Seventh International Workshop on Simulation

21-25 May, 2013

Department of Statistical Sciences, Unit of Rimini

University of Bologna, Italy

Book of Abstracts

Edited by Mariagiulia Matteucci

Quaderni di Dipartimento

Serie Ricerche 2013, n. 3

ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche “Paolo Fortunati”

The Unit of Rimini of the Department of Statistical Sciences of Bologna University in collaboration with the Department of Management and Engineering of the University of Padova, the Department of Statistical Modelling of Saint Petersburg State University and INFORMS Simulation Society are sponsoring the Seventh International Workshop on Simulation. This international conference is devoted to statistical techniques in stochastic simulation, data collection and analysis of scientific experiments and studies representing broad areas of interest. All the previous Workshops took place in St. Petersburg (Russia). The first Workshop took place in May 1994, the second Workshop in June 1996, the third in June 1998, the fourth in June 2001, the fifth in June 2005 and the sixth in June 2009.

Scientific Program Committee

Paola Monari (Italy), Chair
Viatcheslav B. Melas (Russia), Chair
Luigi Salmaso (Italy), Chair

Alexander Andronov (Latvia)
Narayanaswamy Balakrishnan (Canada)
Michel Broniatowski (France)
Ekaterina Bulinskaya (Russia)
Holger Dette (Germany)
Sergei M. Ermakov (Russia)
Valerii Fedorov (USA)
Nancy Flournoy (USA)
Subir Ghosh (USA)
Simone Giannerini (Italy)
Michele La Rocca (Italy)
Stefania Mignani (Italy)
Gennady Mikhailov (Russia)
Valery Nevzorov (Russia)
Michael Nikulin (France)

Ingram Olkin (USA)
Fortunato Pesarin (Italy)
Domenico Piccolo (Italy)
Dieter Rasch(Germany)
Rainer Schwabe (Germany)
John Stufken (USA)

Local Organizing Committee

Stefania Mignani (Italy), Chair
Paola Monari (Italy), Chair

Rosa Arboretti (Italy)
Stefano Bonnini (Italy)
Livio Corain (Italy)
Simone Giannerini(Italy)
Mariagiulia Matteucci (Italy)
Luisa Stracqualursi (Italy)

The conference is scientifically sponsored by:

- Department of Statistical Sciences, University of Bologna
- Department of Management and Engineering,University of Padova

- Department of Statistical Modelling, Saint Petersburg State University
- INFORMS Simulation Society (USA)
- Italian Statistical Society

The conference is sponsored by:

- Comune di Rimini
- Provincia di Rimini
- Regione Emilia-Romagna

List of abstracts

Statistical Estimation of Random Field Thresholds Using Euler Characteristics

Robert J. Adler, Anthea Monod
Technion – Israel Institute of Technology
robert@ee.technion.ac.il, anthea@ee.technion.ac.il

Kevin Bartz, S. C. Kou
Harvard University
kevin@kevinbartz.com, kou@stat.harvard.edu

In many applications of random field models, such as brain imaging and cosmological studies, an important problem is the determination of thresholds: regions where values occur above the threshold indicate significance, while values occurring below do not. In the setting of random fields, the existence of spatial correlation between values renders this problem particularly challenging.

Statistically, this problem can be posed as a test: the null hypothesis H_0 asserts equivalence over a region S between two realizations of a smooth random field $T(\cdot)$. High values of T over S for one of the realizations thus indicate deviation from H_0, inspiring the use of the maximum as a test statistic, $M_S := \sup_{s \in S} T(s)$, which requires its null distribution. In addition, identifying regions with 95% significant values of T requires a 5% threshold t such that $P(M_S > t | H_0) = 0.05$. Obtaining such unknown quantities is difficult since the correlation structure of the random field is required, which is itself also unknown.

In Adler *et al.* (2013), we introduce Lipschitz-Killing curvature (LKC) regression, a new method to produce accurate $(1-\alpha)$ thresholds for random fields without knowledge of the correlation structure. The method borrows from the Euler characteristic heuristic (ECH) (Adler, 2000), a powerful parametric technique for Gaussian random fields to determine null tail probabilities of M_S and thus provides an accurate approximation of the exceedance probability $P(M_S \geq u)$ for large u. The idea

is to fit the observed empirical Euler characteristics φ of excursion sets A_u to the Gaussian kinematic formula (GKF) (Taylor, 2006),

$$E[\varphi(A_u)] = \sum_{i=0}^{\dim(S)} \mathcal{L}_i(S)\rho_i(u). \quad (1)$$

For Gaussian random fields, the ρ_i take an explicit form based on Hermite polynomials, while the $\mathcal{L}_i$ are the LKCs: complex topological quantities that are very challenging to evaluate both theoretically and numerically. Our method quickly and easily provides statistical estimates of the LKCs via generalized least squares (GLS), which then by the GKF (1) and the ECH generate $(1-\alpha)$ thresholds and p-values. Furthermore, LKC regression achieves large gains in speed without loss of accuracy over its main competitor, warping (Taylor & Worsley, 2006).

Keywords: Excursion set, Gaussian kinematic formula, Lipschitz-Killing curvature, generalized least squares regression, significance level.

Acknowledgements: This research was supported in part by US-IL BSF, ISF, NIH-NIGMS, and NSF. Anthea Monod's research was supported by TOPOSYS (FP7-ICT-318493-STREP). The authors would like to thank Jonathan Taylor for helpful discussions at various stages of this work.

References

Adler R.J. (2000): On Excursion Sets, Tube Formulae, and Maxima of Random Fields, *Annals of Applied Probability*, Vol. 10, N. 1, pp. 1–74.

Adler R.J., Bartz K., Kou S.C., Monod A. (2013): Estimating Thresholding Levels for Random Fields via Euler Characteristics, *in preparation.*

Taylor J. (2006): A Gaussian kinematic formula, *Annals of Probability*, Vol. 34, N. 1, pp. 122–158.

Taylor J., Worsley K. (2006): Inference for Magnitudes and Delays of Responses in the FIAC Data Using BRAINSTAT/ FMRISTAT, *Human Brain Mapping*, Vol. 27, N. 5, pp. 434–441.

Random walk in random environment conditioned to be positive: limit theorem for maximum

Afanasyev V.I.
Steklov Institute, Moscow, Russia
viafan@mail.ru

Let (p_i, q_i), $i \in \mathbb{Z}$, be a sequence of independent and identically distributed random vectors such that $p_i + q_i = 1$, $p_i > 0$, $q_i > 0$ for $i \in \mathbb{Z}$. Consider a random walk in the random environment (p_i, q_i), $i \in \mathbb{Z}$. It means that if the random environment is fixed then a moving particle fulfils a transition from the state i to the state $(i+1)$ with probability p_i or to the state $(i-1)$ with probability q_i. Let X_n be a position of the moving particle at time n and $X_0 = 0$.

Suppose that

$$\mathbf{E} \ln \frac{p_0}{q_0} = 0, \quad \mathbf{E} \ln^2 \frac{p_0}{q_0} := \sigma^2, \ 0 < \sigma^2 < \infty. \tag{1}$$

The well-known Ritter theorem establishes that if the condition (1) is valid then, as $n \to \infty$,

$$\frac{\sigma^2 \max_{0 \le i \le n} X_i}{\ln^2 n} \xrightarrow{d} \beta,$$

where β is a positive random value and the symbol $\xrightarrow{d}$ means convergence in distribution.

The conditional version of this theorem is valid (Afanasyev, 2013).

Theorem 1. *If the condition* (1) *is valid then*, *as* $n \to \infty$,

$$\left\{ \frac{\sigma^2 \max_{0 \le i \le n} X_i}{\ln^2 n} \,\middle|\, X_1 \ge 0, \dots, X_n \ge 0 \right\} \xrightarrow{d} \eta,$$

where η is a positive random value and for $x > 0$

$$\mathbf{P}\left(\eta \le x\right) = 4\sqrt{\frac{2}{\pi x}}\sum_{k=1}^{\infty}\exp\left(-\frac{2\left(2k-1\right)^2}{x}\right).$$

Our proof of the theorem is based on the following facts. Let T_n be the passage time of the state $n \in \mathbb{N}$ by the random walk $\{X_n\}$ and $\left\{Z_i^{(n)},\ i \in \mathbb{N}_0\right\}$ be a branching process in random environment with n immigrants (one immigrant in each generation beginning from zero generation) with reproduction distribution $\left\{p_{i+1}q_{i+1}^k,\ k \in \mathbb{N}_0\right\}$ of a representative of the i-th generation, $i \in \mathbb{N}_0$. It is known that for $n \in \mathbb{N}$

$$T_n \stackrel{d}{=} n + 2\sum\nolimits_{i=0}^{+\infty} Z_i^{(n)}. \tag{2}$$

Introduce a process $U_n\left(t\right) = \left(\sigma\sqrt{n}\right)^{-1}\ln\left(Z_{\lfloor nt\rfloor}^{(n)}+1\right),\ t \in \left[0,1\right]$. Theorem 1 is a corollary of the relation (2) and the following theorem.

Theorem 2. *If the condition* (1) *is valid then, as $n \to \infty$,*

$$\left\{U_n \mid\ U_n\left(1\right) = 0\right\} \stackrel{D}{\to} \left|W_0\right|,$$

where W_0 is a Brownian bridge and the symbol $\stackrel{D}{\to}$ means convergence in distribution in the space $D\left[0,1\right]$ with the Skorokhod topology.

Keywords: limit theorems, random walk in random environment, branching process with immigration in random environment.

Acknowledgements: This work was supported by the Program of RAS "Dynamical systems and control theory".

References

Afanasyev V.I. (2013): Conditional limit theorem for maximum of RWRE, *Theory Probab. Appl.*, Vol. 58 (in print).

Queueing Systems with Unreliable Servers in a Random Environment

Larisa Afanasyeva
Lomonosov Moscow State University
l.g.afanaseva@yandex.ru

Elena Bashtova
Lomonosov Moscow State University
bashtovaelena@rambler.ru

We consider a single-server system with an unreliable server. In such a system service is subjected to interruptions that are caused by breakdowns of the server. After breakdown the server is repaired during a random time. We suppose that the service interrupted by the breakdown of the server is continued after its repair from the point at which it was interrupted. Input flow $A(t)$ is assumed to be regenerative in the following sense.

Definition 1. A stochastic flow $A(t)$ is regenerative if there exists an increasing sequence of r.v.'s $\{\theta_j\}_{j=1}^{\infty}$, $\theta_0 = 0$, such that the sequence $\{\kappa_j\}_{j=1}^{\infty} = \{\theta_j - \theta_{j-1},\ A(\theta_{j-1}+t) - A(\theta_{j-1}),\ t \in [0, \theta_j - \theta_{j-1})\}_{j=1}^{\infty}$ consists of i.i.d. random elements.

Let $\xi_j = A(\theta_j) - A(\theta_{j-1})$, $\tau_i = \theta_j - \theta_{j-1}$. We assume that $\mu = \mathsf{E}\tau_1 < \infty$, $a = \mathsf{E}\xi_1 < \infty$ and put $\lambda = a/\mu$.

