e. when parameters of the mean and the covariance functions are unknown. The values of the AER are examined for various combinations of parameters for the bivariate, stationary geometric unisotropic Gaussian random field with exponential covariance function sampled on a regular 2-dimensional lattice.

References

[1] Dučinskas, K. (2009). Approximation of the expected error rate in classification of the Gaussian random field observations. *Statistics and Probability Letters* **79**, 138–144.

[2] Šaltytė-Benth, J., Dučinskas, K. (2005). Linear discriminant analysis of multivariate spatial-temporal regressions. *Scand. J. Statist.* **32**, 281–294.

Estimating the time-varying parameters of stochastic differential equation by maximum principle in finance tasks

Darya Filatova

University of Kielce, Poland, email: daria_filatova@interia.pl

Keywords: estimation, maximum principle, stochastic differential equation, time-varying parameters.

An important area of financial mathematics studies the expected returns and volatilities of the price dynamics of stocks and bonds. The stochastic dynamics of stocks and bonds should be correctly specified since misspecification of a model leads to erroneous valuation and hedging. We have to admit that economic conditions change from time to time, so we assume that return and volatility depend on time as well as on price level for some stock or bond. In this case it is reasonable to use a stochastic differential equation with the time-varying parameters as the model for the description of the price dynamics.

It is not easy to describe the time-varying parameters by means of certain functional forms. Flexible models do not assume any specific form of these functions. This data-analytic approach called nonparametric regression can be found in statistical literature. However, the direct application of the ideas does not bring desired results. The improvements of the identification procedures were presented in [1, 2, 4]. The main idea of these works was based on the discretization of the stochastic differential equation and further approximation of the parameter functions by constants at the discretization points. It is clear that the accuracy of the estimates depends on the accuracy of the discretization method. To overcome this problem we propose to consider the time-varying parameters as an control functions and solve the identification task as an optimal control problem using the maximum principle [3, 5].

In the paper we present the principles of the identification method construction, show its proficiency and give some illustrations.

References

[1] Fan, J., Jiang J., Zhang Ch., Zhou Z. (2003). Time-dependent diffusion models for term structure dynamics. *Statistica Sinica* **13**, 965–992.

[2] Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics. *Statistical Sciences* **20**(4), 317–337.

[3] Filatova D., Grzywaczewski, M., Osmolovskii, N. (2010). Optimal control problem with an integral equation as the control object. *Nonlinear Analysis: Theory, Methods & Applications* **72**(3-4), 1235–1246

[4] Hurn, A.S., Lindsay, K.A., Martin, V.L. (2003). On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. *Journal of Time Series Analysis* **24**, 45–63.

[5] Milyutin, A.A., Osmolovskii, N.P. (1998). Calculus of Variations and Optimal Control. *Translations of Mathematical Monographs* **180**, American Mathematical Society.

On the E-optimality of complete designs under a mixed interference model

Katarzyna Filipiak

Poznań University of Life Sciences, Poland, email: kasfil@up.poznan.pl

Keywords: E-optimal designs, information matrix, mixed interference model, spectral norm.

In the experiments in which the response to a treatment can be affected by other treatments, the interference model with neighbor effects is usually used. It is known, that circular neighbor balanced designs (CNBDs) are universally optimal under such a model, if the neighbor effects are fixed as well as random ([1], [2], [3]). However, such designs cannot exist for each combination of design parameters. In [4] it is shown that in the fixed interference model circular weakly neighbor balanced desings (CWNBDs) are universally optimal over the class of designs with the same number of treatments as experimental units per block and specific number of blocks. It is known, that neither CNBD nor CWNBD can exist if the number of blocks is $p(t-1) \pm 1$, $p \in \mathbb{N}$, with t - number of treatments. The paper [5] gave the structure of the left-neighboring matrix of E-optimal complete block designs under the model with fixed neighbor effects over the classes of designs with $p = 1$. The aim of this paper is to generalize E-optimality results for designs with $p \in \mathbb{N}$ assuming random neighbor effects.

References

[1] Druilhet, P. (1999). Optimality of circular neighbour balanced designs. *J. Statist. Plann. Infer.* **81**, 141–152.

[2] Filipiak, K., Markiewicz, A. (2003). Optimality of circular neighbor balanced designs under mixed effects model. *Statist. Probab. Lett.* **61**, 225–234.

[3] Filipiak, K., Markiewicz, A. (2007). Optimal designs for a mixed interference model. *Metrika* **65**, 369–386.

[4] Filipiak, K., Markiewicz, A. (2011). On universal optimality of circular weakly neighbor balanced designs under an interference model, (submitted).

[5] Filipiak, K., Różański, R., Sawikowska, A., Wojtera-Tyrakowska, D. (2008). On the E-optimality of complete designs under an interference model, *Statist. Probab. Lett.* **78**, 2470–2477.

Spectral properties of information matrices and design optimality

Katarzyna Filipiak and Augustyn Markiewicz

Poznań University of Life Sciences, Poland, email: amark@up.poznan.pl

Keywords: D-optimal design, E-optimal design, eigenvalues, information matrix, interference model, spectral norm.

In statistical research sometimes optimality criteria of experimental designs are formulated as functions of the eigenvalues of nonnegative definite information matrices. The aim of this paper is to characterize the information matrix via its eigenvalues. We are looking for a matrix in a given set such that its smallest nonzero eigenvalue is maximal over the smallest eigenvalues of matrices from this set. Obtained algebraic results are used to determine D-, E-, and universally optimal circular complete block designs under an interference model.

Presented results are based on the following papers: [1] - [3].

References

[1] Filipiak, K., Markiewicz, A. (2011). On universal optimality of circular weakly neighbor balanced designs under an interference model. *Comm. Statist. Theory Methods*, accepted.

[2] Filipiak, K., Markiewicz, A., Różański, R. (2011). Maximal determinant over a certain class of matrices and its application to D-optimality of designs. *Linear Algebra Appl.*, accepted.

[3] Filipiak, K., Różański, R., Sawikowska, A., Wojtera-Tyrakowska, D. (2008). On the E-optimality of complete designs under an interference model. *Statist. Probab. Lett.* **78**, 2470–2477.

Construction of bivariate survival probability functions related to 'micro-shocks' - 'micro-damages' paradigm

Jerzy K. Filus[1] **and Lidia Z. Filus**[2]

[1] *Oakton Community College, USA, email: jfilus@oakton.edu*
[2] *Northeastern Illinois University, USA, email: L-Filus@neiu.edu*

Keywords: confidentiality, disclosure risk, Metropolis algorithm, noise multiplication, prior distribution.

In searching for a proper description of various kind of stochastic dependences among random quantities considered in reliability and biomedical investigations we apply a general method of construction of bivariate probability distributions (or the corresponding joint survival function) of such quantities. High average pulse rate and/or blood pressure, excessive level of cholesterol, or other evidently dependent in magnitude levels of some chemicals in patient's body could serve as examples of such stochastically dependent quantities.

In effort to find general device for underlying stochastic dependences among these indicators we define and employ (on the physical part) the 'micro-shocks' - 'micro-damages' pattern [3] that naturally occurs in some reliability investigations. This reliability pattern can be redefined for a wider range of phenomena such as bio-medical [1], econometric, or other "realities".

In general, we consider random variables X_1, X_2 that interact with each other so that the impact of one of them on the other is mutual in the sense that each variable is an explanatory to the other.

The joint probability distribution of each such pair can model some mutual ("physical", in a very wide sense, not only in a strict sense of the physics theory) interactions.

In the 'micro-shocks' - 'micro-damages' pattern, the realizations x_i of the random variables X_i, accordingly to their sizes influence the hazard rate (or its parameter) of the other random variable X_k, $i, k = 1, 2$ and $i \neq k$. The considered method of construction allows to obtain a joint survival function

$$S(x_1, x_2) = \mathbb{P}(X_1 > x_1, X_2 > x_2)$$

of the random vector (X_1, X_2), given both marginal survival functions $\mathbb{P}(X_1 > x_1)$ and $\mathbb{P}(X_2 > x_2)$.

It turns out that in the simplest case, when both marginals are exponentially distributed, we obtain the common first bivariate exponential Gumbel distribution [5]. In some applications one can consider the method as an extension of what we call "Gumbel device" so that any (not necessarily exponential) two marginal survival functions $\mathbb{P}(X_1 > x_1)$, $\mathbb{P}(X_2 > x_2)$ of X_1, X_2 can be "joint" by what we call "Gumbel dependence factor" $\exp(-cx_1x_2)$, where parameter c is any nonnegative real and the condition $c = 0$ stands for independence.

Realize that this construction preserves given in advance, marginal distributions. Also, one can see that the above dependence factor can be generalized to a wider class of functions. For example, one may consider the Gumbel factors in the following "Weibullian form" $\exp(-cx_1^a x_2^b)$ with positive parameters a, b.

In fact, any arbitrary two continuous marginals (not necessarily from the same class of probability distributions) may "invariantly" be "connected" by a given fixed 'Gumbel factor' to "become" stochastically dependent. In reverse, a fixed pair of marginals can be connected in many different ways each corresponding to one Gumbel factor.

Moreover, the above constructions can easily be extended to higher than two dimensions.

References

[1] Collett, D. (2003). *Modeling Survival Data in Medical Research.* Chapman&Hall, London.

[2] Filus, J. K., Filus, L. Z. (2006). On some new classes of multivariate probability distributions. *Pakistan Journal of Statistics* **22**, 21–42.

[3] Filus, J. K., Filus, L. Z. (2009). 'Microshocks-Microdamages' type of system component interaction; failure models. In: *Proceedings of the 15th ISSAT International Conference on Reliability and Quality in Design*, 24–29.

[4] Filus, J. K., Filus, L. Z., Arnold, B.C. (2010). Families of multivariate distributions involving "triangular" transformations. *Communications in Statistics - Theory and Methods* **39**(1), 107–116.

[5] Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698–707.

Semi-recursive nonparametric algorithms of identification and control

Irina Foox, Irina Glukhova and Gennady Koshkin

Tomsk State University, Russia
e-mail: fooxil@sibmail.com, win32_86@mail.ru, kgm@mail.tsu.ru

Keywords: control, identification, kernel recursive estimator, mean square error.

Let a sequence $(Y_t)_{t=\dots,-1,0,1,\dots}$ be generated by a regression–autoregression

$$Y_t = \Psi(Y_{t-1}, X_t, U_t) + \Phi(Y_{t-1}, X_t, U_t)\xi_t, \tag{1}$$

where (ξ_t) is a sequence of zero mean i.i.d. random variables with unit variance, Y_t is an output variable, X_t, U_t are noncontrolled and controlled input random variables, not depending on (ξ_t), and Ψ and $\Phi > 0$ are unknown functions defined on $\mathbf{R}^3$.

Denote $Z_{t-1} = (Y_{t-1}, X_t, U_t)$. Note that for $x \in \mathbf{R}^3$ we have the conditional expectation $\mathbf{E}(Y_t|Z_{t-1} = x) = \Psi(x)$ and the conditional variance $\mathbf{D}(Y_t|x) = \Phi^2(x)$.

We presume that Assumptions 3.1 and 3.2 from [1] are fulfilled. Then, according to [1:Lemma 3.1], (Y_t) is a strictly stationary process, satisfying the strong mixing condition with a strong mixing coefficient $\alpha(\tau) \le c_0\rho_0^\tau$, $0 < \rho_0 < 1$, $c_0 > 0$. In this case, we can find the MSE of the proposed estimators as in [2].

We estimate $\Psi(x)$ by the statistic

$$\Psi_n(x) = \sum_{t=2}^{n+1} \frac{X_t}{h_t^3}\mathbf{K}\left(\frac{x - Z_{t-1}}{h_t}\right) \Bigg/ \sum_{t=2}^{n+1} \frac{1}{h_t^3}\mathbf{K}\left(\frac{x - Z_{t-1}}{h_t}\right), \tag{2}$$

where $\mathbf{K}(u) = \prod_{i=1}^{3} K(u_i)$ is a three-dimensional product-form kernel, $(h_n) \downarrow 0$ is a number sequence. The conditional variance for model (1) is estimated by a statistic similar to (2).

Consider the stabilization problem of Y_n on the level Y^*. Let Ψ be a simple continuous function. Then we can construct the following control algorithm for the given level Y^* :

$$U_n^* = \frac{\sum_{t=2}^{n+1} \frac{U_t}{h_t^2} K\left(\frac{Y^* - Y_t}{h_t}\right) K\left(\frac{Y^* - Y_{t-1}}{h_t}\right) K\left(\frac{X_{n+1} - X_t}{h_t}\right)}{\sum_{t=2}^{n+1} \frac{1}{h_t^2} K\left(\frac{Y^* - Y_t}{h_t}\right) K\left(\frac{Y^* - Y_{t-1}}{h_t}\right) K\left(\frac{X_{n+1} - X_t}{h_t}\right)}. \tag{3}$$

Simulations and empirical results based on the macroeconomic data of Russian Federation are provided.

Supported by Russian Foundation for Basic Research (project 09-08-00595-a).

References

[1] Masry, E., Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory* **11**(2), 258–289.

[2] Kitaeva, A. V., Koshkin, G. M. (2010). Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression. *Automation and Remote Control* **71**(2), 257–274. DOI: 10.1134/S0005117910020086.

Multivariate and multiple testing of hypotheses - what is preferred in the analysis of clinical trial data and why?

Ekkehard Glimm

Novartis Pharma, Basel, Switzerland

Keywords: clinical trials, confirmatory analysis, multiple testing.

While traditionally, confirmatory clinical trials often had a single univariate clinical endpoint, recent trends show a growing number of such trials with multiple endpoints. The reasons for this are an increased interest in safety parameters, improved biomarker assessment technology and an increased number of trials with active comparators as the control group, where the improvement over the existing standard-of-care is not easily characterized by a single measurement.

Regarding the confirmatory analysis of the treatment effects, we have to make a choice between multivariate and multiple hypothesis testing. This talk will review similarities and differences between the two approaches. In lower-dimensional situations, there often is an interest in the individual endpoints. Hence multiple methods that easily facilitate confirmatory statements about the individual endpoints with strong familywise error rate control (Maurer et al., 2011) are often preferred.

In higher-dimensional situations, multiple methods turn out to be too conservative. Additionally, there is usually less interest in the individual endpoints, such that the multiple testing concept of *consonance* (Gabriel, 1969) is less relevant. In this situation, the advantages of multivariate methods (Srivastava, 2002, 2009) may carry more weight.

References

[1] Gabriel, K.R. (1969). Simultaneous test procedures: some theory of multiple comparisons. *The Annals of Mathematical Statistics* **40**, 224–250.

[2] Maurer, W., Glimm, E., Bretz, F. (2011). Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research* **3**, (in press).

[3] Srivastava, M.S. (2002). *Methods of Multivariate Statistics.* Wiley, New York.

[4] Srivastava, M.S. (2009). A review of multivariate theory for high dimensional data with fewer observations. *Advances in Multivariate Statistical Methods* **9**, 25–52.

On the canonical correlation analysis of bi-allelic genetic markers

Jan Graffelman

Universitat Politècnica de Catalunya, Spain, email: jan.graffelman@upc.edu

Keywords: biplot, generalized inverse, Hardy-Weinberg equilibrium, linkage disequilibrium.