Service times $\{\beta_i\}_{i=1}^{\infty}$ are i.i.d. r.v.'s not depending on $A(t)$. Let $\{u_n^{(1)}\}_{n=1}^{\infty}$ ($\{u_n^{(2)}\}_{n=0}^{\infty}$) be a sequence of working (non-working) intervals of the server. These sequences are independent and each of them consists of i.i.d. r.v.'s. Besides, they do not depend on $A(t)$ and $\{\beta_j\}_{j=0}^{\infty}$. Put $a_i = \mathsf{E}u_n^{(i)} < \infty$, $i = 1, 2$. Let $W(t)$ be the virtual waiting time process and $q(t)$ the number of customers in the system at time t. We introduce a traffic coefficient $\rho = \lambda b \alpha^{-1}$ with $\alpha = \frac{a_1}{a_1 + a_2}$.

For the case $\rho < 1$ we consider a time compression asymptotic describing heavy traffic situation.

We introduce a family of queueing systems $\{S_T\}$ with traffic coefficient ρ_T and let $\rho_T \uparrow 1$, as $T \to \infty$. The input $A_T(t)$ for S_T is defined by the relation $A_T(t) = A(\rho^{-1}(1 - 1/\sqrt{T})t)$. The traffic coefficient for S_T is of the form $\rho_T = 1 - 1/\sqrt{T}$. Denote by $q_T(t)$, $W_T(t)$ the processes $q(t)$, $W(t)$ for the system S_T with input $A_T(t)$.

Theorem 1. Let

$$\mathsf{E}\tau_n^{2+\delta} < \infty,\ \mathsf{E}\xi_n^{2+\delta} < \infty,\ \mathsf{E}\beta_n^{2+\delta} < \infty,\ \mathsf{E}(u_n^{(i)})^{2+\delta} < \infty \quad (1)$$

for some $\delta > 0$ and all $n \geq 1$, $i = 1, 2$. Then the normalized processes

$$\widehat{W}_T(t) = \frac{W_T(tT)}{\sqrt{T}} \text{ and } \widehat{q}_T(t) = \frac{q_T(tT)}{\sqrt{T}}$$

converge weakly to diffusion processes with reflection at the origin and coefficients $(-\sqrt{\alpha}b^{-1}, \widetilde{\sigma}_q^2)$, $(-\sqrt{\alpha}, b^2\widetilde{\sigma}_q^2)$ respectively. Here

$$\widetilde{\sigma}_q^2 = \frac{\alpha}{b^3}\sigma_\beta^2 + \frac{\alpha}{\lambda b}\sigma_A^2 + \frac{1}{b^2}\sigma_S^2,$$

$$\sigma_A^2 = \sigma_\xi^2\mu^{-1} + a^2\sigma_\tau^2\mu^{-3} - 2a\mathsf{cov}(\xi, \tau)\mu^{-2},\ \ \sigma_S^2 = \frac{a_1^2\sigma_2^2 + a_2^2\sigma_1^2}{(a_1 + a_2)^3},$$

and σ_ξ^2, σ_τ^2, $\sigma_i^{(2)}$ are variances of ξ, τ, $u_n^{(i)}$ respectively.

Theorem 2. Let conditions (1) be fulfilled. If $\rho > 1$ then for any finite interval $[0, h]$ the normalized processes

$$\widehat{W}_T(t) = \frac{W(tT) - \alpha(\rho-1)tT}{\widetilde{\sigma}_W\sqrt{T}} \quad \text{and} \quad \widehat{q}_T(t) = \frac{q(tT) - b^{-1}\alpha(\rho-1)tT}{\widetilde{\sigma}_q\sqrt{T}}$$

converge weakly to Wiener processes, as $T \to \infty$.

If $\rho = 1$ then $\widehat{W}_T(t) = \frac{W(tT)}{\widetilde{\sigma}_W\sqrt{T}}$ and $\widehat{q}_T(t) = \frac{q(tT)}{\widetilde{\sigma}_q\sqrt{T}}$ converge weakly to absolute value of Wiener processes, as $T \to \infty$. Here

$$\widetilde{\sigma}_W^2 = b^2\sigma_A^2 + \lambda\sigma_\beta^2 + \sigma_S^2, \quad \widetilde{\sigma}_q^2 = \frac{1}{b^2}\widetilde{\sigma}_W^2.$$

Keywords: queueing system, unreliable server, regenerative input.

Sequential Combining of Expert Information using Mathematica

Patrizia Agati
Department of Statistics - University of Bologna
patrizia.agati@unibo.it

Luisa Stracqualursi, Paola Monari
Department of Statistics - University of Bologna
luisa.stracqualursi@unibo.it, paola.monari@unibo.it

Knowledge-gaining and decision-making in real-world domains often require reasoning under uncertainty. In such contexts, combining information from several, possibly heterogeneous, sources - 'experts', such as numerical models, information systems, witnesses, stakeholders, consultants - can really enhance the accuracy and precision of the 'final' estimate of the unknown quantity (a risk, a probability, a future random event, ...).

Bayesian paradigm offers a coherent perspective from which to address the problem. It just suggests to regard experts' opinions/outputs as data from an experiment (Morris, 1977): a likelihood function may be associated with them — and the Joint Calibration Model (JCM) makes it more easier to assess (Monari and Agati, 2001), (Agati *et al.*, 2007) — and is used to revise the prior knowledge. In such a way, the information combining process just becomes a knowledge updating process.

An issue strictly related to information combining is how to perform an efficient process of sequential consulting. The investigator, indeed, often prefers to consult the experts in successive stages rather than simultaneously: so, s/he avoids wasting time (and money) by interviewing a number of experts that exceeds what s/he needs. At each stage, the investigator can select the 'best' expert to be consulted and choose whether to stop or continue the consulting.

The aim of this paper is to rephrase the Bayesian combining algo-

rithm in a sequential context and use `Mathematica` to implement suitable selecting and stopping rules. The procedure implemented in a notebook file has been investigated in simulation and experimental studies.

Keywords: combining, knowledge, curvature, Mathematica.

References

Agati P., Calò D.G., Stracqualursi L. (2007): A joint calibration model for combining predictive distributions. *Statistica*, Vol. 2, pp. 203-212.

Kullback S. (1959). *Information Theory and Statistics*: Wiley, New York.

Lindley D.V. (1990): The 1988 Wald Memorial Lectures: the present position in Bayesian statistic. *Statistical Science*, Vol.5, pp. 44-89.

McCulloch R. (1989): Local Model Influence. In *Journal of the American Statistical Association*, pp. 473–478.

Monari P., Agati P. (2001): Fiducial inference in combining expert judgements. *Journal of the Italian Statistical Society*, pp. 81–97.

Morris P.A. (1977): Combining expert judgments: a bayesian approach. *Management Science*, Vol.23, pp. 679–693.

Wolfram S. (2003): *The `Mathematica` book. Fifth Edition-*. Wolfram Media.

On Some New Record Schemes

M. Ahsanullah
Rider University, Lawrenceville, NJ, USA.
ahsan@rider.edu

V.B. Nevzorov
St-Petersburg State University, St-Petersburg, Russia.
vanev@mail.ru

Let $X_1, X_2, \ldots$ be a sequence of random variables (r.v.) , $1 = L(1) < L(2) < \ldots$ and $X(n) = \max\{X_1, X_2, \ldots, X_{L(n)}\}$, $n = 1, 2, \ldots$, be correspondingly the upper record times and the upper record values. The classical theory of records which deals with records for independent identically distributed (i.i.d.) X's is very popular and is developed very well. There are a lot of monographs devoted to records (see, for example, Nevzorov (2000), Ahsanullah and Yanev (2008)). Indeed, for more than 60 years of their history the classical records were studied carefully in all directions. The "classical" part of the record theory initiated the necessity to study "nonclassical" models. We suggest here one new "nonclassical" record model- "records with restrictions". This scheme is close to the model of "δ-exceedance records", which was considered by Balakrishnan, Balasubramanian and Panchapakesan (1996).

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables and C be some fixed positive constant. We define $X(1) = X_1$, $L(1) = 1$ and the next after $X(n) = X_{L(n)}$ value $X(n+1)$ is determined as follows. In the sequence $X_{L(n)+1}$, $X_{L(n)+2}$, as the new record variable we take the first of X's (its number we denote as $L(n+1)$) which lies in the random interval $(X(n),\ X(n)+C)$. It means that we ignore all X's which are less than $X(n)$ or which are greater than $X(n)+C$. In one more modification of this record scheme with restrictions the next record $X(n+1)$ coincides with the first X, which is greater than $X(n)$, if X lies in the interval $(X(n),\ X(n)+C)$, and $X(n+1) = X(n)+C$, if

this X is more than $X(n) + C$. Some results are obtained for both of these record schemes.

Keywords: record times, record values, δ-exceedance records, records with restrictions.

Acknowledgements: The work of the first author was partially supported by Summer Research Grant of Rider University. The work of the second author was partially supported by grant NSh-1216.2012.1 and by program 2010-1.1-111-128-033.

References

Ahsanullah M. and Yanev G.P.(2008): *Records and Branching Processes*, Nova Science Publishers, New York.

Balakrishnan N., Balasubramanian K. and Panchapakesan S.(1996): *δ-Exceedance Records*, Journal of Applied Statistical Science, v.4, n.2/3, pp. 123-132.

Nevzorov V.B.(2000): *Records: Mathematical Theory*, Fazis, Moscow (in Russian).

Regression Properties of Sums of Record Values

Akhundov I.
Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Canada
iakhundo@uwaterloo.ca

Nevzorov V.B.
St. Petersburg State University, St. Petersburg, Russia
valnev@mail.ru

Let $X_1, X_2, ...$ be independent random variables (rv's) having a common continuous cumulative distribution function (cdf) $F(x)$, $1 = L(1) < L(2) < \cdots$ and $X(n) = \max\{X_1, X_2, ..., X_{L(n)}\}$ be the upper record times and the upper record values correspondingly. Let $N(n)$ denote the number of records among rv's $X_1, X_2, ..., X_n$ and $\xi_k, k = 1, 2, ...$ be the record indicators, i.e. $\xi_k = 1$ if $X_k = \max\{X_1, X_2, ..., X_k\}$, and $\xi_k = 0$ otherwise. If $S(n)$ denotes the sum of record values generated by $X_1, X_2, ..., X_n$ and $M(n) = \max\{X_1, X_2, ..., X_n\}, n = 1, 2, ...$ then it is not difficult to show that $S(n) = \sum_{k=1}^{n} M(k)\xi_k$. We will also deal with the sums $T(n) = X(1) + X(2) + \cdots + X(n),\ n = 1, 2, ...$

The theory of records is reviewed in details in monographs by Arnold, Balakrishnan and Nagaraja (1998), Nevzorov (2001), Ahsanullah and Nevzorov (2001), and Ahsanullah and Yanev (2008). In Akhundov and Nevzorov (2006), the authors characterize a class of distributions through the regression relation $E(T(n)|X(n+1) = u) = cu + d$ a.s.. In this paper we consider a regression relation which is close to the one given above but characterizes another set of distributions. Instead of sums $T(n)$ we deal with the sums $S(n)$. Indeed, $T(n) = S(L(n))$ where the record times $L(n)$ are rv's. It appears that

$$E(S(n)|X(N(n)) = x) = E(S(n)|M(n) = x)$$

$$= x + \tfrac{n-1}{n}\int_{-\infty}^{x} ydR(y,x) + \tfrac{n-2}{2n}\int_{-\infty}^{x} ydR^2(y,x) + \cdots$$

$$+\frac{1}{n(n-2)}\int_{-\infty}^{x} ydR^{n-1}(y,x),$$

where $R(y,x)=\frac{F(y)}{F(x)}, y \le x$ and $R(y,x)=1$ otherwise. This equality allows us to obtain some new characterizations of probability distributions. The simplest of these characterizations (the case of $n=2$) is given by the following

Theorem. *Let F be a continues cdf. For some $\alpha>1$ and $-\infty<\beta<\infty$ the regression relation $E(S(2)|M(2)=x)=\alpha x+\beta \quad a.s.$ holds if and only if F (up to location and scale parameters) is given by*
a. $F(x)=0, x\le 0$; $F(x)=x^{\delta}, 0<x\le 1$ *and* $F(x)=1, x>1$ *if* $1<\alpha<3/2$;
b. $F(x)=\exp(x), -\infty<x\le 0$ *and* $F(x)=1,\ x>0$ *if* $\alpha=3/2$;
c. $F(x)=(-x)^{\delta}, -\infty<x\le -1$ *and* $F(x)=1,\ x>-1$ *if* $\alpha>3/2$, *where* $\delta=2(\alpha-1)/(3-2\alpha)$. *Keywords:* records, sum, characterization, regression.