Multivariate analysis is becoming increasingly relevant in genetics, due to the automated generation of large databases of genetic markers, single nucleotide polymorphisms (SNPs) in particular. Most SNPs are bi-allelic, and individuals can be characterized generically as AA, AB or BB. Such genotype data can be coded in an indicator matrix. Additional indicators can be defined to indicate whether an individual is a carrier or a non-carrier of a particular allele.

Genetic markers are usually expected to be in Hardy-Weinberg equilibrium which can be assessed by a chi-square or exact test. Such a test concerns the correlation between the two indicators for the *same* marker (within marker correlation).

Correlation between two *different* markers is referred to as linkage disequilibrium in genetics. If the data is represented by indicator matrices, then linkage disequilibrium can be studied by a canonical correlation analysis of two indicator matrices. Generalized inverses can be used to cope with the singularity of covariance matrices. By using the carrier-indicators as supplementary variables, such a canonical analysis is also informative about Hardy-Weinberg equilibrium. Biplots [2] can be used to visualize the results.

In the light of the larger number of markers obtained in genotyping studies, Carroll's [1] generalized canonical correlation analysis can be used to study multiple markers simultaneously.

The various forms of the canonical analysis of genetic markers will be illustrated with several examples in the talk.

References

[1] Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 67th Annual Convention of the American Psychological Association*, 227–228.

[2] Graffelman, J. (2005). Enriched biplots for canonical correlation analysis. *Journal of Applied Statistics* **32**(2), 173–188.

Biplot videos

Michael Greenacre

Universitat Pompeu Fabra, Barcelona, Catalunya email: michael.greenacre@upf.edu

Keywords: biplot, matrix product, regression, singular-value decomposition, triplot.

Most multivariate statistical methods that are used in practice have a common theory of matrix products – such methods include multiple regression, principal component analysis, correspondence analysis, log-ratio analysis, linear discriminant analysis, canonical correlation analysis, as well as several constrained variants of these methods which mix rank reduction with linear constraints, for example redundancy analysis and canonical correspondence analysis. Where there is a matrix product, there is a *biplot*, a type of multivariate scatterplot that graphically represents two sets of objects – usually cases and variables – in a common vector space. In the linearly constrained versions, the constraining variables can be added to the biplot to obtain what is often called a "triplot".

For a couple of years I have been experimenting with dynamic graphics in statistics, producing video animations of models, algorithms and results. The article by Greenacre and Hastie (2010) is a first product of this work, containing four videos embedded in the article where there would otherwise be static figures. The videos illustrate much more clearly the models and results of the complex statistical analyses presented in the article. Other articles with video content as supplementary material are by Greenacre (2010a, 2011).

Mainly as a complement to my book "Biplots in Practice" (Greenacre, 2010b) I have been developing a series of video animations, not only as an educational tool but also opening up new ways of understanding and interpreting multivariate statistical results. In this talk I will take you on a moving-picture journey from the simplest biplot, based on multiple regression, through several illustrations of other well-known multivariate methods, and finally the canonical correspondence analysis of a large ecological data set, including hundreds of cases and hundreds of dependent and independent variables.

References

[1] Greenacre, M. (2010a). Correspondence analysis of raw data. *Ecology* **91**, 958–963.

[2] Greenacre, M. (2010b). *Biplots in Practice*. BBVA Foundation, Madrid. Freely downloadable from `http://www.multivariatestatistics.org`.

[3] Greenacre, M. (2011). Contribution biplots. Working paper 1162, Department of Economics and Business, Pompeu Fabra University. Downloadable from `http://www.econ.upf.edu/en/research/onepaper.php?id=1162`.

[4] Greenacre, M., Hastie, T. (2010). Dynamic visualization of statistical learning algorithms in the context of high-dimensional textual data. *Journal of Web Semantics* **8**, 163–168.

Minimax decision for solution of the problem of aircraft and airline reliability processing results of acceptance full-scale fatigue test of airframe

Maris Hauka and Yuri Paramonov

Aviation Institute, Riga Technical University, Latvia, email: maris.hauka@gmail.com, yuri.paramonov@gmail.com

Keywords: inspection program, Markov chains, minimax, reliability.

Probability of Failure (PF) of fatigue-prone AirCraft (AC) and Failure Rate (FR) of AirLine (AL) for specific inspection program can be calculated using Markov Chains (MC) and Semi-Markov Process (SMP) theory if parameters of corresponding models are known. Exponential approximation of fatigue crack size growth function, $a(t) = a_0 \exp(Qt)$, where a_0, Q are random variables, is used. Estimation of the parameters of the distribution functions of these variables and the choice of final inspection program under condition of limitation of PF and FR can be made using results of observation of some random fatigue crack in full-scale fatigue test of the airframe. For processing of acceptance type test, when redesign of new aircraft should be made if some reliability requirements are not met, the minimax decision is used. The process of operation of AC is considered as absorbing MC with $(n + 4)$ states. The states $E_1, E_2, ..., E_{n+1}$ correspond to AC operation in time intervals $[t_0, t_1), [t_1, t_2), ..., [t_n, t_{SL})$, where n is an inspection number, t_{SL} is specified life (SL), i. e. AC retirement time. States E_{n+2}, E_{n+3}, and E_{n+4} are absorbing states: AC is descarded from service when the SL is reached or fatigue failure (FF), or fatigue crack detection (CD) takes place. In corresponding matrix for operation process of AL the states E_{n+2}, E_{n+3} and E_{n+4} are not absorbing but correspond to return of MC to state E_1(AL operation returns to first interval). In the matrix of transition probabilities of AC, P_{AC}, there are three units in three last lines in diagonal, but for corresponding lines in matrix for AL, P_{AL}, the units are in the first column, corresponding to state E_1. Using P_{AC} we can get the probability of FF of AC and the cumulative distribution function, mean and variance of AC life. Using P_{AL} we can get the stationary probabilities of AL operation $\{\pi_1, ..., \pi_{n+1}, \pi_{n+2}, ..., \pi_{n+4}\}$. Here π_{n+3}defines the part of MC steps, when FF takes place and MC appears in state E_{n+3}. The FR, λ_F , and the gain of this process, g, are calculated using the theory of SMP with reword, taking into accout the reword of succesful operation in one time unit, the cost of acquisition of new AC after SL, FF or CD take place,... If the gain is measured in time unit then $L_{n+3} = g/\pi_3$ is a mean time between FF; the intensity of fatigue failure $\lambda_F = 1/L_{n+3}$. The problem of inspection planning is the choice of the sequance $\{t_1, t_2, ..., t_n, t_{SL}\}$ corresponding to maximum of gain under limitation of AC intensity of fatigue failure. In a numerical example the minimax decision, based on observation of some fatigue crack during acceptance full-scale fatigue test of airframe, is considered.

Markov chain properties in terms of column sums of the transition matrix

Jeffrey Hunter

Auckland University of Technology, email: jeffrey.hunter@aut.ac.nz

Keywords: column sums, generalized matrix inverses, Kemeny constant, Markov chains, mean first passage times, stochastic matrices, stationary distributions.

Questions are posed regarding the influence that the column sums of the transition probabilities of a stochastic matrix (with row sums all one) have on the stationary distribution, the mean first passage times and the Kemeny constant of the associated irreducible discrete time Markov chain. Some new relationships, including some inequalities, and partial answers to the questions, are given using a special generalized matrix inverse that has not previously been considered in the literature on Markov chains.

Multivariate exponential dispersion models

Bent Jørgensen

University of Southern Denmark, Denmark, email: bentj@stat.sdu.dk

Keywords: convolution method, exponential dispersion model, multivariate gamma distribution, multivariate generalized linear model, multivariate Poisson distribution.

In order to develop a general approach for analysis of non-normal multivariate data, it would be desirable to obtain a simple-minded framework that can accommodate a wide variety of different types of data, much like generalized linear models do in the univariate case. There is no shortage of multivariate distributions available, but the main stumbling block so far has been the lack of a suitable multivariate form of exponential dispersion model.

In the univariate case, an exponential dispersion model $\mathrm{ED}(\mu, \sigma^2)$ is a two-parameter family parametrized by the mean μ and dispersion parameter σ^2, with variance $\sigma^2 V(\mu)$, where V denotes the unit variance function. The generalized linear models paradigm is based on combining a link function with a suitable linear model. Estimation uses quasi-likelihood for the regression parameters, and the Pearson statistic for estimating the dispersion parameter.

We consider a new k-variate exponential dispersion model $\mathrm{ED}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ aimed at providing a fully flexible covariance structure corresponding to a mean vector $\boldsymbol{\mu}$ and a positive-definite dispersion matrix $\boldsymbol{\Sigma}$. The covariance matrix is of the form $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma} \odot \boldsymbol{V}(\boldsymbol{\mu})$, where $\odot$ denotes the Hadamard (elementwise) product between two matrices, and $\boldsymbol{V}(\boldsymbol{\mu})$ denotes the (matrix) unit variance function. We consider a multivariate generalized linear model for independent response vectors $\boldsymbol{Y}_i \sim \mathrm{ED}_k(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ defined by $g(\boldsymbol{\mu}_i^\top) = \boldsymbol{x}_i \boldsymbol{B}$, where the link function g is applied coordinatewise to $\boldsymbol{\mu}_i^\top$, $\boldsymbol{x}_i$ is an m-vector of covariates, and $\boldsymbol{B}$ is an $m \times k$ matrix of regression coefficients. We estimate the regression matrix $\boldsymbol{B}$ using a quasi-score function, and we estimate the dispersion matrix $\boldsymbol{\Sigma}$ using a multivariate Pearson statistic defined as a weighted sum of squares and cross-products matrix of residuals. This model specializes to the classical multivariate multiple regression model when g is the identity function and $\mathrm{ED}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution.

The construction of the multivariate exponential dispersion model $\mathrm{ED}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is based on an extended convolution method, which makes the marginal distributions follow a given univariate exponential dispersion model. We illustrate the method by considering multivariate versions of the Poisson and gamma distributions, and discuss some of the challenges faced in the implementation of the method.

Residuals of a linear model for correlated data with measurement errors

Ants Kaasik

University of Tartu, Estonia, email: ants.kaasik@ut.ee

Keywords: correlated observations, errors-in-variables model, phylogenetic analysis.

We consider the case, where a linear model has been set up for variables Y and $X_1, \ldots, X_n$. All the variables in the model are observed with error. When calculating the residuals for such a model, errors in the X variables will have important consequences that cannot be ignored when performing further analyses with these residuals. I show that this can be thought of as measurement errors carrying over to the residuals and the process is analyzed in detail.

Such a model is quite typical in phylogenetic analyses (analysis where traits measured for a species correspond to a sample element and the sample is treated as correlated data because of the shared evolution of the species) in a situation where the relation between two traits is sought and one (or both of them) need to be corrected for the value(s) of some other trait(s). See e.g. [1] and [2] for examples.

References

[1] Revell, L. J. (2009). Size-correction and principal components for interspecific comparative studies. *Evolution* **63**(12), 3258–3268.

[2] Ives, A. R., Midford, P. E., Garland, T. Jr. (2007). Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology* **56**(2), 252–270.

K-nearest neighbors as pricing tool in insurance

Raul Kangro and Kalev Pärna

University of Tartu, Estonia, email: raul.kangro@ut.ee, kalev.parna@ut.ee

Keywords: curse of dimensionality, distance measures, feature selection, k-nearest neighbors, local regression, premium calculation, supervised learning.

The method of k-nearest neighbors (k-NN) is recognized as a simple but powerful toolkit in statistical learning [1], [2]. It can be used both in discrete and continuous decision making known as classification and regression, respectively. In the latter case the k-NN is aimed at estimation of conditional expectation $y(\mathbf{x}) := E(Y|X = \mathbf{x})$ of an output Y given the value of an input vector $\mathbf{x} = (x_1, \dots, x_m)$. In accordance with supervised learning set-up, a training set is given consisting of n pairs $(\mathbf{x_i}, y_i)$ and the problem is to estimate $y(\mathbf{x})$ for a new input $\mathbf{x}$. This is exactly the situation in insurance where the pure premium $y(\mathbf{x})$ for a new client (policy) $\mathbf{x}$ is to be found as conditional mean of loss. Typically the data do not contain any other record with the same $\mathbf{x}$, thus the other data points have to be used in order to estimate $y(\mathbf{x})$. Using the k-NN methodology, one first finds a neighborhood $U_{\mathbf{x}}$ consisting of k samples which are nearest to $\mathbf{x}$ w.r.t a given distance measure d. Secondly, the (weighted) average of Y is calculated over the neighborhood $U_{\mathbf{x}}$ as an estimate of $y(\mathbf{x})$:

$$\hat{y}(\mathbf{x}) := \frac{1}{\sum_{i \in U_{\mathbf{x}}} \alpha_i} \sum_{i \in U_{\mathbf{x}}} \alpha_i \cdot Y_i,$$

where the weights α_i are chosen so that the nearer neighbors contribute more to the average than the more distant ones. We use the distance between the instances $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ in the form

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{m} w_j \cdot d_j(x_{ij}, x_{i'j}),$$

where w_j is the weight of the feature j and $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ (and a zero-one type variable for categorical features).

We address the following key issues related to k-NN method: feature weighting (w_j), distance weighting (α_i), determining the optimum value of the smoothing parameter k. We propose a three-step multiplicative procedure to define w_j which consists of 1) normalization (eliminating the scale effect), 2) accounting for statistical dependence between the feature j and Y, 3) feature selection to obtain a subset of features that performs best. All our optimization procedures are based on cross-validation techniques. The so-called 'curse of dimensionality' is effectively handled by our feature selection process which optimizes the dimension of input.

Finally, comparisons with other methods for estimation of the regression function $y(x)$ (CART, generalized linear regression, use of model distributions) are drawn, which demonstrate high competetiveness of the k-NN method. The conclusions are based on the analysis of a real data set.

References

[1] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[2] Mitchell, T.M. (2001). *Machine-Learning.* McGraw-Hill.

Multivariate model with a Kronecker product covariance structure: general linear model

Pavel A. Kashitsyn

Moscow State University, Russia, email: pavel.kash@gmail.com

Keywords: covariance structure, Kronecker product, multivariate linear model.

Consider a $(p \times n)$-matrix $\mathbf{X} = (X_1, \ldots, X_n)$, where a (pn)-vector $vec\,(\mathbf{X}) = \left(X_1^T, \ldots, X_n^T\right)^T$ is normally distributed with the positive definite covariance matrix Λ and the mean vector $vec\,(\mathbf{M}) = \left(M_1^T, \ldots, M_n^T\right)^T$. Suppose that Λ follows the Kronecker product covariance structure, that is $\Lambda = \Psi \otimes \Sigma$, where $\Psi = (\psi_{ij})$ is an $(n \times n)$-matrix and $\Sigma = (\sigma_{ij})$ is a $(p \times p)$-matrix and the matrices Ψ, Σ are supposed to be positive definite. Such model is considered in [2], where the maximum likelihood estimates (MLE) of the parameters $\mathbf{M}, \Psi, \Sigma$ are obtained. A special case of this model is an intraclass correlation structure model which is considered in [1] and [3].

We consider the general linear model which follows a Kronecker product covariance structure and we want to test the linear hypothesis

$$H : E\mathbf{X} \in \mathcal{L},$$

where $\mathcal{L}$ is a linear subspace. The main result of the research is the following theorem.