Acknowledgements: The second author's work was partially supported by grant NSh-1216.2012.1 and by program 2010-1.1-111-128-033.

References

Ahsanullah M., Nevzorov V.B. (2001): *Ordered Random Variables*, Nova Science Publishers, NY.

Akhundov I., Nevzorov V.B. (2006): Conditional Distributions of record values and characterizations of distributions. *Probability and Statistics, 10* (Notes of Sci Seminar POMI, v.339), 5-14 (in Russian).

Arnold B.C., Balakrishnan N., Nagaraja H.N., (1998): *Records*, Wiley, NY.

Ahsanullah M., Yanev G.P. (2008): *Records and Branching Processes*, Nova Science Publishers, NY.

Nevzorov V.B. (2001): *Records: Mathematical Theory*. Translations of mathematical monographs, v.194. American Math. Soc.

Asymptotic proprieties of a Randomly Reinforced Urn Design targeting any fixed allocation

Giacomo Aletti
Universitá degli Studi di Milano
giacomo.aletti@unimi.it

Andrea Ghiglietti, Anna Maria Paganoni
Politecnico di Milano
andrea.ghiglietti@mail.polimi.it, anna.paganoni@polimi.it

There are many experimental designs in clinical trials, in which the proportion of patients allocated to treatments converges to a fixed value. Some of these procedures are adaptive and the limiting proportion can depend on the treatments' behaviors. Adaptive designs are attractive because they aim to achieve two simultaneous goals: (a) collecting evidence to determine the superior treatment, and (b) increasing the allocation of units to the superior treatment. For a complete literature review on response adaptive designs see Hu, Rosenberger (2006). We focus on a particular class of adaptive designs, described in terms of urn models which are randomly reinforced and presenting a diagonal mean replacement matrix. These models introduced by Durham (1990) for binary responses, were extended to the case of continuous responses by Muliere, Paganoni, Secchi (2006). In that work it was proved that the probability to allocate units to the best treatment converges to one as the sample size increases. Important results on the asymptotic behavior of the urn proportion for this RRU model were developed in Flournoy, May (2009), in the case of reinforcements with different expected values. We construct a new response adaptive design, described in terms of two colors urn model, targeting fixed asymptotic allocations that are function of treatments' performances. The model and the main convergence theorem are presented in Aletti, Ghiglietti, Paganoni (2013). We prove

asymptotic results for the process of colors generated by the urn and for the process of its compositions, concerning almost sure convergence and the convergence rates (Ghiglietti, Paganoni (2012)). Applications to sequential clinical trials, connections with response-adaptive design of experiments are considered as well as simulation studies concerning the power function of a testing hypothesis procedure that naturally arises from this statistical framework are detailed.

Keywords: Response adaptive designs, Clinical trials, Randomly Reinforced Urns.

References

Aletti G., Ghiglietti A., Paganoni A.M. (2013): A modified randomly reinforced urn design. Forthcoming in: *Journal of Applied Probability*, Vol. 50, No. 2. [Available as Mox-report 32/2011, Politecnico di Milano.

Durham S.C., Yu K.F. (1990): Randomized play-the leader rules for sequential sampling from two populations. *Probability in Engineering and Information Science*, 26(4), 355–367.

Flournoy N., May C. (2009): Asymptotic theorems of sequential estimation-adjusted urn models, *The Annals of Statistics*, 37, 1058–1078.

Ghiglietti A., Paganoni A. M. (2012): Statistical properties of two-color randomly reinforced urn design targeting fixed allocations. Available as Mox-report 48/2012, Politecnico di Milano.

Hu F., Rosenberger W. F. (2006): *The Theory of Response-Adaptive Randomization in Clinical Trials*, Wiley, New York.

Muliere P., Paganoni A.M., and Secchi P. (2006). A randomly reinforced urn. *Journal of Statistical Planning and Inference*, 136, 1853–1874.

An Ecological Study of Associations between Cancer Rates and Quality of Air and Streams

Raid Amin
University of West Florida, Pensacola Florida,USA.
ramin@uwf.edu

Nathaniel Hitt , Michael Hendryx , and Mathew Shull
US Geological Survey, Kearneysville, WV, USA.
West Virginia University, Morgantown, WV, USA.
University of West Florida, Pensacola Florida,USA.
nhitt@usgs.gov
mhendryx@hsc.wvu.edu
mls53@students.uwf.edu

This study illustrates the use of the software SaTScan for a spatio-temporal cluster analysis of cancer rates in Florida. The pediatric cancer data are obtained from the Florida Association of Pediatric Tumor Programs. The software SaTScan was developed as a modern disease surveillance tool for the detection of spatial clusters of a disease (Kulldorff 1997). Amin et al. (2010) provided an epidemiological study of pediatric cancer rates in Florida. They used incidence rates for pediatric cancers (2000-2007), adjusted for age and sex.

This study is an extension of Amin et al. (2010). It aims at identifying geographic areas that display high pediatric cancer incidence rates. In addition to the geospatial analysis of cancer rates, we will also analyze air quality data obtained from the Environmental Protection, and data on the quality of streams. The software ArcGIS is used to match cancer, air, and water samples based on their geographical coordinates in Florida. While our main goal still is to detect cancer clusters in Florida, it is very useful to test for possible associations with environmental factors, such as carcinogenic air pollution that is based on the Risk-Screening Environmental Indicators (RSEI),and/or the National Air Toxics Assessment (NATA), from the EPA. This is the only research project that the authors are aware of in which these factors are analyzed together for the USA.

Our approach for disease surveillance uses a sequential methodology as follows:

1. We start out having a cluster analysis for a large region, such as a State. In our project, we analyze data for Florida, using either census tracts or zip codes as the unit of analysis.

2. After we identify large sized clusters, the subsequent cluster analyses are applied to each of the significant clusters separately. In this case, the ”new population” is a cluster region. This way, we identify hot spots.

3. If smaller data units are available, such as in the case of air pollution, we start a new cluster analysis for each regional cluster, using data at the squared km level.

This sequential methodology allows us to go from large to small geographical areas of Florida, and it provides a sound approach to detect clusters.

This project is a disease surveillance study, using statistical methods to identify clusters. In our study, we test for univariate clusters and also for multivariate cluster analysis in SaTScan Multivariate spatial scan statistics for disease surveillance.

Keywords: Cancer, Cluster Analysis, Likelihood Ratio.

References

Amin R., Bohnert A, Holmes L., Rajasekaran A., and Assanasen C. (2010). Cover Page and Research Article in Pediatric Blood and Cancer. 2010 April;54(4): 511-8.Highlighted in ” Childhood cancer clustering in Florida: weighing the evidence,”

Kulldorff M. (1997) A spatial scan statistic. Communications in Statistics: Theory and Methods, 1481-1496.

Model-based clustering of multivariate longitudinal data

Laura Anderlucci, Cinzia Viroli
Department of Statistical Sciences 'Paolo Fortunati' - University of Bologna
laura.anderlucci@unibo.it, cinzia.viroli@unibo.it

Multivariate longitudinal data arise when different individual characteristics are investigated over time. When modelling this kind of data, correlation between measurements on each individual should be taken into account. When the same attributes are observed over time the statistical units can be arranged in a three-way data structure, where the n observations are the rows, the set of p attributes are represented in columns and the different time points, say T, are the layers. In this perspective, each observed unit is a $p \times T$ matrix instead of a conventional p-dimensional vector.

Suppose we are interested in clustering the n observed matrices in some k homogeneous groups, with $k < n$, using the full information of the temporal and the attribute entities. This is not a trivial problem, since correlations between variables could change across time and, vice versa, correlations between occasions can be different for each response. The issue of clustering longitudinal data in a model-based perspective has been recently addressed by McNicholas and Murhpy (2010), who proposed a family of Gaussian mixture models by parameterizing the class conditional covariance matrices via a modified Cholesky decomposition (Newton, 1998). This allows to interpret the observations as deriving from a generalized autoregressive process and to explicitly incorporate their temporal correlation into the model. The approach deals with the case of a single attribute measured over time.

In this work we consider the problem of clustering longitudinal data on multiple response variables. The issue can be addressed by means of matrix-normal distributions (Viroli, 2011). An explicit assumption of this approach is that the total variability can be decomposed into a

'within multiple attributes' and a 'between different times' component. This gives body to a separability condition of the total covariance matrix into two covariance matrices, one referred to the attributes and the other one to the times. According to McNicholas and Murphy (2010) we parameterize the class conditional 'between' matrices through the modified Cholesky decomposition, which has a nice statistical interpretation, since it allows to relate a realization at time t ($t = 1, \ldots, T$) to its past through a linear least-squares representation in a generalized autoregressive process. This mixture model can be fitted using an expectation-maximization (EM) algorithm and model selection can be performed by the BIC and the AIC information criteria. Effectiveness of the proposed approach has been tested through a large simulation study and an application to a sample of data from the Health and Retirement Study (HRS) survey.

Keywords: Multivariate longitudinal data, Model-based clustering, Three-way data.

References

McNicholas, P.,Murphy, B. (2010): Model-based clustering of longitudinal data, *The Canadian Journal of Statistics*, Vol. 38, pp. 153-168.

Newton, H.J. (1988), *TIMESLAB: A Time Series Analyis Laboratory*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Viroli, C. (2011): Finite mixtures of matrix normal distributions for classifying three-way data, *Statistics and Computing*, Vol. 21, pp. 511-522.

Markov-modulated samples and their applications

Alexander M. Andronov
Transport and Telecommunication Institute
lora@mailbox.riga.lv

Let us consider sample elements $\{X_i, i = 1, ..., n\}$, modulated by a finite continuous-time Markov chain (Pacheco, Tang, Prabhu, 2009). For simplicity we say that the elements operate in the so-called random *environment*. The last is described by an "external" time-continuous ergodic Markov chain $J(t), t \geq 0$, with a final state space $E = \{1, 2, ..., k\}$. Let $\lambda_{i,j}$ be the transition rate from state i to state j.