Theorem (orthogonal decomposition).
Let $\mathbf{X} = \|X_1, X_2, \ldots, X_n\|$ be a $(p \times n)$-matrix which follows a Kronecker product covariance structure and $Cov\,(X_i, X_j) = \psi_{ij}\Sigma$, $i, j = \overline{1, n}$, and let matrices Ψ, Σ be positive definite. Let $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m$ be mutually orthogonal submodules of the module $\mathbb{R}_n^p$ over a ring $\mathbb{R}_p^p$ and let a direct sum of these modules be $\mathbb{R}_n^p$:

$$\mathbb{R}_n^p = \mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \ldots \oplus \mathcal{L}_m.$$

Consider a decomposition of a $(p \times n)$-matrix $\mathbf{X}$ on mutually orthogonal components

$$\mathbf{X} = proj_{\mathcal{L}_1}\mathbf{X} + proj_{\mathcal{L}_2}\mathbf{X} + \ldots + proj_{\mathcal{L}_m}\mathbf{X}.$$

Then

a) random $(p \times n)$-matrices $proj_{\mathcal{L}_1}\mathbf{X}$, $proj_{\mathcal{L}_2}\mathbf{X}, \ldots, proj_{\mathcal{L}_m}\mathbf{X}$ are independent and $E proj_{\mathcal{L}_i}\mathbf{X} = proj_{\mathcal{L}_i}E\mathbf{X}$, where projections on submodules $\mathcal{L}_i$, $i = \overline{1, m}$ are determined by a bilinear form Ψ^{-1}.

b) $\left(proj_{\mathcal{L}_i}\mathbf{X}\right)\left(proj_{\mathcal{L}_i}\mathbf{X}\right)^T \sim W_p\left(dim\mathcal{L}_i, \Sigma, \Delta_i\right)$ – a noncentral Wishart distribution with a noncentrality parameter $\Delta_i = \left(proj_{\mathcal{L}_i}E\mathbf{X}\right)\left(proj_{\mathcal{L}_i}E\mathbf{X}\right)^T$. Here projections on submodules $\mathcal{L}_i$, $i = \overline{1, m}$ are determined by a bilinear form Ψ^{-1}.

References

[1] Dinesh, S. B. (1987). Testing hypothesis on the mean vector under an intraclass correlation structure. *Biometrical Journal* **7**, 783–789.

[2] Srivastava, M. S., Nahtman, T., von Rosen, D. (2008). Models with a Kronecker Product Covariance Structure: Estimation and Testing. *Mathematical Methods of Statistics* **17**(4), 357–370.

[3] Srivastava, M.S. (1984). Estimation of intraclass correlations in familial data. *Biometrika* **71**, 177–185.

The asymptotic results for nearly critical branching processes with immigration

Yakubdjan M. Khusanbaev and Gayrat M. Rakhimov

Institute of Mathematics and Information Technology, Tashkent, Uzbekistan,
email: yakubjank@mail.ru, gairat48@gmail.com

Keywords: confidentiality, disclosure risk, Metropolis algorithm, noise multiplication, prior distribution.

Let $\{\xi_{k,j}^{(n)},\ k,j \in \mathbb{N}\}$ and $\{\epsilon_k^{(n)},\ k \in \mathbb{N}\}$ be two independent sequences of non-negative integer-valued and identically distributed random variables for every $n \in \mathbb{N}$. For $n \in \mathbb{N}$ we define a sequence of random variables recursively:

$$X_0^n = 0, \quad X_k^n = \sum_{j=1}^{X_{k-1}^n} \xi_{k,j}^{(n)} + \epsilon_k^{(n)},\ k \in \mathbb{N}.$$

The sequence $\{X_k^n\ k \in \mathbb{N}\}$ is called a branching process with immigration [1]. We assume that $m_n = \mathbb{E}(\xi_{1,1}^{(n)})^2 < \infty$ and $\mathbb{E}(\epsilon_1^{(n)})^2 < \infty$ for all $n \in \mathbb{N}$. The branching process with immigration is called nearly critical if $m_n \to 1$ as $n \to \infty$.

In the papers [2]–[4] asymptotic behavior of the process $X_{[nt]}^n,\ t > 0$ has been investigated in the case $m_n = 1 + \alpha d_n^{-1} + O(d_n^{-1}),\ \alpha \in \mathbb{R}$ as $n \to \infty$, where d_n is a sequence of positive numbers such that $nd_n \to c$ as $n \to \infty$. In this paper we investigate asymptotic behavior of the random process $X_{[nt]}^n,\ t > 0$ when $nd_n \to \infty$ as $n \to \infty$ and prove limit theorems for $X_{[nt]}^n,\ t > 0$. We remark that the obtained results are different from the results in the case $m_n = 1 + \alpha n^{-1} + o(n^{-1})$.

References

[1] Athreya, K.B., Ney P.E. (1972). *Branching processes.* Springer-Verlag, Berlin.

[2] Sriram, T.N. (1994). Invalidity of bootstrap for immigration critical branching process with immigration. *Ann. Statist.* **22**, 1013–1023.

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

[4] Khusanbaev Ya.M. (2009). The convergence of Galton-Watson branching processes with immigration to a diffusion process. *Theory Probab. Math. Statist.* **79**, 179–185.

Recursive kernel estimators of a non-homogeneous Poisson process normalized intensity function and its derivative

Anna V. Kitaeva[1] **and Mihail V. Kolupaev**[2]

[1] *Tomsk Polytechnic University, Tomsk, Russia, email: kit1157@yandex.ru*
[2] *Tomsk State University, Tomsk, Russia, email: al13n@sibmail.com*

Keywords: kernel estimates, mean-square convergence, normalized intensity function, Poisson process.

Poisson processes are known to be useful to model several random phenomena, and the problem of estimation of Poisson intensity functions arises in many diverse areas such as communications, meteorology, insurance, medical sciences, seismology [1, 2]. In the present paper we do not assume any parametric form of the intensity function except some regularity conditions, and suppose that only a single realization of the process is available at time T. We give an extension of the results considered by Kitaeva in [3] to the recursive algorithms. Recursive estimation is particularly useful in large sample size since the result can be easily updated with each additional observation.

Let $\{t_i, i = \overline{1, N}, 0 \leq t_i \leq T\}$ be a realization of a Poisson process having unknown intensity function $\lambda(\cdot)$, where N is the number of points falling into a fixed interval $[0, T]$, $\Lambda(a, b) = \displaystyle\int_a^b \lambda(t)dt$, $K^{(1)}(\cdot)$ be the derivative of the function $K(\cdot) \equiv K^{(0)}(\cdot)$. The density $S(\cdot) = \lambda(\cdot)/\Lambda(0, T)$, that we call normalized intensity function, and its derivative at a point $t \in [0, T]$ are estimated by the following expressions

$$S_N^{(r)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_i^r}K^{(r)}\left(\frac{t-t_i}{h_i}\right) = \left(1-\frac{1}{N}\right)S_{N-1}^{(r)} + \frac{1}{Nh_N^r}K^{(r)}\left(\frac{t-t_N}{h_N}\right)$$

where $r = 0, 1$, $h_n \downarrow 0$ is a sequence of real numbers (bandwidths), $K(\cdot)$ is a kernel function; with the convention that $S_0^{(r)} = 0$. Distinctive feature of the statistics under consideration is due to a random sample size.

Mean-square convergence is proved in a scheme of series under unlimited increasing of the intensity. Simulation studies are carried out to illustrate the convergence and to compare the proposed recursive and the non-recursive estimates.

References

[1] Kingman, J.F.C. (1993). *Poisson Processes.* Clarendon, Oxford.

[2] Daley, D.J., Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes.* Springer, New York.

[3] Kitaeva, A. (2008). Mean-square convergence of a kernel type estimate of the intensity function of an inhomogeneous Poisson process. In: *The Second International Conference 'Problems of Cybernetics and Informatics'. Proceedings* III, 149–152.

Simultaneous confidence region for ρ and σ^2 in the growth curve model with uniform correlation structure

Daniel Klein and Ivan Žežula

Šafárik University, Košice, Slovakia, email: daniel.klein@upjs.sk, ivan.zezula@upjs.sk

Keywords: growth curve model, interclass correlation structure, multivariate linear model, uniform correlation structure.

The basic model we consider is the standard growth curve model with uniform correlation structure:

$$Y = XBZ' + \mathbf{e}, \quad \mathrm{vec}(\mathbf{e}) \sim N\left(0, \Sigma \otimes I_n\right), \quad \Sigma = \sigma^2\left((1-\rho)I_p + \rho\mathbf{1}\mathbf{1}'\right).$$

Here $Y_{n\times p}$ is a matrix of independent p-variate observations, $X_{n\times m}$ is an ANOVA design matrix, $Z_{p\times r}$ is a regression variables matrix, and $\mathbf{e}_{n\times p}$ is a matrix of random errors. As for the unknown parameters, $B_{m\times r}$ is an location parameters matrix, and σ^2, ρ are (scalar) variance parameters. The vec operator stacks elements of a matrix into a vector column-wise.

This model has obtained increasing attention, since it allows to keep the number of variance parameters low even in high dimensional models, and its assumptions are in many situations close to reality. Even if estimators proposed by Žežula (see [2]) and Ye & Wang (see [4]) seem to be quite different, they are identical. Žežula & Klein found their marginal distributions, see [3]. We will present the joint distribution of the two estimators and investigate simultaneous confidence regions for both variance parameters.

References

[1] Klein, D., Žežula, I. (2007). On uniform correlation structure. In: *Mathematical Methods In Economics And Industry, Herany, Slovakia, proceedings*, 94–100.

[2] Žežula, I. (2006). Special variance structures in the growth curve model. *Journal of multivariate analysis* **97**(3), 606–618.

[3] Žežula, I., Klein, D. (2010). Orthogonal decompositions in growth curve models. *Acta et Commentationes Universitatis Tartuensis de Mathematica* **14**, to appear.

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

Life in statistics – Muni S. Srivastava 75

Tõnu Kollo

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

Keywords: high-dimensional analysis, multivariate analysis, quality control, sequential analysis.

On 20 January 2011 Muni Shanker Srivastava turned 75. His magnificent career in statistics started in India. Muni was born in Gonda, Uttar Pradesh and graduated from Lucknow University in India with Bachelor degree in 1955 and Master Degree in 1957. He got his PhD from Stanford University in 1964 under the supervision of Professor Charles Stein. In 1963 he joined the University of Toronto where he found his academic home; from 1972 to 2001 he worked as a full professor and from 2001 he is Professor Emeritus. After the first monographs on multivariate statistics in late 1950-ies there was a long break in publishing new books on Multivariate Analysis. A big step forward was made in 1979 when the monograph "An Introduction to Multivariate Statistics" by M. S. Srivastava and C. G. Khatri came out from North Holland. This book and the following ones: Srivastava & Carter (1983), Sen & Srivastava (1990) and Srivastava (2002) are on desks of any statistician working in multivariate statistics. But these books do not cover wide range of Muni's research activities at all.

Sequential analysis became his first attraction after PhD studies but in early seventies he turned his interests to the multivariate analysis. This fortunate turn resulted in the aformentioned book in 1979. Beside development of asymptotic methods robustness became another important issue in his research. In late eighties quality control theory attracted him. Research in this area gave several principal new analytic results. In 2002 his achievements in multivariate statistics were crowned with the Gold Medal by the Statistical Society of Canada.

What has kept Muni's mind busy in last years? High-dimensional analysis, where the number of variables can exceed the sample size and where most of the classical test statistics are not applicable, has become his new field of interest.

Nineteen supervised PhD dissertations, 177 published papers – these lists are still open.

References

[1] Srivastava, M. S., Khatri, C. G. (1979). *An Introduction to Multivariate Statistics.* North Holland, New York.

[2] Srivastava, M. S., Carter, (1983). *An Introduction to Applied Multivariate Statistics.* North Holland, New York.

[3] Sen, A. K., Srivastava, M. S. (1990). *Regression Analysis, Theory, Methods & Applications.* Springer-Verlag, New York.

[4] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* Wiley, New York.

A modified principal component test for high-dimensional data

Siegfried Kropf[1], **Guo-Chun Ding**[2], **Holger Heuer**[2] **and Kornelia Smalla**[2]

[1] *University of Magdeburg, Germany, email: siegfried.kropf@med.ovgu.de*
[2] *Julius Kühn-Institut Braunschweig, Germany, email: guo-chun.ding@jki.bund.de, holger.heuer@jki.bund.de, kornelia.smalla@jki.bund.de*

Keywords: high-dimensional data, principal component test, rotation test.

Modern biochemical analysis techniques often deliver high-dimensional observation vectors, while only small sample sizes are feasible. As an example we consider a microarray (PhyloChip) data set for comparing the bacterial community structures in the rhizosphere of three potato cultivars grown at two sites (cf. to [4] for details).

In [3], Läuter and colleagues proposed a PC test that calculates the principal components from the total sums and squares and cross products matrix $\mathbf{W}$ of the data and carries out a test on the basis of the low-dimensional principal components. For multivariate normal data this yields left-spherically distributed components and hence an exact multivariate test for the usual multivariate test statistics. Another proposal for an exact multivariate test in this situation is the 50-50-MANOVA test by Langsrud in [1].

For extreme relations of sample size n and number of variables p (n=18 and p=2432 in the example), however, there arises a problem regarding the power. The sample size restricts the number q of principal components enclosed in the test. But omitting essential components may yield a loss of power.

Therefore, we use a modified test statistic

$$\tilde{F} = \frac{\left(\sum_{i=1}^{q} \lambda_i h_{ii}\right)/\nu_h}{\left(\sum_{i=1}^{q} \lambda_i g_{ii}\right)/\nu_g}\ ,$$

where the λ_i are the eigenvalues of $\mathbf{W}$, h_{ii} and g_{ii} are the hypothesis and residual related sums of squares, and ν_h and ν_g are the corresponding degrees of freedom as one would use in a univariate test. This test statistic does no longer follow an F-distribution under the null hypothesis, but one can find a Satterthwaite approximation and one can still use properties of left-spherically distributed data to derive an exact test on the basis of rotation tests as introduced by Langsrud in [2].

The power of the resulting tests is compared in the example and in simulation studies, demonstrating the good performance of the new test in this high-dimensional setting.

References

[1] Langsrud, Ø. (2005). 50-50-Multivariate analysis of variance for collinear responses, *The Statistician* **51**, 305–317.

[2] Langsrud, Ø. (2005). Rotation tests, *Statistics and Computing* **15**, 53–60.

[3] Läuter, J., Glimm, E., Kropf, S. (1996). New multivariate tests for data with an inherent structure, *Biometrical Journal* **38**, 5–23.

[4] Weinert, N., Piceno, Y., Ding, G.-C., Meinecke, R., Heuer, H., Berg, G., Schloter, M., Andersen, G., Smalla, K. (2011). PhyloChip hybridization uncovered an enormous bacterial diversity in the rhizosphere of different potato cultivars: many common and few cultivar-dependent taxa, *FEMS Microbiology Ecology* **75**, 497–506.

Limited expected value function and its applications in insurance mathematics

Meelis Käärik and Heidi Kadarik

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

Keywords: distribution fitting, insurance principles, limited expected value function.