Additionally, n binary identical elements are considered. Each component can be in two states: *up* (1) and *down* (0). The elements of system fail one by one, in random order. For a fixed state $i \in E$, all n elements have the same failure rate $\gamma_i(t)$ and are stochastically independent. When the external process changes its state from i to j at some random instant t, all elements, which are alive at time t, continue their life with new failure rate $\gamma_j(t)$. If on interval (t_0, t) the random environment has state $i \in E$, then the residual lifetime $\tau_r - t_0$ (*up*-state) of the r-th component, $r = 1, 2, ..., n$, has a cumulative distribution function (CDF) with failure rate $\gamma_i(t)$ for time moment t, and the variables $\{\tau_r - t_0, r = 1, 2, ..., n\}$ are independent.

In the above described process elements of a sample $\{X_i, i = 1, ..., n\}$ are no i.i.d. anymore, as it is assumed in the classical sampling theory. Let $N(t)$ be a number of elements which are in the *up* state at time moment t. Expressions for $p_{r,i,j}(t_0, t) = P\{N(t) = r, J(t) = j | N(t_0) = r, J(t_0) = i\}$, for $r \in \{1, ..., n\}$, $i, j, \in E$, are used.

Our paper is devoted to a problem of statistical estimation of the described process parameters. For that we make the following suppositions. Firstly, parameters of the Markov-modulated processes $\{\lambda_{i,j}\}$ are known. Secondly, with respect to hazard rates $\gamma_i(t)$, a paramet-

ric setting takes place: all $\gamma_i(t)$ are known, accurate to m parameters $\beta^{(i)} = (\beta_{1,i}, \beta_{2,i}, ..., \beta_{m,i})^T$, so we will write $\gamma_i(t; \beta^{(i)})$. Further, we use the $(m \times k)$-matrix $\beta = (\beta^{(1)}, \beta^{(2)}, ..., \beta^{(m)})$ of unknown parameters. Thirdly, with respect to the available sample: sample elements are fixed corresponding to their appearance, so the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are fixed. Finally, states of the random environment $J(t)$ are known only for time moments $0, X_{(1)}, X_{(2)}, ..., X_{(n)}$.

The maximum likelihood estimates (Rao, 1965), (Turkington, 2002) for the unknown parameters β are derived. The elaborated technique is illustrated by a reliability estimation of the coherent series-parallel system (Gertsbakh, Shpungin, 2010), (Samaniego, 2007), which operates in an alternative random environment.

Keywords: Markov-modulated samples, maximum likelihood estimates.

References

Gertsbakh I.B., Shpungin Y. (2010): *Models of Network Reliability: Analysis, combinatorics, and Monte Carlo*, CRC press, Boca Raton-London-New York.

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

Rao C.R. (1965): *Linear Statistical Inference and its Application*, John Wiley and Sons, inc., New Your-London-Sidney.

Samaniego F. (2007): *System Signatures and their Application in Engineering Reliability*, Springer, New York-Berlin.

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

Markov-Modulated Linear Regression

Alexander M. Andronov
Transport and Telecommunication Institute
lora@mailbox.riga.lv

Nadezda Spiridovska, Irina Yatskiv
Transport and Telecommunication Institute
Spiridovska.N@tsi.lv, ivl@tsi.lv

Classical linear regression (Turkington, 2002) is of the form $Y_i = x_i\beta + Z_i,\ i = 1, \ldots, n$, where Y_i is scale response, $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})$ is $1 \times k$ vector of known regressors, β is $k \times 1$ vector of unknown coefficients, Z_i is scale disturbance. The usual assumptions take place: the disturbances Z_i are independently, identically normal distributed with mean zero and unknown variance σ^2, the $n \times k$ matrix $X = (x_{i,\nu}) = (x_i^T)^T$ has rank $r(X) = k$, so $(X^TX)^{-1} =$ exists.

Now we suppose that model (1) corresponds to one unit of a continue time, $Z_i(t)$ is Brown motion and responses $Y_i(t)$ are time-additive. Then for $t > 0$

$$Y_i(t) = x_i\beta t + Z_i\sqrt{t},\ i = 1, \ldots, n, \tag{1}$$

where disturbance Z_i are (as earlier) independently, identically normal distributed with mean zero and unknown variance σ^2.

Additionally we suppose that model (1) operates in the so-called *external environment*, which has final state space E. For the fixed state $s_j \in S$, $j = 1, \ldots, m$, parameters β of model (1) are $\beta_j = (\beta_{1,j}, \ldots, \beta_{k,j})^T$, but as earlier $\{Z_i\}$ are stochastically independent, normal distributed with mean zero and variances σ^2. Let $(t_{i,1}, \ldots, t_{i,m})$ be the $1 \times m$ vector, where component $t_{i,j}$ means a sojourn time for response Y_i in the state $s_j \in S$. If $t_i = t_{i,1} + \cdots + t_{i,m}$ then

$$Y_i(t_i) = x_i \sum_{j=1}^{m} \beta_j t_{i,j} + Z_i\sqrt{t_i},\ i = 1, \ldots, n. \tag{2}$$

Further we suppose that the external environment is random one and is described by a continuous-time Markov chain $J(t)$, $t \geq 0$, with finite state set $S = \{1, 2, \ldots, m\}$ (Pacheco, Tang, Prabhu, 2009). Let $\lambda_{i,j}$ be the known transition rate from state s_i to state s_j, and $\Lambda_i = \sum\limits_{j \neq i} \lambda_{i,j}$. Now in the formula (2), random sojourn time $T_{i,j}$ in the state s_j for the i-th realization must be used instead of $t_{i,j}$.

We have n independent realizations of this Markov chain. It is supposed that for the i-th realization the following data are available: total observation time $t_i = T_{i,j} + \ldots + T_{i,m}$, initial and final states of $J(\cdot)$, and the response $Y_i = Y_i(t_i)$ from (2). Additionally we have a knowledge on parameters $\{\lambda_{i,j}\}$ of the modulated Markov chain $J(\cdot)$. One allows us calculating average sojourn time $E\big(T_{i,j} | t_i, J_{i,0}, J_{i,\tau(i)}\big)$ in the state s_j during time t_j for the i-th realization, given fixed initial and final states $J_{i,0}$ and $J_{i,\tau(i)}$ of Markov chain $J(\cdot)$. These averages are used in the formula (2) instead of unknown values $\{t_{i,j}\}$.

On this basis, one must be estimated unknown parameters: the $k \times m$ matrix $\beta = (\beta_1 \ldots \beta_m) = (\beta_{\nu,j})$ and the variance σ^2. Corresponding statistical procedures and numerical examples are considered in the paper.

Keywords: Markov-modulated samples, maximum likelihood estimates.

References

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

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

Predictive Hierarchic Modelling of Operational Characteristics in Clinical Trials

Vladimir Anisimov
Predictive Analytics, Innovation, Quintiles, UK
Vladimir.Anisimov@quintiles.com

Statistical design and trial operational characteristics are affected by stochasticity in patient's enrolment and various event's appearance. The complexity of clinical trials and risk-based monitoring of various characteristics require developing new predictive analytic techniques.

An analytic methodology for predictive patient's enrolment modelling is developed by Anisimov & Fedorov (Statistics in Medicine, 2007). It uses delayed Poisson processes with gamma distributed rates (Poisson-gamma model) and empirical Bayesian technique. This approach accounts for stochasticity in enrolment, variation of enrolment rates between different centres and allows predicting enrolment over time and time to complete trial using either planned or interim data.

This methodology was developed further to account for random delays and closure of clinical centres and modelling events in trials with waiting time to response (oncology), Anisimov (2011ab). To model more complicated hierarchic processes including follow-up patients, different associated events including dropout and related costs, a new technique that uses evolving stochastic processes is proposed. It allows to derive closed-form solutions for many practical cases, thus, does not require Monte Carlo simulation.

Assume that patients arrive at centre i according to doubly stochastic Poisson process (Cox process) with rate $\lambda_i(t)$ that can be random and time dependent. Consider a sequence of arrival times $t_{1i} < t_{2i} < ...$ and a family of stochastic processes $\xi_{ki}(t,\theta), 0 \le t \le \tau_{ki}$, $(\xi_{ki}(t,\theta) = 0, t < 0)$, where θ is some unknown parameter, τ_{ki} is a random lifetime and $\xi_{ki}(\cdot)$ are independent at different i and k with distributions

not depending on k. Consider sums of evolving processes:

$$Z_i(t) = \sum_{k:\ t_{ki} \le t} \xi_{ki}(t - t_{ki}, \theta),\ Z(t) = \sum_i Z_i(t).$$

In this way we can describe various operational characteristics.

In particular, consider modelling follow-up and lost patients. Suppose that in centre i each patient upon arrival can either stay in the trial during follow-up period L or drop-out during this period with a given rate μ_i (possibly random). Define $\xi_{ki}(t) = 1, 0 \le t \le \tau_{ki}$, and $\xi_{ki}(t) = 0, t > \tau_{ki}$, where τ_{ik} are distributed as $\min(Ex(\mu_i), L)$ and $Ex(\mu)$ is exponentially distributed with parameter μ. Then $Z_i(t)$ represents the number of follow-up patients in centre i at time t. In a similar way we can represent lost patients and also model the related costs.

To describe the evolution of multiple events, e.g., recurrence, death and lost to follow-up in oncology trials, we can use for $\xi_{ki}(t)$ finite Markov or semi-Markov absorbing processes (Anisimov, 2011b).

For these models the main predictive characteristics are derived in a closed form. Estimation of the unknown parameter θ is also considered.

Keywords: clinical trial, patient enrolment, hierarchic modelling, analytical technique.

Acknowledgements: This work was supported by Quintiles, Innovation Unit.

References

Anisimov V.V. (2011a): Statistical modeling of clinical trials (recruitment and randomization), *Communications in Statistics - Theory and Methods*, Vol. 40, N. 19-20, pp. 3684-3699.

Anisimov V.V. (2011b): Predictive event modelling in multicentre clinical trials with waiting time to response, *Pharmaceutical Statistics*, Vol. 10, Iss. 6, pp. 517-522.

Empirical convergence bounds for Quasi-Monte Carlo integration

Anton A. Antonov
Saint Petersburg State University
tonytonov@gmail.com

Sergej M. Ermakov
Saint Petersburg State University
sergei.ermakov@gmail.com

Quasi-Monte Carlo integration techniques have recently gained significant popularity over Monte Carlo schemes, since the theoretical rate of convergence of the quasi-random approach is higher, especially in multidimensional case. However, Monte Carlo utilizes confidence intervals in order to control the integration error without any extra function evaluations, and no similar concept is available when using Quasi-Monte Carlo. The only possible general upper bound is given by the Koksma-Hlawka inequality, but the discrepancy estimation is too computationally complex to be considered as applicable, whereas introducing an artificial randomness into Quasi-Monte Carlo (i.e. random shift) requires a significant number of extra integrand evaluations.

The main idea behind confidence intervals is the fact that cumulative means of independent trials converge to the normal distribution (central limit theorem). Furthermore, this limit holds under significantly weaker conditions, including the case with dependent random variables. The latter, however, requires separate variance estimation.

In our research we split an integration domain $\mathfrak{X} \in \mathbb{R}^{\mathrm{d}}$ into $N = 2^s$ non-intersecting sub-domains $\mathfrak{X}_1, \mathfrak{X}_2, \ldots, \mathfrak{X}_N$ of equal volume and focus on a set of Haar functions, forming an orthonormal system in $L_2[\mathfrak{X}]$. By using an apparatus, described in (Ermakov, 1977), we build an interpolatory quadrature formula S_N, which is exact for the Haar system. This property allows us to use the formula in conjunction with the gen-

eralized Sobol sequence and to simultaneously build an upper boundary for the integration error. A straightforward variance analysis shows that it is possible to obtain an analogue of the confidence interval, based on a set of simultaneous estimates $\{\hat{\alpha}_i\}_{i=1}^{N}$, where $\alpha_i = \int\limits_{\mathfrak{x}_i} f(x)dx$.

The variance of the formula is then given by

$$DS_N = \frac{1}{N}\left\{\int\limits_{\mathfrak{x}} f^2(x)dx - \left(\int\limits_{\mathfrak{x}} f(x)dx\right)^2 - \sum_{i<j}(\alpha_i - \alpha_j)^2\right\}.$$

The proposed algorithm does not require extra integrand evaluations and its variance is never greater than the traditional Monte Carlo variance. Moreover, it is applicable regardless of the dimension d, provided by the proper usage of generalized Sobol sequences. Computational examples indicate that carefully chosen parameters lead to a both accurate and narrow empirical boundary for the integration error.

Keywords: Quasi-Monte Carlo, Sobol sequence, Haar functions, confidence interval

Acknowledgements: The work is supported by the RFBR grant N°11-01-00769a

References

- Ermakov S. M. (1975): *Die Monte Carlo Methode und verwandte Fragen*, VEB Deutscher Verlag der Wissenschaften, Berlin, in German.
- Sobol' I. M. (1969): *Multidimensional Quadrature Formulas and Haar Functions*, Nauka, Moscow, in Russian.

The influence of the dependency structure in combination-based permutation tests

Rosa Arboretti*, Iulia Cichi*, Luigi Salmaso‡, Vasco Boatto*, Luigino Barisan*
*Department of Land, Environment, Agriculture, University of Padova, Italy and ‡Forestry and Department of Management and Engineering, University of Padova, Italy
rosa.arboretti@unipd.it, iulia.cichi@unipd.it, luigi.salmaso@unipd.it vasco.boatto@unipd.it, luigino.barisan@unipd.it

With reference to multivariate permutation tests we deal with the method of NonParametric Combination (NPC) of a finite number of dependent permutation tests. A quite important problem usually occurs in several multidimensional applications when variables, which are to be analyzed, are correlated and their associated regression forms are different (linear, quadratic, exponential, general monotonic, etc.).In this paper we show that the nonparametric combining strategy is suitable to cover almost all real situations of practical interest since the dependence relations among partial tests are implicitly captured by the combining procedure itself. One open problem related to NPC-based tests is the possibility for the experimenter to manage with the impact of the dependency structure on the possible significance of combined tests. We will explore this problem from different points of view (different kinds of dependency, different combination functions and an increasing number of correlated variables). We also present an application example on a real case study from wine studies.

Keywords: correlation, permutation tests, nonparametric combination, NPC.

References

Agresti A. (2002): *Categorical Data Analysis*, Wiley, New Jersey.

Pesarin F., Salmaso L. (2010): *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.

Tempesta T., Arboretti Giancristofaro R., Corain L., Salmaso L., Tomasi D., Boatto V., (2010), The Importance of Landscape in Wine Quality Perception: an Integrated Approach Using Choice-Based Conjoint Analysis and Combination-Based Permutation Tests, *Food Quality and Preference*, **21**, 827-836.

Effects of correlation structures on nonparametric rankings

Rosa Arboretti
Dept. Land, Environment, Agriculture and Forestry, University of Padova, Italy
rosa.arboretti@unipd.it

Kelcey Jasen
Department of Statistics, University of Central Florida, Florida
Kelcey.Jasen@ucf.edu

Mario Bolzan
Department of Statistics, University of Padova, Italy
mario.bolzan@unipd.it

Evaluating customer satisfaction is an important process in continuous improvement of businesses and evaluating the quality of products or services. The methodology proposed uses customer satisfaction survey results to obtain a final ranking of the individual qualities and analyzes the subsets of relationships between the qualities. An evaluation collected in the form of rankings, or likert scales, rarely naturally satisfies parametric assumptions and thus a nonparametric approach is desirable. The pooling of preference ratings using the nonparametric combination ranking methodology has an underlying dependency structure which is nonparametric. Permutation tests are performed on the correlations between all possible subsets of rankings. An analysis of the correlations between rankings leads to an understanding of which qualities are dominating the global ranking and which hold lesser importance.

Keywords: global ranking, permutation tests, nonparametric combination.

References

Arboretti Giancristofaro R., Corain L., Gomiero D., Mattiello F. (2010): Nonparametric Multivariate Ranking Methods for Global Performance Indexes, *Quaderni di Statistica*, Vol. 12, pp. 79-106.

Arboretti G. R., Bonnini S., Salmaso L., (2009). Employment status and education/employment relationship of PhD graduates from the University of Ferrara. Journal of Applied Statistics, 36, 12, pp. 1329-1344.

Pesarin F., Salmaso L. (2012): *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.

Queuing modeling and simulation analysis of bed occupancy control problems in healthcare

Cristina Azcarate, Fermin Mallor
Dept. Statistics and OR, Public University of Navarre, Spain
cazcarate@unavarra.es, mallor@unavarra.es

Julio Barado
Hospital of Navarre, Pamplona, Spain
SX017905@navarra.es

The type of equipment and clinical staff required by intensive care units (ICU) makes them very costly to run. The inevitable periodic bed shortages have no desirable consequences: admissions and discharges of patients are triaged, and then the number of patients who are rejected from admission increases and the length of stay (LoS) gets shortened. A high quality of service means a low percentage of rejected patients and a LoS in the ICU as long as necessary. To reach high levels for the two mentioned objectives efficient bed management policies are necessary. Several papers have used simulation techniques to find a solution to the ICU bed management problem. Although some of these studies suggest early discharge as a bed management tool, they do not include it in their models. Mallor and Azcarate (2013) demonstrated in a real setting that patient LoS is not independent of the ICU workload but it can be influenced by the ICU bed occupancy level. As a consequence they pointed out the need to include these discharge policies to get a valid simulation model.

In this paper we consider the problem of obtaining efficient bed-management policies for the ICU. To achieve this objective we consider the mathematical representation of an ICU as a $G/G/c/c$ queue model with several types of customers, each one with its own arrival pattern and service time. We first study a simplified version of the ICU (by adopting Markovian assumptions on arrivals and service times) that will

allow us to obtain properties related with the management and LoS of patients.

Queueing theory has studied the problem of resource allocation under uncertainty (Gross and Harris, 2008) to provide queue designs and control policies that optimize some measure of interest. Here, we deal with a different queueing control problem in which neither the arrival rates nor the number of servers can be modified. The control of the bed occupancy is addressed by modifying the service time rates, making them dependent on the system state: individual service time μ_i, when i beds are occupied, $i = 1, ..., c$. The goals to be achieved are the two quality of service (QoS) components already mentioned: to minimize the probability of rejecting a patient (because a full ICU) and to minimize the shortening of the patient's LoS. The first objective has a clear mathematical formulation: $Minimize\ p_c$. For the second objective we propose four different formulations, all of them leading to nonlinear optimization problems. We obtain the solutions to these problems and compare them in terms of the two dimensions of the QoS.

Keywords: simulation, queuing models, intensive care unit, bed occupancy.

Acknowledgements: This work was partially supported by the Spanish Ministry of Economy and Competitiveness, project MTM2012-36025.

References

Gross, D., Harris, C.M. (2008): *Fundamentals of queueing theory*, John Wiley and Sons, 4th Ed.

Mallor, F., Azcarate, C. (2013): Combining optimization with simulation to obtain credible models for intensive care units, *Annals of Operations Research*, DOI 10.1007/s10479-011-1035-8.

Left Truncated and Right Censored Lifetime Data: Simulation and Analysis

Narayanaswamy Balakrishnan
McMaster University, Hamilton, Ontario, Canada
bala@univmail.cis.mcmaster.ca

Left truncated and right censored data arise naturally in many lifetime studies. After providing a real motivating example from a reliability study of power transformers, I shall describe the basic model, form of data and some results on likelihood method of inference, and specifically the details of an efficient EM algorithm. Then, I shall describe a simulation algorithm and present results evaluating the performance of the method of estimation as well as likelihood-based model discrimination from an extensive Monte Carlo simulation study. Finally, I will conclude with an illustrative example and some suggestions for further work.

Spline based ROC curves and surfaces for biomarkers with an upper or a lower limit of detection

Leonidas E. Bantis
University of the Aegean, Dept. of Statistics and Actuarial-Financial Mathematics
lbantis@aegean.gr

John V. Tsimikas, Stelios D. Georgiou
University of the Aegean, Dept. of Statistics and Actuarial-Financial Mathematics
tsimikas@aegean.gr, stgeorgiou@aegean.gr

Receiver Operating Characteristic (ROC) curves are a common tool for assessing the accuracy of an ordinal or continuous biomarker when two groups are to be distinguished (usually the healthy and the diseased group). However, it might be the case that biomarker measurements are censored due to a lower (and more rarely an upper) limit of detection (LOD). This is usually due to practical limitations regarding the nature/mechanism of the biomarker. We study a spline based approach for ROC estimation for which monotonicity constraints are imposed. The monotone spline is fitted to the cumulative hazard function of the marker measurements. Under the proposed technique the problem reduces to a restricted least squares one, with linear restrictions on the parameters. Hence, convex optimization is involved and convergence is guaranteed. The presence of covariates might affect the discriminatory capability of a biomarker. Our approach can accommodate observed covariates with the use of the well known Cox model typically used in survival analysis. The extension of this modeling approach for the estimation of an ROC surface that refers to the discriminatory capability of a biomarker in distinguishing between three groups is straightforward. The advantages of the presented approach is that it avoids strict para-

metric assumptions unlike usual maximum likelihood techniques, it is computationally stable, and it can be used for measurements that lie on the real line, unlike simple imputation techniques that assume positive measurements and induce bias in estimating the volume under the ROC surface. The method is evaluated and compared to other known methods via simulations. Estimation of the area under the curve (AUC) which is the most commonly used index that summarizes the overall marker's diagnostic ability, is shown to be more efficient with the proposed approach compared to the other simple imputation based ones. Furthermore, our approach can be particularly useful when interest lies in high FPR ranges and the partial AUC is to be estimated. Similar simulation results are obtained for the volume under the ROC surface (VUS) when examining the three class case.

Keywords: area under the curve, limit of detection, Receiver Operating Characteristic, spline.