One of the common problems in insurance mathematics is that we usually do not see the actual loss variable but certain truncated version of it: the claims payments are limited by the sum insured, reinsurance treaties limit the actual claim size for initial insurer, also (fixed amount) deductibles set limits for policy holders, etc. In all these cases a function called *limited expected value function*, defined by

$$E[X;x] = E\min(X,x),$$

where X is a random variable (claim size), plays an important role. There are many well-known characteristics in insurance that are calculated using this function, which motivated us to study this topic more closely. We reveal some essential properties of this function and describe some important practical applications where it is used.

We also introduce the method of limited expected value function for measuring the goodness of fit between empirical and theoretical distributions. This is one of the many uses of the limited expected value function and it suits particularly well to the insurance data as it can also take into account censored data (if necessary). Also, this method can be used as an alternative or additional tool in case the data is complex and other goodness of fit tests do not give reliable results. The main disadvantage of this method is that the behavior of corresponding test statistic is not thoroughly studied, there are no certain criteria to tell us when the value of this statistic is good enough to say that a proposed distribution fits empirical data well. This problem is of our special interest, several simulations with different distributions are carried out to find out the reference values.

This research is supported by Estonian Science Foundation Grant No 7313.

On estimation of loss distributions and risk measures

Meelis Käärik and Anastassia Žegulova

Institute of Math. Statistics, University of Tartu, Estonia, email: meelis.kaarik@ut.ee, anastassia.zhegulova@ut.ee

Keywords: estimation of distributions, risk measures, theory of extreme values.

Estimation of certain loss distribution and analyzing its properties is a key issue in several finance-mathematical and actuarial applications. A special interest usually lies on the tail of the loss distribution, which allows us to answer important questions related to solvency of the insurer. It is common to apply the tools of extreme value theory and generalized Pareto distribution in problems related to heavy-tailed data (see, e.g., Coles, 2001, McNeil et al., 2005), this also provides conservative estimates for certain risk measures such as value at risk and expected shortfall.

Our main goal is to study third party liability claims data obtained from Estonian Traffic Insurance Fund (ETIF) where the observation period is one year (from mid 2006 to mid 2007). The data is quite typical for insurance claims containing very many observations and being heavy-tailed. In our approach the fitting consists of two parts: for main part of the distribution we use log-normal fit (which was the most suitable based on our previous studies) and generalized Pareto distribution is used for the tail. Main emphasis of the fitting techniques is on the proper threshold selection. We examine a wide range of thresholds and seek for stability of parameter estimates, compare the mean residual life plots and study the behaviour of risk measures at fixed tresholds. Additionally we compare our model with composite lognormal-Pareto model in respect of risk measures proposed by Cooray and Ananda (2005).

The work is supported by Estonian Science Foundation Grant No 7313.

References

[1] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer, London.

[2] McNeil, A., Frey, R., Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press, Princeton.

[3] Cooray, K., Ananda, M.M.A. (2005). Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal* **5**, 321–334.

Estimability and restrictions in linear models

Lynn R. LaMotte

Louisiana State University Health Sciences Center – New Orleans

Keywords: linear conditions, testable hypotheses.

In a linear model $X\beta$ for the mean vector of a multivariate random variable Y, estimable functions of β are customarily defined as those functions $P'\beta$ for which unbiased linear estimators exist. Testable hypotheses about β, on the other hand, are linear relations that restrict the model.

Statistical computing packages, like SAS and ANOVA programs in R, refuse to deal with non-estimable functions. That an estimator is not unbiased doesn't seem to be so fatal a flaw that we should be forbidden even to look at it. Then why is non-estimability so bad?

By examining the relation between estimability and restrictions, this paper will show that the mean vector carries no information at all about non-estimable functions, and that therefore any statement about a non-estimable function based on the mean vector is false. As part of this development, useful representations of linear restrictions on affine sets are shown.

Generalizations of Ward's method, or, norms and netball - what's the connection?

Alan Lee and Bobby Willcox

University of Auckland, email: lee@stat.auckland.ac.nz

Keywords: distance matrix, Minkowski distance, Ward's method.

In this talk, we consider several generalizations of the popular Ward's method for agglomerative hierarchical clustering. Our work was motivated by clustering software, such as the R function `hclust`, which accepts a distance matrix as input and applies Ward's definition of inter-cluster distance to produce a clustering. The standard version of Ward's method uses squared Euclidean distance to form the distance matrix. We explore the properties and effect on the clustering of using other definitions of distance, such as the Minkowski distance and powers of the Minkowski distance. We explore the effect of these on several examples and find that using powers of the Manhattan metric is particularly effective.

Memory properties of aggregated autoregressive processes and fields

Remigijus Leipus

Vilnius University, email: remigijus.leipus@mif.vu.lt

Keywords: aggregation, autoregressive process, autoregressive field, spectral density.

The aggregation of the first order autoregressive models, AR(1), with random coefficients is investigated. We study two cases of the underlying model:

(i) autoregressive sequence on $\mathbf{Z}$;

(ii) autoregressive field on $\mathbf{Z}^2$ with various configurations of neighbors.

Asymptotics of the spectral density of the resulting random process or field is studied. We show that, depending on the law of the AR coefficients, the aggregated process/field can exhibit short or long memory structure.

Estimation under restrictions built upon biased initial estimators

Natalja Lepik

University of Tartu, Estonia, email: natalja.lepik@ut.ee

Keywords: restriction estimator, survey sampling.

The users of official statistics often require that sample-based estimates satisfy certain restrictions. In the domain's case it is required that the estimates of domain totals sum up to the population total or to its estimate. For example, in time domains, quarterly estimates have to sum up to the yearly total. The relationships holding for the true population parameters do not necessarily hold for the respective estimates. This inconsistency of estimates is annoying for statistics users. On the other hand, known relationships between population parameters is a kind of auxiliary information. Involving this information into estimation process presumably improves estimates. Our goal is to define consistent domain estimators that are more accurate than the initial inconsistent domain estimators.

One solution to the problem of finding estimates under restrictions is the general restriction estimator (GR) proposed by Knottnerus (2003). His estimator is based on the unbiased initial estimators and is unbiased itself. The advantage of the GR estimator is the variance minimizing property in a class of linear estimators. Sõstra (2007) has developed the GR estimator for estimating domain totals under summation restriction. Optimality property of the domain GR estimator is studied in Sõstra and Traat (2009). In all these works, the unbiasedness or asymptotic unbiasedness of initial estimators is assumed.

It is well known that there are many useful estimators that are biased. For example, the model-based small area estimators are design-biased. The synthetic estimator can be biased on the domain level. Even the widely used GREG estimator is only asymptotically unbiased. In this thesis we will allow the vector of initial estimators $\hat{\boldsymbol{\theta}}$ to be biased, and will construct three new restriction estimators, based on the biased initial estimators:

$$\hat{\boldsymbol{\theta}}_{GR1} = (\mathbb{I} - \mathbf{KR})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}),$$

$$\hat{\boldsymbol{\theta}}_{GR2} = (\mathbb{I} - \mathbf{K}^*\mathbf{R})\hat{\boldsymbol{\theta}},$$

$$\hat{\boldsymbol{\theta}}_{GR3} = (\mathbb{I} - \mathbf{K}^*\mathbf{R})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}),$$

where $\mathbf{K} = \mathbf{VR}'(\mathbf{RVR}')^{-1}$, $\mathbf{K}^* = \mathbf{MR}'(\mathbf{RMR}')^{-1}$, $\mathbf{V}$ and $\mathbf{M}$ are accordingly the covariance and the MSE matrices of the initial estimator-vector; $\boldsymbol{b}$ is its bias.

References

[1] Knottnerus, P. (2003). *Sample Survey Theory. Some Pythagorean Perspectives.* Wiley, New York.

[2] Sõstra, K. (2007). Restriction estimation for domains. *Doctoral Dissertation.* Tartu University Press, Tartu.

[3] Sõstra, K., Traat, I. (2009). Optimal domain estimation under summation restriction. *Journal of Statistical Planning and Inference* **139**, 3928-3941.

Combined permutation invariant covariance matrix and estimation in multilevel models

Yuli Liang[1], **Tatjana von Rosen**[1] **and Dietrich von Rosen**[2]

[1] *Stockholm University, Sweden, email: yuli.liang@stat.su.se; tatjana.vonrosen@stat.su.se*
[2] *Swedish University of Agricultural Sciences, Sweden,*
email: Dietrich.von.rosen@et.slu.se

Keywords: maximum likelihood estimator, multilevel models, shift and non-shift invariance, spectral decomposition.

The objective of this paper is to combine shift and non-shift invariance in multilevel models. The random factors are described via their covariance matrices and it is shown that the two types of invariance imply two specific structures for the covariance matrices: block circular Toeplitz and block compound symmetry. Useful results are obtained for the spectrum of such permutation invariant covariance matrices and model reparameterization is performed by putting restrictions on the spectrum. Spectral decomposition is exploited to derive explicit maximum likelihood estimators of the variance and covariance components.

Non-linearity in the returns of the Colombian coffee prices

Miguel Lira Vidrio

University of Guadalajara, México, email: mlira007@gmail.com

Keywords: BDS, coffee price, GARCH, non-linearity.

The commodities suffer from shocks that affect supply and demand, which destabilizes the market, and makes it difficult to predict the behavior of the series. Therefore we have tried to model the series using generalized autoregressive conditional heteroscedasticity (GARCH) [1] to study this behavior.

In this paper the statistical behavior of daily returns of coffee prices ranging from March 1, 1990 until October 28, 2010 is studied. The assumption of linearity, through an unusual test, the test of Brock, Dechert and Scheinkman, (BDS) [2] is reviewed.

The result reveals that the series does not meet the independence criteria and identical distribution. The errors do not follow a normal distribution so it would be difficult to forecast, and would also reject the theory of efficient markets, for which it would be necessary to compare this model with others to compare the results.

References

[1] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

[2] Brock, W., Dechert, W., Scheinkman, J., LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometrics Review* **15**, 197–235.

A comparison of least squares model averaging estimators

Erkki P. Liski[1] **and Antti Liski**[2]

[1] *University of Tampere, Finland, email: erkki.liski@uta.fi*
[2] *Tampere University of Technology, Finland, email: antti.liski@tut.fi*

Keywords: mean square error, model selection, restricted least squares, risk profile, shrinkage estimator.

Our framework is the linear regression model

$$y = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 \mathbf{I}_n), \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ matrix of explanatory variables that we want to keep in the model on theoretical or other grounds. An $n \times m$ matrix $\mathbf{Z}$ contains m additional explanatory variables which we add in the model only if they are supposed to improve estimation of β. To minimize the mean square error of estimation, a balance must be attained between the bias due to omitted variables and the variance due to parameter estimation. Magnus et al. (2010) presented a technique of averaging least squares estimators over models such that the resulting estimator can be presented as a shrinkage estimator. Among this type of estimators we wish to find those that have good risk profile, i.e. the risk is close to the efficiency bound. In general, shrinkage estimators have better risk profile over Post Model Selection (PMS) estimators, and they avoid an unbounded risk.

Shrinkage estimators are computationally superior over the PMS estimators and the model averaging estimators which require estimation of models weights. Computing time of shrinkage estimators increases only linearly with m, the number of auxiliary regressors, while computing time of the PMS estimators is of order 2^m. Thus the shrinkage technique can be easily applied to large data sets when the number of auxiliary regressors is large. We apply the technique on hip fracture patients data to compare treatment costs between hospital districts in Finland.

References

[1] Magnus, J. R., Powell, O., Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* **154**, 139–153.

Statistical analysis of cosmic microwave background data

Anatoliy Malyarenko

Mälardalen University, Sweden, email: anatoliy.malyarenko@mdh.se

Keywords: cosmic microwave background, isotropic Gaussian random field, tensor bundle.

The Cosmic Microwave Background (CMB) consists of photons that began to travel freely when the Universe was approximately 379000 years old. The CMB is completely characterised by its intensity tensor. A CMB detector measures the CMB's intensity tensor P that depends on the direction of observations, $\mathbf{n}$. Mathematically, $\mathbf{n}$ is a point on the sphere S^2, while $P(\mathbf{n})$ is a section of the special tensor bundle over S^2, namely, the tensor product $TS^2 \otimes T^*S^2$ of the tangent bundle TS^2 by the cotangent bundle T^*S^2.

In cosmological models, it is usually assumed that the CMB is a single realisation of a random field. In other words, $P(\mathbf{n})$ is a random section of the bundle $TS^2 \otimes T^*S^2$. A variant of the rigourous mathematical theory of random sections of vector and tensor bundles was built by Malyarenko [1].

After performing primary statistical analysis of raw CMB data we obtain a pixel map $P(\mathbf{n}_j)$. The next step is to perform statistical tests in order to accept or reject the following standard cosmological hypothesis:

$P(\mathbf{n})$ is an isotropic Gaussian random field.

Until now, almost all performed tests used only a part of information that contains in the intensity tensor, namely, its trace, the total intensity $I(\mathbf{n}_j)$.

In our presentation, we will discuss how to use the complete intensity tensor map in the statistical tests.

References

[1] Malyarenko, A. (2011). Invariant random fields in vector bundles and application to cosmology. *Annales de l'Institut Henri Poincaré* (in press).

Methods and algorithms for estimation of statistical characteristics of the MPLS network

Nikolay Medvedev and Philipp Panfilov

Bauman Moscow State Technical University, Russia,
email: medved@bmstu.ru, ponf@bmstu.ru

Keywords: Chvatal-Sankoff conjecture, longest common subsequence, optimal alignment, variance bound.

Proposed algorithms of MPLS [1] networks performance estimation allows to calculate with given accuracy following network characteristics:

- the number of routers in the current network work cycle;
- network topology (the number of senders, routers, connectivity, etc.);
- load (in) for each channel;
- performance for each of the routers;
- size of the queues at routers in a given network;
- total number of packets passing through the router in a session;
- the total number of lost packets on each router for a session;
- percentage of packet loss for each router in a session;
- average packet delay at each router for a session;
- total number of packets of each stream in the past for each of the switched paths for a session;
- percentage of packet loss for each thread in the past for each of the switched paths for a session.

To find the shortest path from sender to receiver, an algorithm is proposed based on use of elements of Dijkstra algorithm for finding the minimum distance in a weighted graph. As a result of the algorithm, we obtain the path length (length, the sum of all weights on this way). The path length is equal to 0 if the initial vertex is finite and equal to -1 if the path does not exist. For service organization streams of packets from certain users in the MPLS network can be allocated separate resources [2]. MPLS network operator, to analyze the capabilities of its network, to serve the above-mentioned flow parameters to provide guaranteed service. An example of application of the developed methods and algorithms to calculate the network routing and traffic optimization is given. As an example, we calculate now the best way for dynamic routing for a different set of restrictions on routing. In the case where there are no requirements flow network bandwidth, length is unlimited, that is the only requirement for routing is the absence of cycles, then we use the function

$$a^{sd,l} = \begin{cases} 1, & \text{when } (s,d) \text{ routers are connected by the edge } l, \\ 0, & \text{otherwise.} \end{cases}$$

To determine the parameters of quality of service it is necessary to classify application traffic by the following characteristics: the relative predictability of the data transmission speed, the sensitivity of traffic to the packet delay, traffic sensitivity to losses and distortions of the package [3]. Three criteria for classification applications correspond to three groups of parameters used in defining and specifying the required quality of service: the parameters of bandwidth, delay settings and parameters of transmission reliability.