Simulating correlated ordinal and discrete variables with assigned marginal distributions

Alessandro Barbiero
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano
alessandro.barbiero@unimi.it

Pier Alda Ferrari
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano
pieralda.ferrari@unimi.it

Stochastic simulation is a significant aspect of statistical research. Model building, parameter estimation, hypothesis tests, and other statistical tools require verification to assess their validity and reliability, typically via simulated data. In many research fields, data sets often include ordinal variables, e.g. measured on a Likert scale, or count variables. This work aims to give a contribution on these topics by proposing a procedure for simulating samples from ordinal and discrete variables with assigned marginal distributions and association structure. Up to now, a few methodologies that address this problem have appeared in the literature. Ruscio and Kaczetow (2008) introduced an iterative algorithm for simulating multivariate non-normal data (discrete or continous). They first construct a huge artificial population whose components are independent samples from the desired marginal distributions and then reorder them in order to catch the target correlations; the desired samples are drawn from this final population as simple random samples. Demirtas (2006) proposed a method for generating ordinal data by simulating correlated binary data and transforming them into ordinal data, but the procedure is complex and computationally expensive, since it requires the iterative generation of large samples of binary data.

More recently, Ferrari and Barbiero (2012) proposed a method (called GenOrd) able to generate correlated point-scale rv (i.e. rv whose support is of the type $1, 2, \ldots, k$) with marginal distributions and Pearson's correlations assigned by the user. A sample is drawn from a standard multivariate normal rv with correlation matrix $\mathbf{R}^N$ and then discretized to yield a sample of discrete/ordinal data meeting the prescribed marginal distributions. The matrix $\mathbf{R}^N$ ensuring the assigned correlation matrix $\mathbf{R}^D$ on the target variables is computed through a recursive algorithm.

In this work, we show that this method is also able to generate discrete variables with any finite support and/or association structure expressed in terms of Spearman's correlations as well. We also propose a modification of GenOrd for dealing with discrete variables defined on infinite support, which needs to be truncated when computing $\mathbf{R}^N$, since the method requires a finite number of cut-points when discretizing the multivariate normal rv. The performances of the techniques above described are compared through a Monte Carlo study and assessed in terms of computational efficiency and precision under various settings. Examples of application of the new proposal concerning inferential issues are provided to show its utility and usability even by non-experts.

Keywords: multivariate random variable, Pearson's correlation, Spearman's correlation.

References

Demirtas H. (2006): A method for multivariate ordinal data generation given marginal distributions and correlations, *Journal of Statistical Computation and Simulation*, 76, pp.1017-1025.

Ferrari P.A., Barbiero A. (2012): Simulating Ordinal Data, *Multivariate Behavioral Research*, 47:4, pp.566-589.

Ruscio J., Kaczetow W. (2008): Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, 43:3, pp.355-381.

Reliability Modeling with Hidden Markov and semi-Markov Chains

Vlad Stefan Barbu
Laboratoire de Mathématiques Raphael Salem, Université de Rouen, France
barbu@univ-rouen.fr

In this talk, we are interested in hidden models of Markovian and semi-Markovian type, in associated reliability and survival analysis topics and in some related estimation procedures.

First, we will present a canonical system for which hidden Markov or semi-Markov processes are appropriate modeling tools and we introduce the corresponding notation and definitions.

Second, we are interested in presenting reliability problems for which the hidden (semi-)Markov processes are adapted modeling tools. Examples are mainly related to software reliability modeling (cf., e.g, Ledoux, 2003; Durand and Gaudoin, 2005; Gaudoin and Ledoux, 2007). The problem of obtaining the reliability indicators in terms of the basic characteristics of the models is addressed.

Third, we are interested in the estimation of such models, which is usually obtained through an algorithmic approach. In our case, the corresponding estimators are obtained through an EM algorithm, that we briefly describe.

The interest of the type of stochastic processes that we present in our talk comes: on the one hand, from the wide range of applications for which these processes are a flexible modeling tools; on the other hand, from the important generalization that the hidden semi-Markov processes bring as compared to the hidden Markov processes, that are too restrictive for a certain number of applications (see, e.g., Barbu and Limnios, 2008).

Keywords: hidden Markov and semi-Markov chains, survival analysis, relia-

bility theory, statistical estimation.

References

Barbu V.S., Limnios N. (2008): *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications - Their use in Reliability and DNA Analysis*, Lecture Notes in Statistics, Vol. 191, Springer, New York.

Durand J.-B., Gaudoin O. (2005): Software reliability modelling and prediction with hidden Markov chains, *Statistical Modelling*, Vol. 5, N. 4, pp. 75–93.

Gaudoin O., Ledoux J. (2007): *Modélisation aléatoire en fiabilité des logiciels*, Hermès, Paris (in French).

Ledoux J. (2003): Software reliability modeling. In: Pham H. (Eds.) *Handbook of Reliability Engineering*, Springer, New York, pp. 213-234.

Adaptive quadrature for likelihood inference in dynamic latent variable models

Francesco Bartolucci
Department of Economics, Finance and Statistics, University of Perugia (IT)
bart@stat.unipg.it

Silvia Cagnone
Department of Statistical Sciences, University of Bologna (IT)
silvia.cagnone@unibo.it

Dynamic latent variable models represent a useful and flexible tool in the study of macro and micro-econometric data. Here we refer to two particular dynamic latent variable models, the Stochastic Volatility (SV) models applied in the analysis of financial time series and the Limited Dependent Variable (LDV) models for analyzing panel data. Both models allow us to well capture the variability present in the data through an autoregressive structure and at the same time they are more parsimonious than other models. Nevertheless, the estimation procedure of these models presents some computational difficulties related to the presence of the latent variables. Since these variables are unobservable, they have to be integrated out from the likelihood function and an analytical solution does not exist. Among the different estimation procedures discussed in the literature, a common method for both models is based on the direct maximum likelihood estimation using a non-linear filter algorithm. This algorithm allows to rephrase the likelihood function as a product of univariate integrals that have to be approximated. In this regard, for the SV model Fridman and Harris (1989) proposed to use the Gauss Legendre numerical quadrature, whereas Bartolucci and De Luca (2003) applied a rectangular quadrature. As for the LDV models Heiss (2008) approximated the uni-dimensional integrals by means of the Gauss Hermite (GH) quadrature.

In this work we propose to use the Adaptive Gaussian Hermite (AGH)

numerical quadrature (Liu and Pierce, 1994) to approximate the unidimensional integrals resulting from the non-linear filtering algorithm applied in the estimation of both the SV and LVD models. This numerical method appears to be superior to the other approximations mainly for two different reasons. It requires only few quadrature points to get accurate estimates and it does not risk to miss the maximum since it well captures the peak of the integrand involved in the likelihood function. A wide simulation study is carried out in order to evaluate the performance of AGH under different conditions of study and we compare its performance with the other approximation methods.

Keywords: Gauss Hermite quadrature, autoregressive structure, non linear filter algorithm.

References

Bartolucci F., De Luca G. (2003): Likelihood-based inference for asymmetric stochastic volatility models, *Computational Statistical and Data Analysis*, Vol. 42, pp. 445-449.

Fridman M., and Harris L. (1989): A Maximum likelihood approach for non-Gaussian stochastic volatility models, *Journal of Business and Economic Statistics*, Vol 16, pp. 284-291.

Heiss F. (2008): Sequential numerical Integration in nonlinear state space models for microeconometric panel, *Journal of Applied Econometrics*, Vol.23, pp. 373-389.

Liu, Q. and Pierce, D.A. (1994): A note on Gauss-Hermite quadrature, *Biometrika* , Vol.81, pp. 624-629

Bayes-optimal importance sampling for computer experiments

Julien Bect and Emmanuel Vazquez
Suplec, Gif-sur-Yvette, France
{julien.bect, emmanuel.vazquez}@supelec.fr

Let $h : \mathfrak{X} \to \mathbb{R}$ denote a function that is expensive to evaluate, for instance the response, or a function of the response, of a time-consuming deterministic computer program with input space $\mathfrak{X}$. Given a probability density function π with respect to a reference measure μ on $\mathfrak{X}$, we consider the problem of constructing a good unbiased estimator of the average $\theta = \int h\,\pi\,\mathrm{d}\mu$ using importance sampling (IS):

$$\widehat{\theta}_m \;=\; \frac{1}{m}\sum_{i=1}^{m} h(X_i)\,w(X_i)\,, \qquad X_i \overset{\text{iid}}{\sim} q(x), \qquad w = \frac{\pi}{q}. \tag{1}$$

Following a path that is now classical in the design and analysis of computer experiments, it has recently been proposed—in the special case where h is the indicator function of a rare event—to make use of a Gaussian process model of the expensive computer code, in order to select a good proposal density q (Dubourg et al., 2011; Barbillon et al., 2012).

As a first contribution to this line of research, we revisit the classical question of choosing an optimal proposal density for (1), from the point of view of Bayesian numerical analysis (Ritter, 2000). We prove that the optimal proposal distribution for the quadratic risk is proportial to $g(x)\,\pi(x)$, where $g(x)^2 = \mathbb{E}\left(h(x)^2\right)$, the expectation being taken with respect to the prior on h. The proposal distribution used by Dubourg et al. (2011) and Barbillon et al. (2012) were, thus, sub-optimal.

Unfortunately, it turns out that the estimator (1), with q the optimal proposal just determined, cannot be used directly since 1) the optimal density is only known up to a multiplicative constant, and 2) we do not know in general how to draw independent samples from q. As a second

contribution, we explain how to construct an approximate IS estimator using Sequential Monte Carlo simulations (Del Moral et al., 2006). In the case where h is the indicator function of a static rare event, the proposed approach is reminiscent of the Bayesian Subset Simulation algorithm of Li et al. (2012). We prove that the approximate IS estimator is still unbiased, and converges in distribution to $\widehat{\theta}_m$ when the number of particles goes to infinity (under the condition that h and w are continuous $q\mu$-almost everywhere).

Keywords: Importance Sampling, Gaussian process, Bayesian Numerical Analysis, Rare events, Subset Simulation.

References

P. Barbillon, Y. Auffray, and J.-M. Marin (2012): Bornes de probabilités d'événements rares dans le contexte des événements simulés. In *44èmes Journées de Statisque (JdS 2012), 21-25 mai 2012, Bruxelles, Belgique*.

P. Del Moral, A. Doucet, and A. Jasra (2006): Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, Vol. 68, N. 3, pp. 411–436.

V. Dubourg, F. Deheeger, and B. Sudret (2011): Metamodel-based importance sampling for structural reliability analysis. Preprint arXiv:1105.0562v2.

L. Li, J. Bect, and E. Vazquez (2012): Bayesian subset simulation: a kriging-based subset simulation algorithm for the estimation of small probabilities of failure. In *Proceedings of PSAM 11 & ESREL 2012, 25-29 June 2012, Helsinki, Finland*.

K. Ritter (2000): *Average-case Analysis of Numerical Problems*, Lecture Notes in Mathematics, Vol. 1733, Springer Verlag.