References

[1] Lahoud, S., Texier, G., Toutain, L. (2001). *Classification and Evaluation of Constraint-Based Routing Algorithms for MPLS Traffic Engineering.*

[2] Ma, Q., Steenkiste, P. (1997). On path selection for traffic with bandwidth guarantees. In: *Proceedings of IEEE International Conference on Network Protocols.*

[3] Elwalid, A., Jin, C., Low, S. H., Widjaja, I. (2001). MPLS adaptive traffic engineering. In: *Proceedings IEEE INFOCOM 2001.*

Structured families of models with Commutative Orthogonal Block Structure

João T. Mexia

Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa
CMA - Centro de Matemática e Aplicações

Keywords: COBS, structured families.

When models with the same structure correspond to the treatments of a base design we have a structured family of models. The joint analysis of such models will enable the study of the action of the factors in the base design on the models on the family.

When the models in the family are mixed with the same variance components the family will be isomorphic. Then the study of the actions of the factors in the base design will be centered on the estimable vectors of the models in the family.

We will consider such a study for isomorphic families of models with Commutative Orthogonal Block Structure (COBS). The family of variance-covariance matrices for such models will be

$$\boldsymbol{V} = \left\{ \sum_{j=1}^{m} \gamma_j \boldsymbol{Q}_j \right\}$$

where the $\boldsymbol{Q}_1, \ldots \boldsymbol{Q}_m$ are pairwise orthogonal orthogonal projection matrices such that $\sum_{j=1}^{m} \boldsymbol{Q}_j = \boldsymbol{I}_n$, so the model will have Orthogonal Block Structure. Moreover we will assume that the orthogonal projection matrix on the space spanned by the mean vectors commute with $\boldsymbol{Q}_1, \ldots \boldsymbol{Q}_m$.

References

[1] Fonseca, M., Mexia, J. T., Zmyślony, R. (2008). Inference in normal models with commutative orthogonal block structure. *Acta et Commentationes Universitatis Tartunesis de Mathematica* **12**, 3–16.

[2] Nelder, J.A. (1965a). The Analysis of Randomized Experiments with Orthogonal Block Structure. I - Block Structure and the Null Analysis of Variance. In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **283**(1393), 147–162.

[3] Nelder, J.A. (1965b). The Analysis of Randomized Experiments with Orthogonal Block Structure. II - Treatment, Structure and the General Analysis of Variance. In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **283**(1393), 163–178.

[4] Zmyślony, R. (1978). A characterization of Best Linear Unbiased Estimators in the general linear model. *Lecture Notes in Statistics* **2**, 365–373.

Depth-based classification for functional data

Stanislav Nagy

Charles University, Prague, Czech Republic, email: s.nagy@volny.cz

Keywords: band depth, classification, data depth, functional data.

The nonparametric classification of data from a subspace of continuous functions $C([0,1])$ will be discussed. Special attention will be paid to depth-based classification rule and its possible generalizations. The decision rule is related to the concept of data depth, which is in this case a functional

$$D : C([0,1]) \to [0,1].$$

Depth is a measure of centrality of an observation with respect to a data set or a distribution. Recently several authors proposed their notions of depth for functional data (Fraiman and Muniz [2], López-Pintado and Romo [5]). These depth functionals are invariant with respect to a domain permutation

$$T : C([0,1]) \to C([0,1]) : x(t) \mapsto x(\phi(t)),$$

where ϕ is a bijection of $[0,1]$ and $t \in [0,1]$. Thus, none of the established depth functionals is able do deal with the shape of functions.

This problem will be demonstrated in a functional classification task. A new class of depth functionals, K-band depths for $K \in \mathbb{N}$ will be utilized in order to handle it. The simplicial depth described by Liu [4] along with Fraiman-Muniz method are employed to involve derivatives into depth computation. The performance of the new approach is compared to similar results obtained by Cuveas et al. [1] in a simulation study of functional data supervised classification. We show that proper derivative using in combination with DD-plot (Depth-Depth plot) techniques proposed by Li et al. [3] is a powerful tool not only for the classification of functional observations.

References

[1] Cuevas, A., Febrero, M., Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* **22**(3), 481–496.

[2] Fraiman, R., Muniz, G. (2001). Trimmed means for functional data. *Test* **10**(2), 419–440.

[3] Li, J., Cuesta-Albertos, J. A., Liu, R. Y. (2010). DD-Classifier: Nonparametric Classification Procedure Based on DD-plot, preprint.

[4] Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics* **18**(1), 405–414.

[5] López-Pintado, S., Romo, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104**(486), 718–734.

Estimation of parameters in the extended growth curve model with a linearly structured covariance matrix

Joseph Nzabanita[1], Martin Ohlson[1] and Dietrich von Rosen[2]

[1] *Linköping University, Sweden, email: joseph.nzabanita@liu.se, martin.ohlson@liu.se*
[2] *Swedish University of Agricultural Sciences, Sweden, email: Dietrich.von.Rosen@et.slu.se*

Keywords: estimation, extended growth curve model, linearly structured covariance matrix, residuals.

The extended growth curve model with two terms and a linearly structured covariance matrix is considered. In general there is no problem to estimate the covariance matrix when it is completely unknown. However, problems arise when one has to take into account that there exists a structure generated by a few number of parameters. In this paper an estimation procedure that handles linearly structured covariance matrices is proposed. The idea is first to estimate the covariance matrix when it should be used to define an inner product in a regression space and thereafter re-estimate it when it should be interpreted as a dispersion matrix. This idea is exploited by decomposing the residual space, the orthogonal complement to the design space, into three orthogonal subspaces. Studying residuals obtained from projections of observations on these subspaces yields explicit consistent estimators of the covariance matrix. An explicit consistent estimator of the mean is also proposed and numerical examples are given.

Cornish-Fisher expansions using sample cumulants and monotonic transformations

Haruhiko Ogasawara

Otaru University of Commerce, Japan, email: hogasa@res.otaru-uc.ac.jp

Keywords: asymptotic cumulants, asymptotic expansions, Cornish-Fisher expansion, monotonic transformation, studentization.

General formulas of the asymptotic cumulants of a studentized parameter estimator are given up to the fourth order with the added higher-order asymptotic variance. Using the sample counterparts of the asymptotic cumulants, formulas for the Cornish-Fisher expansions with the third-order accuracy are obtained. Some new methods of monotonic transformations of the studentized estimator are presented. In addition, similar transformations of a fixed normal deviate are proposed up to the same order with some asymptotic comparisons to the transformations of the studentized estimator. Applications to the mean and the binomial proportion are shown with a numerical illustration for estimation of the proportion.

Profile analysis for a growth curve model

Martin Ohlson[1] and Muni S. Srivastava[2]

[1] *Department of Mathematics, Linköping University, Sweden,*
email: martin.ohlson@liu.se
[2] *Department of Statistics, University of Toronto, Canada,*
email: srivastava@utstat.toronto.edu

Keywords: growth curve model, profile analysis.

In this talk, we consider profile analysis of several groups where subvectors of the mean vectors are equal. This leads to a profile analysis in a growth curve model. The likelihood ratio statistics are given for the three hypotheses known in literature as parallelism, level hypothesis and flatness. Furthermore, exact and asymptotic distributions are given in the relevant cases.

References

[1] Ohlson, M., Srivastava, M. S. (2010). Profile analysis for a Growth Curve Model. *Journal of the Japan Statistical Society* **40**(1), 1–21.

[2] Srivastava, M. S. (1987). Profile analysis of several groups. *Communications in Statistics - Theory and Methods* **16**(3), 909–926.

Polynomial smoothing of discrete sparsely observed distribution

Paulo Eduardo Oliveira

CMUC, University of Coimbra, Portugal, email: paulo@mat.uc.pt

Keywords: discrete data, local polynomial, sparse observations.

Let $\mathbf{P}$ be a probability distribution on a support with N cells arranged, for simplicity, in a table $\mathbf{C} = (C_{i,j})$, where $i = 1\ldots, K$, $j = 1,\ldots L$. Observation counts are described by $\mathbf{N} = (N_{i,j})$, or equivalently, by the empirical probability distribution $\overline{\mathbf{P}} = (\overline{P}_{i,j} = N_{i,j}/n)$, where $n = \sum_{i,j} N_{i,j}$. Rearranging the rows in order to have a N-dimensional vector, $\mathbf{N}$ is multinomially distributed.

This talk is concerned with the estimation of $\mathbf{P} = (P_{i,j})$ with a special interest when the sample size n is small. Moreover, having in mind a few applications, some partial knowledge of the distribution might be available and we should integrate this into the estimation. We will assume the knowledge of the marginal distribution of $\mathbf{P}$, that is, for some given Π_i, $i = 1,\ldots,K$, $\Pi_i = \sum_{j=1}^{L} P_{i,j}$. The general idea in constructing estimators is to adapt polynomial smoothing to this framework.

To avoid computational difficulties with border and edge effects, we consider a replication of the tables $\mathbf{C}$, $\mathbf{P}$ and $\mathbf{N}$, enlarging them by reflecting cells with respect to each one of the four borders and edges. Defining the appropriate matrices in the usual way, local polynomials can be represented as the minimizers of $H_{i,j} = \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)^t \mathbf{W}_{i,j} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,j}\beta_{i,j}\right)$, for each i and j. This procedure does not integrate knowledge of the marginal distribution. Moreover, it may produce non-acceptable estimates, especially when n is small, as is our case of interest. We propose to correct this in two different ways. The first one is to introduce in the minimization problem a constraint, forcing the solution to agree with the marginal distribution:

$$\begin{aligned} &\text{minimize} \sum_{\ell=1}^{L} H_{i,\ell} \\ &\text{subject to } \textstyle\sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1,\ldots,K. \end{aligned}$$

The second one changes the minimizing function by considering relative errors:

$$\begin{aligned} &\text{minimize } H_i^* = \sum_{\ell=1}^{L} \frac{1}{\beta_{0,i,\ell}} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right)^t \mathbf{W}_{i,\ell} \left(\overrightarrow{\mathbf{P}} - \mathbf{X}_{i,\ell}\beta_{i,\ell}\right), \\ &\text{subject to } \sum_{j=1}^{L} \beta_{0,i,j} = \Pi_i, \quad i = 1,\ldots,K. \end{aligned}$$

We characterize the obtained estimators, describe their behaviour and relations, and present some numerical results showing their performance. Finally, we note that this approach is easily extended to higher dimensional supports.

Life distributions in survival analysis and reliability: Structure of semiparametric families

Ingram Olkin

Stanford University

Keywords: life distributions, semiparametric distributions, survival analysis.

Semiparametric families are families that have both a real parameter and a parameter that is itself a distribution. A number of semiparametric parametric families suitable for lifetime data in survival or reliability are introduced: scale, power, frailty (proportional hazards), age, moment, and others. Interesting results on stochastic orderings are obtained for these families. The coincidence of two families provides a characterization of the underlying distribution. Some of the characterization results provide a rationale for the use of certain families. In this talk we provide an overview of these semiparametric families, and present several characterizations.

This work is a joint effort with Albert W. Marshall.

References

[1] Marshall, A.W., Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families.* Springer, New York.

Majorization and its applications

Ingram Olkin

Stanford University

Keywords: majorization, partial order.

Majorization is a partial order that surprisingly arises naturally in economics, chemistry, physics, political science, and more. Of particular interest are applications in probability and statistics. In this talk, I will provide some historical origins as they arise in applications, present some of the key theorems, and list examples in different fields.

References

[1] Marshall, A.W., Olkin, I., Arnold, B.C. (2011). *Inequalities: Theory of Majorization and Its Applications. Second edition.* Springer, New York.

Asymptotics for the normalized spectral function

Jolanta Pielaszkiewicz, Martin Ohlson, Dietrich von Rosen

Linköping University, Sweden, email: jolanta.pielaszkiewicz@liu.se, maohl@mai.liu.se, Dietrich.von.rosen@et.slu.se

Keywords: asymptotics, p/n, spectral function, Stieltjes transform.

The Stieltjes transform will be used to obtain asymptotics for the normalized spectral function of a quadratic form $AA' + \frac{1}{n}XX'$, where A is non-random matrix and $X \sim N_{p,n}(0, \Sigma, \Psi)$, where p and n are respectively - the number of variables and observations, such that $\frac{p}{n} \to c > 0$.
The density will be derived explicitly for the case $A = 0$ and $X \sim N_{p,n}(0, \sigma^2 I, I)$.

References

[1] Girko, V.L., von Rosen, D. (1994). Asymptotics for the normalized spectral function of matrix quadratic form. *Random Oper. and Stoch. Equ.* **2**(2), 153–161.

[2] Girko, V.L. (1990). *Theory of Random Determinants.* Kluwer, Dordrecht.

Using Edgeworth expansion approximating two- and three-dimensional probability distribution functions

Margus Pihlak

Tallinn University of Technolgy, Estonia, email: margusp@staff.ttu.ee

Keywords: Edgeworth expansion, Hermite matrix polynomials.

In this talk we present the techniques for approximating unknown distribution function with a well-known and well-studied distribution function. The development of approximation technique is closely related with development of matrix algebra. We also present some newer results of matrix algebra. For more detailed presentation of this kind matrix algebra see [1] , [3], [2], for example. Some results on Edgeworth expansions are presented in [4] where a two-dimensional distribution function is approximated by means of the Edgeworth type expansion. In this presentation we generalize the Edgeworth expansion to the three-dimensional case. This presentation is supported by Estonian Science Foundation Grant 7656.

References

[1] Harville, A. (1997). *Matrix Algebra from a Statistican's Perspective*. Springer, New York.

[2] Kollo, T., von Rosen D. (2005). *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht.

[3] Pihlak, M. (2004). Matrix integral. *Linear Algebra and Its Applications* **388**, 315–325.

[4] Pihlak, M. (2008). Approximation of multivariate distribution functions. *Mathematica Slovaca* **58**, 635–652.

Twenty times IWMS: A short history

Simo Puntanen[1] **and George P. H. Styan**[2]

[1] *University of Tampere, Finland, email: simo.puntanen@uta.fi*
[2] *McGill University, Montréal, Canada, email: styan@math.mcgill.ca*

Keywords: IWMS-series.

We briefly describe the development of the International Workshop on Matrices and Statistics (IWMS) series. The first IWMS was held in Tampere, Finland, 6–8 August 1990, and the 20th IWMS is being held in Tartu, Estonia, 26 June – 1 July 2011. An illustrated summary of the talk is presented in the end of this Abstract volume.

Distribution of an arbitrary linear transformation of internally Studentized residuals of multivariate regression with elliptical errors

Seppo Pynnonen

University of Vaasa, Finland

Keywords: general linear hypothesis, left-spherical elliptical distribution, outliers.

This paper derives the matrix-variate joint distribution of an arbitrary non-singular linear transformation of Studentized residuals from multivariate regression with elliptically distributed errors. The joint distributions of the major commonly utilized Studentized versions of (multivariate) regression residuals are obtained as special cases of the matrix-variate distribution introduced in the paper. Applications in regression diagnostics and testing general linear hypothesis are briefly discussed.

Multivariate models with space varying memory

Alfredas Račkauskas[1] and Charles Suquet[2]

[1] *Vilnius University, Lithuania, email: alfredas.rackauskas@mif.vu.lt*
[2] *University Lille I, France, email: charles.suquet@univ-lille1.fr*

Keywords: fractional Brownian motion, functional central limit theorem, Hilbert space, linear processes, long memory.