Probabilistic counterparts of nonlinear parabolic systems

Belopolskaya Yana
St.Petersburg State University for Architecture
and Civil Engineering
yana.belopolskaya@gmail.com

The main object of our investigation here is a probabilistic representation of a solution to the Cauchy problem for a class of nonlinear parabolic systems including nonlinear systems of two types and their combination. Namely, given functions $a \in R^d$, $A \in R^{d\times d}$, $c \in R^{d_1\times d_1}$, $C \in R^{d\times d_1\times d_1}$ depending on $x, u, \nabla u$ we consider the Cauchy problem

$$\frac{\partial u_m}{\partial s} + \frac{1}{2} TrA^*\nabla^2 u_m A + \langle a, \nabla u_m\rangle + \sum_{l=1}^{d_1}[C_{ml}A^*\nabla u_l \qquad (1)$$

$$+c_{ml}u_l] + g_m(x,u,\nabla u) = 0, \quad u_m(T,x) = 0, \quad \text{and}$$

$$\frac{\partial u^q}{\partial s} + \frac{1}{2}\mathrm{Tr}A^{q*}\nabla^2 u^q A^q + \langle a^q, \nabla u^q\rangle + \sum_{m=1}^{k}\gamma^{qm}u^m + \qquad (2)$$

$$g^q(x,u,\nabla u^q) = 0, \quad u^q(T,x) = u_0^q(x),$$

with a, A defined on $R^d \times V \times R \times R^d$, $V = \{1,\dots,d_1\}$ and $\gamma \in R^{k\times k}$ defined on R^d. Here d, d_1, k are given positive integers. To obtain a probabilistic counterpart of (1) given a standard Wiener process $w(t) \in R^d$, we consider a system of stochastic equations

$$d\xi(t) = A(\xi(t), u(t,\xi(t))dw(t), \quad \xi(s) = x \in R^d, \qquad (3)$$

$$d\eta(t) = c(\xi(t), u(t,\xi(t)))\eta(t)dt + C(\xi(t), u(t,\xi(t)))(\eta(t), dw(t)), \qquad (4)$$

$$\langle h, u(s,x)\rangle = E_{s,x,h}\left[\langle \eta(T), u_0(\xi(T))\rangle\right] + \qquad (5)$$

$$E_{s,x,h}\left[\int_s^T \langle \eta(\theta), g(\xi(\theta), u(\theta,\xi(\theta)))\rangle d\theta\right], \text{ where } \eta(s) = h \in R^{d_1}.$$

Under certain conditions (5) defines a unique classical (local in time) solution of a semilinear form of (1). By a differential continuation this approach can be extended to (1) and to some systems of fully nonlinear PDEs. To get a probabilistic representation of a solution for (2) given a Markov chain $\nu(t)$ with generator $[\Gamma f]_q = \sum_{q'=1}^{d_1} \gamma_{qq'}[f_{q'} - f_q]$) we set $u^\nu(s,x) \equiv u(s,x,\nu)$ and consider

$$d\xi(t) = a^{\nu(t)}(\xi(t), u^{\nu(t)}(t,\xi(t)))dt + A^{\nu(t)}(\xi(t), u^{\nu(t)}(t,\xi(t)))dw(t),$$
$$u^q(s,x) = E_{s,x,q}\left[u_0^{\nu(T)}(\xi(T)) + \int_s^T \tilde{g}^{\nu(t)}(\xi(t), u^{\nu(t)}(t,\xi(t))dt)\right], \tag{6}$$

where $\tilde{g}^q = \exp\{\int_s^t \gamma_{qq}(\theta)d\theta\}g^q$. Finally, we combine the above approaches to derive probabilistic representations for more complicate systems of parabolic equations. Probabilistic representations (5) and (6) can be used to construct numerical solutions of the Cauchy problem for (1) and (2). To obtain viscosity solutions of (1) or (2) we apply an approach based on the Pardoux-Peng theory of backward stochastic differential equations and the above results (see Belopolskaya (2011)).

Keywords: Markov chains, processes, nonlinear parabolic systems

Acknowledgements: The support of RFBR Grant No. 12-01-00457 and project 1.370.2011 Minobrnauki is gratefully acknowledged

References

Belopolskaya Ya. (2011): Probabilistic approaches to nonlinear parabolic equations in jet-bundles, *Global and Stochastic Analysis,* Vol. 1, N. 1, pp. 3-40.

Strengths and weaknesses with non-linear mixed effect modelling approaches for making inference in drug development

Martin Bergstrand
Department of Pharmaceutical Biosciences, Uppsala University, Sweden
martin.bergstrand@farmbio.uu.se

Mats O Karlsson
Department of Pharmaceutical Biosciences, Uppsala University, Sweden
mats.karlsson@farmbio.uu.se

Application of pharmacometric, non-linear mixed effect (NLME), models to analyze longitudinal data from clinical trials have advantages both with respect to the type of information gained and the statistical power for making inference (1,2) could make drug development more efficient. It could be of particular value in small population groups where practical limitations severely hamper the possibility to sufficiently power a study based on a more conventional approach (3). The possibility to combine different sources of information from multiple studies with varying design is another advantage with this approach that is likely to be of particular importance to small populations.

The advantages with a pharmacometric approach for planning and analyzing clinical trials are illustrated with two case studies from the therapeutic areas diabetes and rheumatoid arthritis (1,5,6). Likelihood ratio tests based on NLME models applied to longitudinal data, of one or more connected outcome variables, were compared to t-tests for group-wise comparison of change from baseline data. In the case studies, the NLME based analysis resulted in several-fold reductions of expected sample sizes for a given power to detect clinically relevant drug effects. The case studies were also used to illustrate how the expected utility of different study designs and evaluation options, such as incorporation pharmacokinetic measurements, can be assessed.

The primary opposition towards application of a NLME modelling approach typically lies in the difficulty to control the assumptions made. Strategies to assess the validity of model assumptions and/or relax such assumptions, by accounting for model and parameter uncertainty, will be discussed in the context of the case studies.

Keywords: Pharmacometrics, Small populations, NLME, Power.

References

Karlsson K.E., Vong C., Bergstrand M., Jonsson E.N., Karlsson M.O. (2013): Comparisons of Analysis Methods for Proof-of-Concept Trials, *CPT: Pharmacometrics and Systems Pharmacology*, Vol. 2, e23.

Jonsson, E.N., Sheiner L.B. (2002): More efficient clinical trials through use of scientific model-based statistical tests, *Clinical pharmacology and therapeutics*, Vol. 72, N 6, pp. 603-14.

Lesko L.J. (2012): Quantitative analysis to guide orphan drug development, *Clinical Pharmacology and Therapeutics*, Vol. 92, N 2, pp. 258-61.

Hamren B., Bjork E., Sunzel M., Karlsson M. (2008): Models for plasma glucose, HbA1c, and hemoglobin interrelationships in patients with type 2 diabetes following tesaglitazar treatment, *Clinical Pharmacology and Therapeutics*, Vol. 84, N. 2, pp.228-35.

Lacroix B.D., Lovern M.R., Stockis A., Sargentini-Maier M.L., Karlsson M.O., Friberg L.E. (2009): A pharmacodynamic Markov mixed-effects model for determining the effect of exposure to certolizumab pegol on the ACR20 score in patients with rheumatoid arthritis, *Clinical Pharmacology and Therapeutics*, Vol. 86, N.4, pp. 387-95.

Discrete Simulation of Stochastic Delay Differential Equations

Harish S. Bhat, Nitesh Kumar, and R. W. M. A. Madushani
Applied Mathematics Unit
University of California, Merced
5200 N. Lake Rd., Merced, CA 95343 USA
hbhat@ucmerced.edu

The stochastic delay differential equation

$$dX_t = \alpha X_{t-\tau} dt + dW_t, \tag{1}$$

where W_t is the standard Wiener process, has been proposed as a simplified, macroscopic model for human balance control (Milton, 2011). We seek a numerical method to compute the probability density function of X_t that does not require computing sample paths of (1). For a stochastic differential equation with no delay, this task could be completed by solving the associated Fokker-Planck equation; however, the Fokker-Planck equation associated with stochastic delay equations is circular and has thus far been of limited use in numerical solution procedures.

Our strategy for solving (1) consists of discretizing the equation both in time and in probability space, i.e., converting all random variables in the problem from continuous to discrete. This yields a delayed random walk, the probability mass function of which can be computed using two methods that we develop: a recursive method and a tree method. The recursive method involves unraveling the time-discretization of (1) into a sum of random variables, and then computing the distribution of this sum. The tree method consists of incrementally growing a tree of all possible sample paths of (1) together with the respective probabilities along each path. Rather than grow full trees (Bhat and Kumar, 2012), we grow approximate trees in which paths that differ by a small tolerance are allowed to coalesce. Both methods we propose can be used to solve (1) even if W_t is replaced by a different stochastic process. We analyze

the accuracy and efficiency of these two competing numerical methods, and we judge which method can be more readily generalized to solve a more complex stochastic delay model that incorporates feedback control (Milton et al., 2009). Due to its discontinuous nature, this latter model cannot be treated using standard numerical methods for stochastic delay differential equations (Mao, 2003).

Keywords: stochastic differential equations, time delay, biophysical modeling, numerical analysis, delayed random walks

Acknowledgements: This work was supported by a grant from the Graduate Research Council of the University of California, Merced.

References

Bhat H.S. and Kumar N. (2012): Spectral Solution of Delayed Random Walks, *Physical Review E*, Vol. 86, No. 4, 045701.

Mao X. (2003): Numerical Solutions of Stochastic Functional Differential Equations, *LMS Journal of Computation and Mathematics*, Vol. 6, pp. 141-161.

Milton J.G. (2011): The Delayed and Noisy Nervous System: Implications for Neural Control, *Journal of Neural Engineering*, Vol. 8, No. 6, 065005.

Milton J.G., Townsend J.L., King M.A., Ohira T. (2009): Balancing with Positive Feedback: The Case for Discontinuous Control, *Philosophical Transactions of the Royal Society A*, Vol. 367, No. 1891, pp. 1181-1193.

An Affine Invariant k-Nearest Neighbor

Gérard Biau
Université Pierre et Marie Curie, Paris, France
gerard.biau@upmc.fr

Luc Devroye
McGill University, Montreal, Canada
lucdevroye@gmail.com

Vida Dujmović
Carleton University, Ottawa, Canada
gerard.biau@upmc.fr

Adam Krzyżak
Concordia University, Montreal, Canada
gerard.biau@upmc.fr

We design a data-dependent metric in $\mathbb{R}^d$ and use it to define the k-nearest neighbors of a given point. Our metric is invariant under all affine transformations. We show that, with this metric, the standard k-nearest neighbor regression estimate is asymptotically consistent under the usual conditions on k, and minimal requirements on the input data.

Keywords: Nonparametric estimation, Regression function estimation, Affine invariance, Nearest neighbor methods, Mathematical statistics.

Acknowledgements: This work was supported by NSERC grants.

References

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

M. Hallin and D. Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30:1103–1133, 2002.

T.P. Hettmansperger, J. Möttönen, and H. Oja. The geometry of the affine invariant multivariate sign and rank methods. *Journal of Nonparametric Statistics*, 11:271–285, 1998.

H. Oja and D. Paindaveine. Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323, 2005.

E. Ollila, H. Oja, and V. Koivunen. Estimates of regression coefficients based on lift rank covariance matrix. *Journal of the American Statistical Association*, 98:90–98, 2003.