In the talk we shall discuss long memory phenomenon of multidimensional time series. We consider an operator fractional Brownian motion with values in a finite or infinite dimensional Hilbert space defined via operator-valued Hurst exponent. We prove that this process is a limiting one for polygonal lines constructed from partial sums of time series having space varying long memory. The talk is based on the paper [1].

References

[1] Račkauskas, A., Ch. Suquet (2011). Operator fractional Brownian motion as limit of polygonal line processes in Hilbert space. *Stochastics and Dynamics* **11**(1), 1–22.

Multivariate linear models and profile analysis

Dietrich von Rosen

Swedish University of Agricultural Sciences, email: dietrich.von.rosen@slu.se

Keywords: extended growth curve model, Growth Curve model, profile analysis.

We give an overview presentation of multivariate linear normal models theory and applications. In particular profile analysis, partial least squares and spatio-temporal models are considered. Moreover, the Growth Curve model, sum of profiles model and the extended growth curve model will be discussed.

References

[1] Kollo, T., von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices.* Springer, Dordrecht.

On the identification of sets of informative peptides on antigen microarrays

Tatjana von Rosen[1], Marju Valge[2], Triin Võrno[2] and Ene Käärik[2]

[1] *University of Tartu, Stockholm University, email: Tatjana.vonRosen@stat.su.se,*
[2] *University of Tartu, Estonia, email: Ene.Kaarik@ut.ee*

Keywords: high-dimensional data, profile analysis, selection of variables.

Within the last decade, the development of antigen chip technology has enabled the simultaneous measurement of thousands of peptides in biological samples. Finding sets of peptides which can uniquely characterize TB patients and healthy individuals, can, for example, help to develop better diagnostic tests and to identify candidate vaccine antigens. In this work we use the procedure for variables selection which is applicable in a high-dimensional setting worked out by Läuter et al (2009) and new identification technique of control peptides that gives high mean response and low coefficient of variation across all replicates (Ngo et al, 2009).

Using profile analysis (of sets of peptides), as the next step, will facilitate the study of behavior of the immune system of vaccinated individuals over time.

References

[1] Läuter, J., Horn, F., Rosołowski, M., Glimm, E. (2009). High-dimensional data analysis: Selection of variables, data compression and graphics - Application to gene expression. *Biometrical Journal* **51**(2), 235-251.

[2] Ngo, Y., Advani, R., Valentini, D., Gaseitsiwe, S., Mahdavifar, S., Maeurer, M., Reilly, M. (1994). Identification and testing of control peptides for antigen microarrays. *Journal of Immunological Methods* **343**(2), 68–78.

On the reliability of Errors-in-Variables Models

Burkhard Schaffrin and Sibel Uzun

The Ohio State University, Columbus/OH, USA,
email: schaffrin.1@osu.edu, uzun.1@buckeyemail.osu.edu

Keywords: correlated observations, Errors-in-Variables (EIV) Model, outlier detection, redundancy numbers, reliability, Total Least-Squares (TLS)

Reliability has been quantified in a simple Gauss-Markov Model (GMM) by Baarda [1] for the application to geodetic networks as the potential to detect outliers – with a specified significance and power – by testing the Least-Squares residuals for their zero expectation property after an adjustment assuming "no outliers". It was shown that, under homo-scedastic conditions, the so-called "redundancy numbers" could very well serve as indicators for the "local reliability" of an (individual) observation. In contrast, the maximum effect of any undetectible outlier on the estimated parameters would indicate "global reliability."

This concept has been extended successfully to the case of correlated observations by Schaffrin [3] quite a while ago. However, no attempt has been made so far to extend Baarda's results to the (homoscedastic) Errors-in-Variables (EIV) Model for which Golub and van Loan [2] had found their – now famous – algorithm to generate the Total Least-Squares (TLS) solution, together with all the residuals. More recently, this algorithm has been generalized by Schaffrin and Wieser [4] to the case where a truly – not just element-wise – Weighted TLS solution can be computed when the covariance matrix has the structure of a Kronecker-Zehfuss product.

Here, an attempt will be made to define reliability measures within such an EIV Model, in analogy to Baarda's original approach.

References

[1] Baarda, W. (1968). A testing procedure for use in geodetic networks. *Neth.Geodetic Comm., Public. on Geodesy*, New Series **2**(5), Delft/NL.

[2] Golub, G.H., van Loan, C. F. (1980). An analysis of the Total Least-Squares problem. *SIAM J. Numer. Anal.* **17**(6), 883–893.

[3] Schaffrin, B. (1997). Reliability measures for correlated observations. *J. of Surveying Engrg* **123**(3), 126–137.

[4] Schaffrin, B., Wieser, A. (2008). On weighted Total Least-Squares adjustment for linear regression. *J. of Geodesy* **82**, 415–421.

Privacy protection and quantile estimation under noise multiplication

Bimal Sinha[1], Tapan Nayak[2], Laura Zayatz[3]

[1] *University of Maryland Baltimore County and US Census Bureau*
email: sinha@umbc.edu
[2] *George Washington University and US Census Bureau, email: tapan@gwu.edu*
[3] *US Census Bureau, email: laura.zayatz@census.gov*

Keywords: confidentiality, disclosure risk, Metropolis algorithm, noise multiplication, prior distribution.

We will address two inferential aspects of noise multiplied magnitude microdata. First, in the context of disclosure risk assessment of tabular magnitude data, we study the consequences of noise multiplication of original microdata when an intruder tries to speculate a target unit's value in a cell based on knowledge of the noise perturbed cell total and values of some original units within the cell. Second, we discuss statistical methods to infer about a *quantile* of a microdata set based on their noise perturbed values. An application with income data will be presented.

References

[1] Blumenthal, S., Cohen, A. (1968). Estimation of the larger translation parameter. *Ann. Math. Statist.* **39**, 502–516.

[2] Blumenthal, S., Cohen, A. (1968). Estimation of the larger of two normal means. *Ann. Math. Statist.* **39**, 861–876.

[3] Brand, R. (2002). Microdata protection through noise addition. In: *Inference Control in Statistical Databases*, Ed. J. Domingo-Ferrer. Berlin: Springer, 97–116.

[4] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Ed.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier.

[5] Dudewicz, E.J. (1971). Maximum likelihood estimators for non-1-1 functions. *Trabojos de Estadistica y de Investigacion Operativa* **22**, 65–70.

[6] Dudewicz, E.J. (1971). Maximum likelihood estimators for ranked means. *Z. Wahrscheinlichkeitstheorie Verw. Grb.* **19**, 29-42.

[7] Duncan, G.T., Fienberg, S.E. (1999). Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In: *Eurostat Statistical Data Protection '98 Lisbon*, Luxembourg: Eurostat, 351–362.

[8] Duncan, G.T., Stokes, S.L. (2004). Disclosure risk vs. data utility: The R-U confidentiality map as applied to topcoding. *Chance* **17**, 16–20.

[9] Duncan, G. T., Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *J. Amer. Statist. Assoc.* **95**, 720–729.

[10] Elfessi, A., Pal, N. (1992). Estimation of the smaller and larger of two uniform scale parameters. *Commun. Statist. - Theory Meth.* **21**, 2997–3015.

[11] Evans, T., Zayatz, L., Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *J. Official Statist.* **4**, 537-551.

[12] Fong, D.K.H. (1987). *Ranking and Estimation of Exchangeable Means in Balanced and Unbalanced Models: A Bayesian Approach.* Ph.D. thesis, Purdue University.

[13] Fong, D.K.H. (1992). A Bayesian approach to the estimation of the largest normal mean. *J. Statist. Comput. Simul.* **40**, 119–133.

[14] Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *J. Official Statist.*, 383–406.

[15] Givens, G.H., Hoeting, J.A. (2005). *Computational Statistics.* New York: John Wiley.

[16] Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *Amer. Statist.* **60**, 224–232.

[17] Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In: *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 303–308.

[18] Kim, J.J., Winkler, W.E. (2003). Multiplicative noise for masking continuous data. Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., April 2003.

[19] Kumar, S., Sharma, D. (1993). Unbiased inestimability of ordered parameters. *Statistics* **24**, 137–142.

[20] Little, R.J.A. (1993). Statistical analysis of masked data. *J. Official Statist.* **9**, 407–426.

[21] Misra, N., Anand, R., Singh, H. (1997). Estimation of the smaller and larger scale parameters of two exponential distributions. *Statistics and Decisions* **15**, 75–98.

[22] Nayak, T. Sinha, B., Zayatz, L. (2010). Statistical properties of multiplicative noise masking for confidentiality protection. To appear in *J. Official Statist.*

[23] Shao, J. (1999). *Mathematical Statistics.* New York: Springer.

[24] Tendick, P. (1991). Optimal noise addition for preserving confidentiality in multivariate data. *J. Statist. Plann. Inference* **27**, 341–353.

[25] van Eeden, C. (2006). *Restricted Parameter Space Estimation Problems.* New York: Springer.

[26] Willenborg, L.C.R.J., De Waal, T. (2001). *Elements of Statistical Disclosure Control.* New York: Springer.

[27] Yancey, W.E., Winkler, W.E., Creecy, R.H. (2002). Disclosure risk assessment in perturbative microdata protection. In: *Inference Control in Statistical Databases*, Ed. J. Domingo-Ferrer, Springer, 135–152.

A two sample test in high dimension with fewer observations

Muni S. Srivastava

University of Toronto

Keywords: asymptotic normality, equality of two mean vectors, fewer observations, high dimension, unequal covariance matrices.

In this paper, we propose a test for testing the equality of mean vectors of two groups with unequal covariance matrices based on N_1 and N_2 independently distributed p-dimensional observation vectors. This test is invariant under the transformation of the observation vectors by any $p \times p$ diagonal matrix. There are no tests available in the literature that has this invariance property.

The asymptotic distribution of the test statistics is given as $(N_1, N_2, p) \to \infty$, where $(N_1/N_2) \to k \in (0, \infty)$ but (N_1/p) and (N_2/p) may go to zero or infinity.

Caïssan squares: the magic of chess

George P. H. Styan

McGill University, Canada, email: styan@math.mcgill.ca

Keywords: alternate couplets property, bibliography, Caïssa, EP, 4-ply, involution-associated magic matrices, involutory matrix, knight-Nasik, magic key, most-perfect, pandiagonal, philatelic items, postage stamps, rhomboid, "Ursus".

We study various properties of $n \times n$ Caïssan magic squares. A magic square is Caïssan whenever it is pandiagonal and knight-Nasik, so that all paths of length n by a chess bishop are magic (pandiagonal) and by a (regular) chess knight are magic (CSP2-magic).

Following the seminal 1881 article [4] by "Ursus" in *The Queen,* we show that 4-ply magic matrices, or equivalently magic matrices with the "alternate-couplets" property, have rank at most equal to 3. We also show that an $n \times n$ magic matrix $\mathbf{M}$ with rank 3 and index 1 is EP if and only if $\mathbf{M}^2$ is symmetric. We identify and study 46080 Caïssan beauties—Caïssan magic squares which are also CSP3-magic; a CSP3-path is made by a special knight that leaps over 3 instead of 2 squares. We find that just 192 of these Caïssan beauties are EP. We generalize an algorithm given by Cavendish [2:(1894)] for generating Caïssan beauties and find these are all EP. We also study the n-queens problem first posed with $n = 8$ by Bezzel [1:(1848)] and the Firth–Zukertort "magic chess board" due to Firth [3:(1887)].

An extensive annotated and illustrated bibliography of over 300 items, many with hyperlinks, ends our report. We give special attention to items by (or connected with) "Ursus": Henry James Kesson (b. c. 1844), Andrew Hollingworth Frost (1819–1907), Charles Planck (1856–1935), and Pavle Bidev (1912–1988). We have tried to illustrate our findings as much as possible, and whenever feasible with images of postage stamps or other philatelic items.

References

[1] [Max Friedrich Wilhelm Bezzel (1824–1871)] (1848). Vor einige Zeit wurden uns von einem Schachfreunde zwei Fragen vorgelegt .... *Schachzeitung der Berliner Schachgesellschaft* **3**, p. 363.

[2] "Cavendish" [Henry Jones (1831–1899)] (1894). *Recreations with Magic Squares: the eight queens' problem solved by magic squares and domino squares,* Thomas de la Rue, London.

[3] Firth, W. A. [William A. Firth (c. 1815–1890)] (1887). *The Magic Square,* printed by R. Carswell & Son, Belfast.

[4] "Ursus" [Henry James Kesson (b. c. 1844)] (1881). Caïssan magic squares. *The Queen: The Lady's Newspaper & Court Chronicle,* **70**, p. 142 (August 6, 1881), pp. 276–277 (September 10, 1881) & p. 391 (October 15, 1881).

NIG Lévy process in asset price modelling: case of Estonian companies

Dean Teneng

Institute of Mathematical Statistics, University of Tartu, Estonia, email: teneng@ut.ee

Keywords: asset price, Lévy process, NIG.

As an asset is traded at fair value, its varying price trace out an interesting time series reflecting in a general way the asset's value and underlying economic activites [1]. This time series exhibit price jumps, clustering and a host of other properties [3] not usually captured by log-normal models [2].

Interestingly, Lévy processes offer the possibility of distinguishing jumps, diffusion, drift [3] and the laxity to answer questions on frequency, continuity, etc. An important feature of Normal Inverse Gaussian-Lévy (NIG-Lévy) model is its path richness i.e. it can model so many small jumps such that it does not need a Brownian motion component to capture these. Hence limitations arising from Brownian motion based models are almost eliminated. Secondly, the unique characteristics listed above are reflected in the Lévy triplet. These are easily introduced in the modeling picture by moments method i.e. just matching the theoretical and empirical descriptive statistics and extracting parameters for NIG since NIG has a well-behaved characteristic function.

We use a simple R-code to reproduce the price trajectories of 10 Estonian companies.

References

[1] Gentle, J. E., Hardle, W. K. Modelling asset prices [and Discussion]. *http://sfb649.wiwi.hu-berlin.de*

[2] Teneng, D. (2011) Limitations of the Black-Scholes model. *International Research Journal of Finance and Economics*, in print.

[3] Teneng, D. (2010). Path properties of Lévy processes In: *Proceedings of First International Scientific Conference of Students and Young Scientists Theoretical and Applied Aspects of Cybernetics, Feb 21-25, Kiev-Ukraine*, 207–210.

On exact testing problems in linear models with two variance-covariance components

Julia Volaufova and Lynn R. LaMotte

Louisiana State University Health Sciences Center – New Orleans

Keywords: accuracy of p-value, approximate test, exact test, fixed effects, variance components.

Linear models with variance-covariance components are used in a wide variety of applications. A special case of models with two variance-covariance components has been studied extensively for decades. Most often the objective of inference is testing linear hypotheses about the mean of the response. Even assuming multivariate normality, it is not clear what test to recommend except in a few special settings, such as balanced or orthogonal designs. Here we shall investigate a simultaneous hypothesis on the mean and on the between-subject variance component (see also Crainiceanu & Ruppert (2004)) and in that setting special cases of hypotheses will be studied. Some special cases will be mentioned as well. We shall illustrate some statistical properties of test procedures, such as accuracy of p-values and powers of approximate and exact tests obtained by simulation.

References

[1] Crainiceanu, C.M., Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **66**(1), 165–185.

Stochastic forecast of the population using models for fertility and mortality

Mare Vähi

University of Tartu, Estonia, email: mare.vahi@ut.ee

Keywords: beta-distribution, forecast, predictive distribution, uncertainty.

Probabilistic population forecast is based on stochastic population renewal using forecasts of fertility and mortality. Fitting the suitable family of distributions for modelling the changes in fertility distribution is the first step. Most commonly used distributions are beta distribution, gamma distribution and Hadwiger distribution. The beta distribution and mixtures of beta distributions show excellent fit to the one year age-specific fertility rate distributions. The second step is to estimate the parameters of the distribution of mortality. The fertility model and mortality model are then used in simulation of future fertility and mortality to obtain forecasts of the population. The method is demonstrated using Estonian data for the period 1991 - 2009.

References

[1] Alho, J. M., Spencer, B. D. (2005). *Statistical Demography and Forecasting.* Springer Science, New York.

[2] Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting* **22**(3), 547–581.

A study on the distribution of the residual in the Growth Curve Model

Jianxin Wei[1], **Ying Li**[2] **and Dietrich von Rosen**[2,3]

[1] *Uppsala University, Sweden, email: jianxin.wei@statistik.uu.se*
[2] *Swedish University of Agricultural Sciences, Sweden, email: ying.li@slu.se*
[3] *Linköping University, Sweden, email: dietrich.von.rosen@slu.se*

Keywords: bootstrap, Edgeworth expansion, Growth Curve Model, residuals.

The Growth Curve Model was introduced by Potthoff & Roy (1964). The maximum likelihood estimators of the parameters in the model have been obtained. The study of the properties of the estimators has taken place over many years. However, the distributions of the residuals are still unknown and are interesting to consider from both, a theoretical and a practical point of view. In this paper, the approximation of the distribution of the residuals in the Growth Curve Model is derived via Edgeworth expansions and by aid of bootstrap methodology. A simulation study is included to verify the results, and further to compare the MLE and least square estimator in the growth curve model.

References

[1] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

[2] Kollo, T., Roos, A., von Rosen, D. (2007). Approximation of the distribution of the location parameter in the Growth Curve Model. *Scandinavian Journal of Statistics* **34**(3), 499–510.

[3] Potthoff, R.F., Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for Growth Curve problems. *Biometrika* **51**, 313–326.

[4] von Rosen, D. (1995). Residuals in the Growth Curve Model. *Annals of the Institute of Statistical Mathematics* **47**(1), 129–136.

Some results related to the general Gauß-Markov model
–
weak complementarity, non-testability and missing observations

Hans Joachim Werner

University of Bonn, Germany, email: hjw.de@uni-bonn.de

Keywords: Gauß-Markov model, missing observations, non-testability, weak complementarity.

In the literature, COMPLEMENTARY matrices have been studied because of their importance when analyzing OVER-PARAMETRIZED linear statistical models. In this talk, the somewhat more general concept of WEAK COMPLEMENTARITY is considered. Observing the fact that the usual F-TEST in ANOVA is applicable only for "TESTABLE" hypotheses, that in practice however - e.g. in non-orthogonal settings or incomplete layouts - NON-TESTABLE hypotheses can be of importance, a variant of the F-TEST is discussed that allows to decide for significant deviations also in NON-TESTABLE situations and to detect NON-TESTABILITY, too.

On the performance of the restricted estimators

M. Revan Özkale

Cukurova University, Turkey, email: mrevan@cu.edu.tr

Keywords: mean square error, restricted estimators.

The ordinary least squares (OLS) estimator is often used to estimate the parameters in linear regression models. Multicollinearity among the columns of the explanatory variables is known to cause severe distortion of the OLS estimates of the parameters. Therefore, alternative methods to solve the multicollinearity problem are preferred. One of the methods for solving the multicollinearity problem is through the use of non-sample information on the parameters which results in the restricted estimators. This study provides the results on the performance of the restricted two parameter estimator (see [3]), which includes the restricted ridge (see [1]), restricted Liu and restricted shrunken estimators as special cases, over the restricted least squares and the OLS estimators under the matrix mean square error (MSE) criterion when the restrictions are not correct and when they are correct. Theoretical results are evaluated via a numerical example based on Webster et al. [4] and the behavior of the restricted estimators is examined by the surface plot of the scalar MSE on the data set.

References

[1] Groß, J. (2003). Restricted ridge estimation. *Statistics and Probability Letters* **65**, 57–64.

[2] Harville, D. A. (1997). *Matrix Algebra From A Statistician's Perspective.* Springer-Verlag, New York.

[3] Özkale, M. R., Kaçıranlar, S. (2007). The restricted and unrestricted two parameter estimators. *Communications in Statistics: Theory and Methods* **36**(15), 2707–2725.

[4] Webster, T. J., Gunst, R. F., Mason, R. L. (1974). Latent root regression analysis. *Technometrics* **16**, 513–522.

A comparison of three estimators in GMANOVA model when the sample size is fewer than the dimension

Hirokazu Yanagihara[1] **and Muni S. Srivastava**[2]

[1] *Hiroshima University, Japan, email:yanagi@math.sci.hiroshima-u.ac.jp*
[2] *University of Toronto, Canada, email:srivasta@utstat.toronto.edu*

Keywords: GMANOVA model, high dimensional data, least square estimator, Moore-Penrose inverse, ridge type estimator.

A generalized multivariate analysis of variance (GMANOVA) model proposed by Potthoff and Roy [1] is one of statistical models suitable for a longitudinal data. The matrix form of GMANOVA model is given by

$$Y \sim N_{n\times p}(A\Xi X', \Sigma \otimes I_n),$$

where Y is an $n\times p$ response variables matrix, A is an $n\times k$ between-individuals explanatory variables matrix with the full rank k $(< n)$, X is a $p \times q$ within-individuals explanatory variables matrix with the full rank q $(\leq p)$, and Ξ is a $k\times q$ unknown regression coefficients matrix. It is a known fact that a maximum likelihood estimator (MLE) of regression coefficients Ξ in the GMANOVA model is defined as

$$\hat{\Xi}_{\mathrm{ML}} = (A'A)^{-1}A'YS^{-1}X(X'S^{-1}X)^{-1},$$

where $S = Y'\{I_n - A(A'A)^{-1}A'\}Y/(n-k)$. However, if the dimension p is larger than the sample size n, the MLE of Ξ cannot be defined because the inverse matrix of S does not exist then. Hence, we avoid such an undesirable situation by the following three methods:

(1) We ignore S^{-1}. This is corresponding to the least square estimator (LS) of Ξ, i.e.,

$$\hat{\Xi}_{\mathrm{LS}} = (A'A)^{-1}A'YX(X'X)^{-1}.$$

(2) We replace S^{-1} with the inverse matrix of the ridge type estimator of S, i.e.,

$$\hat{\Xi}_{\mathrm{R}} = (A'A)^{-1}A'YS_\lambda^{-1}X(X'S_\lambda^{-1}X)^{-1},$$

where $S_\lambda = S + \lambda I_p/(n-k)$ and $\lambda = \mathrm{tr}(S)/\sqrt{p}$. This ridge type estimator of S proposed by Srivastava and Kubokawa [2].

(3) We replace S^{-1} with the Moore-Penrose inverse of S, i.e.,

$$\hat{\Xi}_{\mathrm{MP}} = (A'A)^{-1}A'YS^+X(X'S^+X)^{-1}.$$

An aim of this paper is to compare with above three estimators theoretically and numerically when the dimension p is larger than the sample size n.

References

[1] Potthoff, R. F., Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313–325.

[2] Srivastava, M. S., Kubokawa, T. (2008). Akaike information criterion for selecting components of the mean vector in high dimensional data with fewer observations. *Journal of the Japan Statistical Society* **38**, 259–283.

Tartu Conferences on Multivariate Statistics: A short retrospective view

Tõnu Kollo

University of Tartu, Estonia, email: 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 all still active in research 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.

Photo 1: 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, Taivo Arak, Donatas Surgailis, Šarunas Raudis, Boris G. Mirkin. All the Soviet Union wide conferences were attended by more than a hundred participants, and there was always tight competition to have your talk included into 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 "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 "Matrices in Statistics", the general organiser of which was Professor George P. H. Styan.

Photo 2: E.-M.Tiit at the Opening Section (1994).

PHOTO 3: I. Olkin giving a talk (1994).

PHOTO 4: I. Olkin giving a talk (1994).

Photo 5: From left: Y. Fujikoshi, Mrs. C. R. Rao, C. R. Rao (1994).

Photo 6: From left: T. Kollo, K.-T. Fang, D. v. Rosen (1994).

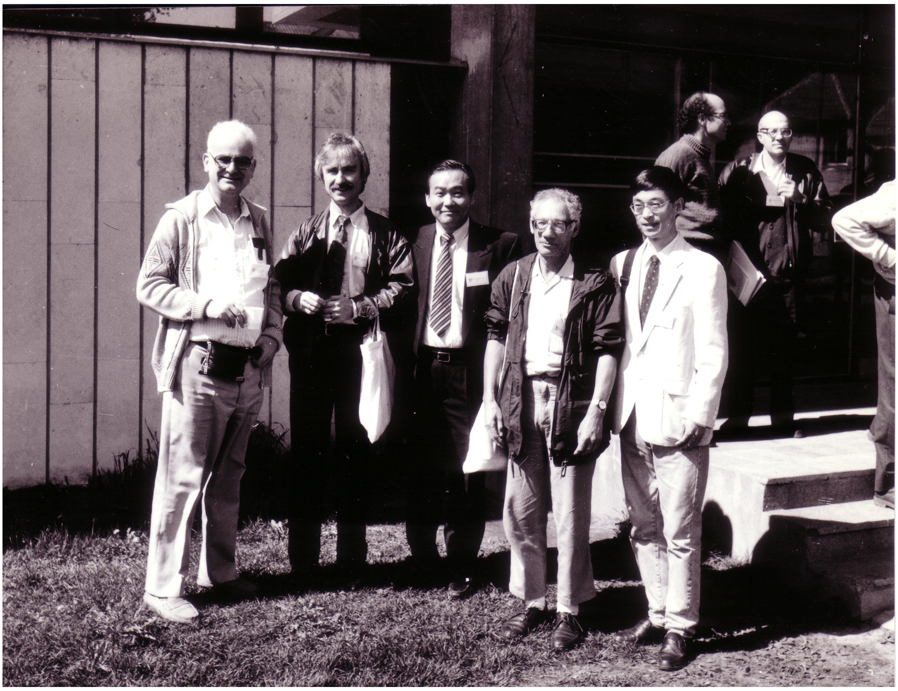

PHOTO 7: From left: G. P. H. Styan, H. J. Werner, E. I. Im, H. Neudecker, S. Liu (1994).

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 T. 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.

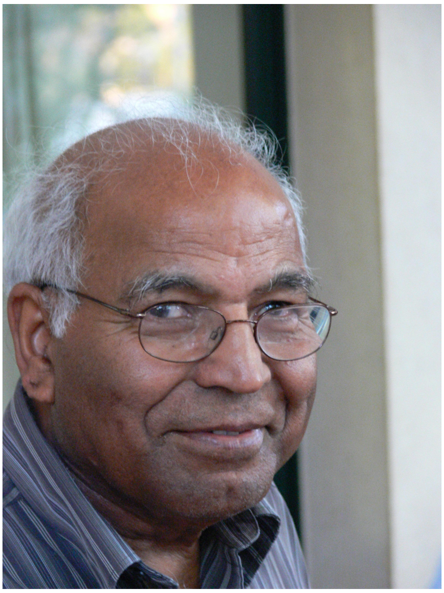

PHOTO 8: M. S. Srivastava (2003).

PHOTO 9: B. Arnold (standing) and N. Balakrishnan (2007).

The previous 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.

Programme Committee wishes all of you fruitful ideas and enjoyable time in Tartu.

Tõnu Kollo
Vice-Chair of the Programme Committee

A short history of the International Workshop on Matrices and Statistics (IWMS)

Simo Puntanen[1] and George P. H. Styan[2]

[1]*School of Information Sciences, University of Tampere, Finland, email: simo.puntanen@uta.fi*
[2]*McGill University, Canada, email: styan@math.mcgill.ca*

Abstract

We present a short history of the International Workshop on Matrices and Statistics (IWMS). The first IWMS was held in Tampere, Finland, in 1990, and the Workshop this year, the 20th IWMS, is being held in Tartu, Estonia, 27–30 June 2011. We have established an open-access website for the IWMS at the University of Tampere: `http://mtl.uta.fi/iwms/` where we plan to put all associated reports and photographs of the IWMS from 1990 onwards, including those published in *Image: The Bulletin of the International Linear Algebra Society.*

The first workshop in the "International Workshop on Matrices and Statistics" (IWMS) series took place at the University of Tampere in Tampere, Finland, 6–8 August 1990. This workshop was organized by a local committee from the Statistics Unit of the Department of Mathematical Sciences at the University of Tampere. The key persons in the organizing committee were Pentti Huuhtanen, Erkki Liski, Tapio Nummi, Tarmo Pukkila, Simo Puntanen, and George P. H. Styan. There was no idea at that time that this would be the beginning of an almost annual series of meetings. This first IWMS was actually called "The International Workshop on Linear Models, Experimental Designs, and Related Matrix Theory". Since 1990 the name has changed twice, and in 1998 the IWMS became the "International Workshop on Matrices and Statistics", following a suggestion by C. Radhakrishna Rao.

In 1990 in Tampere there were 98 participants from 18 different countries. The Keynote Address in 1990 was given by C. Radhakrishna Rao. The invited speakers were

Jerzy K. Baksalary	Sujit Kumar Mitra	Friedrich Pukelsheim
R. Dennis Cook	Seppo Mustonen	Jagdish N. Srivastava
Yadolah Dodge	Heinz Neudecker	George P. H. Styan
Shanti S. Gupta	Ingram Olkin	

The organizers of the group meetings were

Jerzy K. Baksalary	Sanpei Kageyama	Kirti R. Shah
Tadeusz Caliński	Jürgen Kleffe	George P. H. Styan
R. Dennis Cook,	Sujit Kumar Mitra	Götz Trenkler
R. William Farebrother	Seppo Mustonen	Song-Gui Wang
Yasunori Fujikoshi	Friedrich Pukelsheim	Haruo Yanai
T. P. Hettmansperger	Jorma Rissanen	

Many of these persons have also been active participants in later workshops. George P. H. Styan has missed only one IWMS in 1990–2010 and thereby has the highest score in the attended IWMSs.

The following is an up-to-date version of the aims of the IWMS:

> The purpose of the IWMS is to stimulate research and, in an informal setting, to foster the interaction of researchers in the interface between statistics and

matrix theory. The Workshop will provide a forum through which statisticians may be better informed of the latest developments and newest techniques in linear algebra and matrix theory and may exchange ideas with researchers from a wide variety of countries.

Quite soon after the 2nd IWMS in Auckland, New Zealand, in 1992, the organizing system for the IWMS found its form as two committees: International and Local. The International Organizing Committee (IOC) for several years comprised R. William Farebrother (UK), Simo Puntanen (Finland), George P. H. Styan (Canada), Hans Joachim Werner (Germany). Recently, also S. Ejaz Ahmed (Canada), Jeffrey J. Hunter (New Zealand), Augustyn Markiewicz (Poland), Götz Trenkler (Germany), Júlia Volaufová (USA), and Dietrich von Rosen (Sweden), have joined the IOC; in 2008 George P. H. Styan was named Honorary Chair of the IOC of the IWMS. It is of course worth emphasizing that a most demanding task and responsibility for the meeting arrangements belongs to the local organizing committee.

The IWMS series has had three *ILAS Lecturers*: Gene H. Golub (1999), Jerzy K. Baksalary (2003), and Ravindra B. Bapat (2008), and two *Nokia Lecturers*: Ingram Olkin (2004) and C. Radhakrishna Rao (2005). As the IWMS Birthday Boys have been celebrated T. W. Anderson (80, 90), Ingram Olkin (80), C. Radhakrishna Rao (80), George P. H. Styan (60, 65), and a Special Session for Tarmo Pukkila (60) was held in 2006. Memorial Sessions have been held for Bernhard Flury (1999), Sujit Kumar Mitra (2004), and Jerzy K. Baksalary (2005).

We now present a list of the 19 Workshops that have been held from 1990 to 2010, as well as the 2011 Workshop in Tartu, Estonia. The photographs 1, 6, 7, 9 and 12 are taken by the University of Tampere photographer, photograph 4 by Hazel Hunter and the others by Simo Puntanen.

PHOTO 1: Group of participants in IWMS-1990, Tampere; C. Radhakrishna Rao inviting more people to the picture.

1990/1: International Workshop on Linear Models, Experimental Designs, and Related Matrix Theory Tampere, Finland, 6–8 August 1990, $n = 98$.
Chair of the Organizing Committee: Erkki Liski. Programme.

1992/2: [2nd] International Workshop on Matrix Methods for Statistics, Auckland, New Zealand, 4–5

December 1992, $n = 23$.
Chair of the Organizing Committee: Alastair J. Scott. Report in *Image.*

1994/3: Tartu Satellite Workshop on Matrices in Statistics, Tartu, Estonia, 28 May 1994.
Local Chairs: Ene-Margit Tiit & Hannu Niemi, IOC Chair: George P. H. Styan.

1995/4: 4th International Workshop on Matrix Methods for Statistics, Montréal, Québec, Canada, 15–16 July 1995, $n = 70$.
Local and IOC Chair: George P. H. Styan. Report in *Image.*

PHOTO 2: IWMS-4, Montréal, 15–16 July 1995.

1996/5: 5th International Workshop on Matrix Methods for Statistics, Shrewsbury, England, 18–19 July 1996, $n = 28$.
Local and IOC Chair: R. William Farebrother. Programme. Report in *Image.*

1997/6: 6th International Workshop on Matrix Methods for Statistics, Istanbul, Turkey, 16–17 August 1997, $n = 40$.
Local Chair: Fikri Akdeniz, IOC Chair: Hans Joachim Werner. Report in *Image.*

1998/7: 7th International Workshop on Matrices and Statistics, Fort Lauderdale, Florida, USA, 11–14 December 1998, $n = 78$,
in celebration of *T. W. Anderson's 80th birthday.* Programme.
Local Chair: Fuzhen Zhang, IOC Chair: George P. H. Styan. Report in *Image.*

PHOTO 3: George P. H. Styan, T. W. Anderson, Fuzhen Zhang; Fort Lauderdale, December 1998.

1999/8: 8th International Workshop on Matrices and Statistics, Tampere, Finland, 7–8 August 1999, $n = 95$.
☐ The ILAS Lecturer: *Gene H. Golub.*
Memorial Session: *Bernhard Flury* (1951–1999).
Local and IOC Chair: Simo Puntanen. Programme.
`http://www.uta.fi/laitokset/mattiet/workshop99/`, Report in *Image.*

2000/9: 9th International Workshop on Matrices and Statistics, Hyderabad, India, 9–13 December 2000, $n = 100$,
in celebration of *C. Radhakrishna Rao's* 80th birthday.
Local Chairs: S. B. Rao, P. Bhimasankaram, IOC Chair: Hans Joachim Werner.
Programme. Report in *Image.*

2001/10: 10th International Workshop on Matrices and Statistics, Voorburg, The Netherlands, 2–3 August 2001, $n = 54$. Report in *Image.*
Local Chair: Patrick J. F. Groenen, IOC Chair: George P. H. Styan.

2002/11: 11th International Workshop on Matrices and Statistics, Lyngby, Denmark, 29–31 August 2002, $n = 65$,
in celebration of *George P. H. Styan's* 65th birthday. Report in *Image.*
Local Chair: Knut Conradsen, IOC Chair: Hans Joachim Werner.

2003/12: 12th International Workshop on Matrices and Statistics, Dortmund, Germany, 5–8 August 2003, $n = 45$.
☐ The ILAS Lecturer: *Jerzy K. Baksalary.*
Programme. Report in *Image.*
Local Chair: Götz Trenkler, IOC Chair: Hans Joachim Werner.

2004/13: 13th International Workshop on Matrices and Statistics, Bedlewo, Poznań, Poland, 18–21 August 2004, $n = 82$,
in celebration of *Ingram Olkin's* 80th birthday,
☐ The Nokia Lecturer: *Ingram Olkin.*
Memorial Session: *Sujit Kumar Mitra* (1932–2004).
Local Chair: Augustyn Markiewicz, IOC Chair: Simo Puntanen.
`http://matrix04.amu.edu.pl/`, Programme. Report in *Image.* Poster.

PHOTO 4: IWMS-13, Bedlewo, Poznań, Poland, 18–21 August 2004.

2005/14: 14th International Workshop on Matrices and Statistics, Massey University, Albany Campus, Auckland, New Zealand, 30 March – 1 April 2005, $n = 50$.
☐ The Nokia Lecturer: *C. Radhakrishna Rao.*
Memorial Session: *Jerzy K. Baksalary* (1944–2005).
Local Chair: Jeffrey J. Hunter, IOC Chair: George P. H. Styan.
`http://iwms2005.massey.ac.nz/`, Announcement. Programme. Report in *Image.* Flyer.

2006/15: 15th International Workshop on Matrices and Statistics, Uppsala, Sweden, 13–17 June 2006, $n = 68$.
Special Session for *Tarmo Pukkila's* 60th birthday.
Local Chair: Dietrich von Rosen, IOC Chair: Hans Joachim Werner.
`http://www.bt.slu.se/iwms2006/iwms06.html`, Programme. Report in *Image.*

Photo 5: IWMS-15, Uppsala, Sweden, 13–17 June 2006.

2007/16: 16th International Workshop on Matrices and Statistics, Windsor, Ontario, Canada, 1–3 June 2007, $n = 74$,
in celebration of *George P. H. Styan's* 70th birthday,
Local Chair: S. Ejaz Ahmed, IOC Chair: George P. H. Styan.
`http://www.uwindsor.ca/units/iwms/main.nsf`, Programme. Poster.

2008/17: 17th International Workshop on Matrices and Statistics, Tomar, Portugal, 22–26 July 2008, $n = 80$,
in celebration of *T. W. Anderson's* 90th birthday.
☐ The ILAS Lecturer: *Ravindra B. Bapat*.
Local Chair: João T. Mexia, IOC Chair: Simo Puntanen.
`http://www.iwms08.ipt.pt/`, Programme. Report in *Image*. Poster.

2009/18: 18th International Workshop on Matrices and Statistics, Smolenice Castle, Slovakia, 23–27 June 2009, $n = 67$.
Local Chair: Viktor Witkovský, IOC Chair: Júlia Volaufová.
`http://www.um.sav.sk/en/iwms2009.html`, Programme. Poster.

2010/19: 19th International Workshop on Matrices and Statistics, Shanghai, China, 5–8 June 2010, $n = 186$.
Local Chair: Yonghui Liu, IOC Chair: Jeffrey J. Hunter.
`http://www1.shfc.edu.cn/iwms/index.asp`, Programme. Report in *Image*.

2011/20: 20th International Workshop on Matrices and Statistics, with the Tartu 9th Conference on Multivariate Statistics, Tartu, Estonia, 27–30 June 2011,
in celebration of *Muni S. Srivastava's* 75th birthday.
Local Chair: Kalev Pärna, Programme Committee Chair: Dietrich von Rosen,
Vice-Chair: Tõnu Kollo.
`http://www.ms.ut.ee/tartu11/`

Selected refereed papers presented at the IWMS have been (or are about to be) published in the following journal special issues:

1992: Third Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 176 (1992), viii + 289 pp. (Includes 8 research papers presented at the Workshop held in Tampere, Finland, 6–8 August 1990.) Preface. DOI.
Jerzy K. Baksalary & George P. H. Styan, eds.

1993: *Journal of Statistical Planning and Inference*, vol. 36, no. 2–3 (1993), pp. 127–432. (24 research papers presented at the Workshop held in Tampere, Finland, 6–8 August 1990.) Preface. Author index. DOI.
Jerzy K. Baksalary & George P. H. Styan, eds.

1994: Fourth Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 210 (1994), 273 pp. Preface. DOI.
Jeffrey J. Hunter, Simo Puntanen & George P. H. Styan, eds.

1996: Fifth Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, In Celebration of C. Radhakrishna Rao's 75th Birthday. vol. 237/238 (1996), vii + 273 pp. Author index.
Ravindra B. Bapat, George P. H. Styan & Hans Joachim Werner, eds.

1997: Sixth Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 264 (1997), ix + 506 pp. Preface. DOI.
R. William Farebrother, Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

1999: Seventh Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 289 (1999), iv + 344 pp. Preface. DOI.
R. William Farebrother, Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

2000: Eighth Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 321 (2000), xi + 412 pp. Preface. DOI.
Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

2002: Ninth Special Issue on Linear Algebra and Statistics: *Linear Algebra and its Applications*, vol. 354 (2002), xii + 291 pp. Preface. DOI.
Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

2004: Tenth Special Issue on Linear Algebra and Statistics, Part 1: *Linear Algebra and its Applications*, vol. 388 (2004), 400 pp. Preface. DOI.
Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

2005: Tenth Special Issue on Linear Algebra and Statistics, Part 2: *Linear Algebra and its Applications*, vol. 410 (2005), 290 pp. Preface. DOI.
Simo Puntanen, George P. H. Styan & Hans Joachim Werner, eds.

2005: *Research Letters in the Information and Mathematical Sciences*, vol. 8 (2005), v + 228 pp. Special Issue: Proceedings of the 14th International Workshop on Matrices and Statistics, Auckland, New Zealand, 30 March–1 April 2005. Foreword. Available online.
Jeffrey J. Hunter & George P. H. Styan, eds.

2006: *Linear Algebra and its Applications*, vol. 417 (2006), Proceedings of the 13th International Workshop on Matrices and Statistics, Bedlewo, Poznań, Poland, 18–21 August 2004. Preface. DOI.
Ludwig Elsner, Augustyn Markiewicz & Tomasz Szulc, eds.

2009: *Linear Algebra and its Applications*, vol. 430, no. 10 (2009), pp. 2563–2834, Proceedings of the 16th International Workshop on Matrices and Statistics, Windsor, Ontario, Canada, 1–3 June 2007. Preface. DOI.
S. Ejaz Ahmed, Jeffrey J. Hunter, George P. H. Styan & Götz Trenkler, eds.

2010: *Acta et Commentationes Universitatis Tartuensis de Mathematica*, vol. 14, 2010. Proceedings of the 18th International Workshop on Matrices and Statistics, Smolenice Castle, Slovakia, 23–27 June 2009.
Tõnu Kollo, Dietrich von Rosen, Viktor Witkovský & Júlia Volaufová, eds.

2011: *Numerical Linear Algebra with Applications* will publish a special issue as the Proceedings of the 19th International Workshop on Matrices and Statistics, Shanghai, China, 5–8 June 2010.
Åke Björck, Maya Neytcheva, Musheng Wei & Yonghui Liu, eds.

2011: *Acta et Commentationes Universitatis Tartuensis de Mathematica*, also welcomes papers presented at the Shanghai IWMS.
Tõnu Kollo & Dietrich von Rosen, eds.

2012: IWMS-20 with the Tartu 9th Conference on Multivariate Statistics: Special volume by World Scientific from selected papers of the Conference, Special volume of *Acta et Commentationes Universitatis Tartuensis de Mathematica*.

Photo 6: Jerzy K. Baksalary giving a talk in Tampere, August 1990.

Photo 7: R. Dennis Cook, Norman Draper, Nye John, George P. H. Styan; Tampere, August 1990.

Photo 8: Ravindra B. Bapat, Tomar, July 2008.

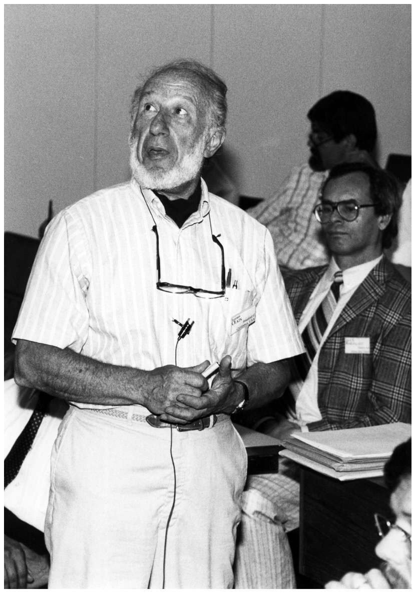

Photo 9: Ingram Olkin, Tampere, August 1990.

Photo 10: C. Radhakrishna Rao, Hyderabad, December 2000.

Photo 11: C. Radhakrishna Rao and Bhargavi Rao, Hyderabad, December 2000.

PHOTO 12: Jerzy K. Baksalary, Tadeusz Caliński, Sujit Kumar Mitra; Tampere, August 1990.

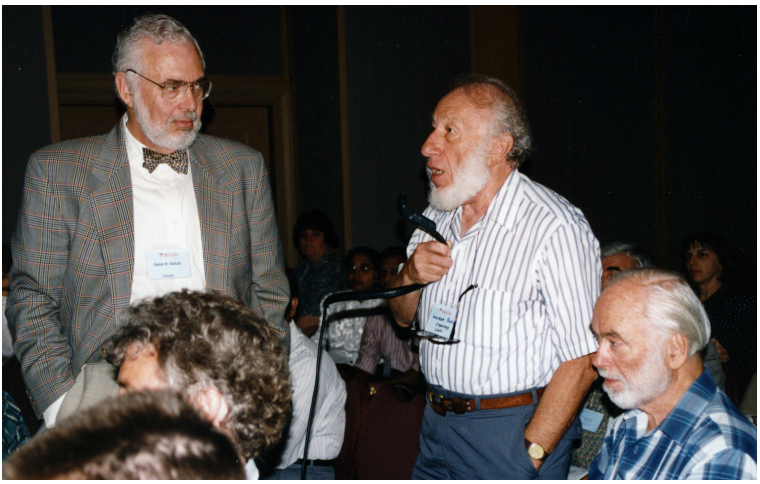

PHOTO 13: Gene H. Golub, Ingram Olkin, T. W. Anderson; Montréal, July 1995.

PHOTO 14: Group of participants in Bedlewo, August 2004.

PHOTO 15: Enjoying the conference banquet (and the view to Detroit) in Windsor, June 2007.

Photo 16: In an after-dinner session in Smolenice Castle in July 2009, Tõnu Kollo (smiling in the picture) tentatively agreed to organize the IWMS-2011 in Tartu. Left: Soile Puntanen, right: Miroslav Fiedler.

Index