R.H. Randles. A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84:1045–1050, 1989.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

Models with cross-effect of survival functions in the analysis of patients with multiple myeloma

Alexander Bitukov, Oleg Rukavitsyn
The Hematology Center, Main Military Clinical Hospital name of N.N.Burdenko, Moscow, Russia
82465@bk.ru

Ekaterina Chimitova, Boris Lemeshko, Mariya Vedernikova
Novosibirsk State Technical University, Novosibirsk, Russia
ekaterina.chimitova@gmail.com

Mikhail Nikulin
Universite Victor Segalen, Bordeaux, France
mikhail.nikouline@u-bordeaux2.fr

Accelerated life models are used more and more often in oncology and hematology studies for the problems of relating lifetime distribution to explanatory variables; see Klein and Moeschberger (1997), Piantadosi (1997) and Zeng and Lin (2007). The survival functions for different values of the covariates according to the Cox proportional hazard (PH) model do not intersect. However, in practice this condition often does not hold. Then we need to apply some more complicated models which allow decreasing, increasing or nonmonotonic behavior of the ratio of hazard rate functions. Following Bagdonavicius, Levuliene and Nikulin (2009) and Nikulin and Wu (2006) we give examples to illustrate and compare possible applications of the Hsieh model (see Hsieh (2001)) and the simple cross effect (SCE) model (see Bagdonavicius and Nikulin (2002)), both of them are particularly useful for the analysis of survival data with one crossing point.

The research of various schemes of chemotherapy for the patients with multiple myeloma has been carried out. The purpose of the investigation is to compare the response time to the treatment in two groups of patients who received different treatment. As the Kaplan-Meier estimates of distribution functions in two groups intersect, the Cox PH

model can be inappropriate for these data. For this reason we propose using the models with cross-effect for relating the distribution of response time to the scheme of chemotherapy, type of the response, etc.

A very important practical result of our analysis is the establishment of the influence of Bortezomibe on the speed of the achievement of the response. We have ascertained the fact that responses such as a complete response, partial response, minimal response, stabilization and progression of the disease in the group of patients treated by Bortezomibe were achieved faster than in the control group.

Keywords: lifetime analysis, censored data, regression models, cross-effect.

References

Bagdonavicius, V. and Nikulin, M. (2002). *Accelerated Life Models*. Boca Raton: Chapman and Hall/CRC.

Bagdonavicus, V., Levuliene, R. and Nikulin, M. (2009). Testing absence of hazard rates crossings, *Comptes Rendus de l'Academie des Sciences de Paris*, Ser. I, 346, 7-8, 445-450.

Hsieh, F. (2001). On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society*, B 63, 63-79.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis*, New York: Springer.

Nikulin, M. and Wu, H.-D. (2006). Flexible regression models for carcinogenesis data. *Probability and Statistics, Steklov Mathematical Institute in St.Petersburg, RAS*, 78-101.

Piantadosi, S. (1997). *Clinical Trials*, J.Wiley : New York.

Zeng, D. and Lin, D.Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of Royal Statistical Society*, B 69, 1-30.

Two notions of consistency useful in permutation testing

Stefano Bonnini
Department of Economics and Management, University of Ferrara, Italy
stefano.bonnini@unife.it

Fortunato Pesarin
Department of Statistics, University of Padova, Italy
pesarin@stat.unipd.it

Consistency of some nonparametric tests has been studied by several authors under the assumption that the population variance is finite and/or in the presence of some violations of the data exchangeability between samples. Since main inferential conclusions of permutation tests concern the actual data set, we consider the notion of consistency in a weak version (i.e. in probability), i.e. when data behave in accordance with the alternative, as the standardized noncentrality of a test statistic diverges the rejection probability tends to one for all $\alpha > 0$. Here, we characterize weak consistency of permutation tests within two quite different but complementary settings. According to the traditional notion, the first assumes population mean is finite and without assuming existence of population variance considers the rejecting behavior for large sample sizes. The second assumes sample sizes are fixed (and possibly small) and considers on each subject a large number of informative variables. Moreover, since permutation test statistics do not require to be standardized, we need not assuming that data are homoscedastic in the alternative. Some application examples to mostly used test statistics are discussed. A simulation study and some hints for robust testing procedures are also presented.

As a guide and without loss of generality, we refer to univariate one-sided two independent sample designs with real or ordered categorical variables. Extensions to nominal categorical variables, one-sample,

multi-sample, and multivariate designs are straightforward, the latter being obtained by the nonparametric combination (NPC) of dependent permutation tests (Pesarin, 2001; Pesarin and Salmaso, 2010).

Keywords: heteroscedastic alternatives, nonparametric combination, random effects, weak finite-sample consistency, weak traditional consistency.

References

Finos L., Salmaso L. (2007): FDR- and FWE-controlling methods using data-driven weights, *Journal of Statistical Planning and Inference*, Vol.137, pp. 3859-3870.

Lehmann E.L. (1951): Consistency and unbiasedness of certain nonparametric tests, *Annals of Mathematical Statistics*, Vol.22, pp. 165-179.

Lehmann E.L. (1986): *Testing Statistical Hypotheses*, Wiley(2nd ed.), New York.

Pesarin F. (2001): *Multivariate Permutation Tests With Applications in Biostatistics*, Wiley, Chichester.

Pesarin F., Salmaso L. (2009): Finite-sample consistency of combination -based permutation tests with application to repeated measures designs, *Journal of Nonparametric Statistics*.

Pesarin F., Salmaso L. (2010): *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley, Chichester.

Pesarin F., Salmaso L. (2011): A new characterization of weak consistency of permutation tests, *Journal of Statistical Planning and Inference*, (forthcoming).

Attribute Agreement Analysis in a Forensic Handwriting Study

Michele Boulanger
Dept. of International Business, Rollins College
mboulanger@rollins.com

Mark E. Johnson
Dept. of Statistics, University of Central Florida
mejohnso@mail.ucf.edu

Thomas W. Vastrick
Forensic Document Examiner
vastrick@yahoo.com

A study is described that involves enhancing the scientific underpinnings of forensic handwriting analysis. Forensic document examiners are individuals who carefully scrutinize handwriting samples in a variety of civil and criminal cases and attest to the originators of hand written documents (wills, ransom notes, letters found at crime scenes, and so forth). A large-scale study is described that addresses the concerns of the US National Academy of Sciences regarding handwriting analyses. Data from multiple written specimens that have been reviewed by multiple examiners is described and the preliminary results are given. Challenges in the analyses and future directions are given.

Keywords: forensics, handwriting, attribute agreement analysis.

Acknowledgements: This work was supported by a National Institute of Justice grant through the National Center for Forensic Science at the University of Central Florida.

References

Strengthening Forensic Science in the United States: A Path Forward, Committee on Identifying the Needs of the Forensic Sciences Community; Committee on Applied and Theoretical Statistics, National Research Council, 2009, 254 pages.

Small variance estimators for rare event probabilities

Broniatowski Michel
LSTA, University Paris 6
michel.broniatowski@upmc.fr

Caron Virgile
LSTA, University Paris 6
virgile.caron@upmc.fr

Improving Importance Sampling estimators for rare event probabilities requires sharp approximations of conditional densities. This is achieved for events defined through large exceedances of the empirical mean of summands of a random walk, in the domain of large or moderate deviations. The approximation of conditional density of the trajectory of the random walk is handled on long runs. The length of those runs which is compatible with a given accuracy is discussed; simulated results are presented, which enlight the gain of the present approach over classical IS schemes. When the conditioning event is in the zone of the central limit theorem it provides a tool for statistical inference in the sense that it produces an effective way to implement the Rao-Blackwell theorem for the improvement of estimators.
Keywords: Gibbs principle, conditioned random walk, large deviation, Importance sampling, Rao-Blackwell theorem.

References

Blanchet, J.H., Leder K. and Glynn, P. W. (2009). Efficient simulation of light-tailed sums: an old-folk song sung to a faster new tune.... *Monte Carlo and quasi-Monte Carlo methods 2008,* P. L'Ecuyer and A.B. Owen Eds, Springer-Verlag, 227-248.

Broniatowski M. et Caron V. (2011): Long runs under a condi-

tional limit distribution. Soumis. http://fr.arxiv.org/pdf/1202.0731v1.

Broniatowski M. et Caron, V. (2013): Small variance estimators for rare event probabilities. *ACM TOMACS Special Issue on Monte Carlo Methods in Statistics,* (23):1 (article 7).

Broniatowski M. et Caron V. (2013): Conditional inference in parametric models. Soumis. http://arxiv.org/pdf/1202.0944v1.pdf.

Bucklew, J.A., Ney, P., and Sadowsky, J.S., 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Ann. Probab.*, 27:44-59.

Dembo, A. and Zetouni, O. (1996): Refinements of the Gibbs conditioning principle. *Probab. Theory Related Fields,* 104:1–14.

Diaconis, P. and Freedman, D.A. (1988): Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. *J. Theoret. Probab.,* 1:381–410.

Upper Bounds for the Error in Some Interpolation and Extrapolation Designs

Michel Broniatowski
Université Pierre et Marie Curie (Paris VI), France
michel.broniatowski@upmc.fr

Giorgio Celant
Department of Statistics, University of Padova, Italy
giorgio.celant@unipd.it

This article deals with probabilistic upper bounds for the error in functional estimation defined on some interpolation and extrapolation designs, when the function to estimate is supposed to be analytic. The error pertaining to the estimate may depend on various factors: the frequency of observations on the knots, the position and number of the knots, and also on the error committed when approximating the function through its Taylor expansion. When the number of observations is fixed, then all these parameters are determined by the choice of the design and by the choice estimator of the unknown function.
The scope of the article is therefore to determine a rule for the minimal number of observation required to achieve an upper bound of the error on the estimate with a given maximal probability.

Keywords: extrapolation designs; design of experiments; functional estimation.

References

Broniatowski, M., and Celant, G., 2007. Optimality and bias of some interpolation and extrapolation designs. *J. Statist. Plann.*

Inference **137**, 858–868.

Celant, G., 2003. Extrapolation and optimal designs for accelerated runs. *Ann. I.S.U.P.* **47**, 51–84.

Celant, G., 2002. Plans accélérés optimaux: estimation de la vie moyenne d'un système. *C. R. Math. Acad. Sci. Paris* **335**, 69–72.

Hoel, P.G., Levine, A., 1964. Optimal spacing and weighting in polynomial prediction. *Ann. Math. Statist.* **35**, 1553-1560.

Hoel, P.G., 1965. Optimum designs for polynomial extrapolation. *Ann. Math. Statist.* **36**, 1483–1493.

Luttmann, F.W. and Rivlin, T.J., 1965. Some numerical experiments in the theory of polynomial interpolation. *IBM J. Res. Develop.* **9**, 187–191.

Natanson, I.P. *Constructive Function Theory*, Vol. III, Ungar, New York, 1965.

Rivlin, T.J., 1969. *An introduction to the approximation of functions*. Blaisdell Publishing Co. Ginn and Co.

Spruill, M.C., 1984. Optimal designs for minimax extrapolation. *J. Multivariate Anal.*, **15**, 1, 52–62.

Spruill, M.C., 1987. Optimal designs for interpolation. *J. Statist. Plann. Inference* **16**, 219–229.

Spruill, M.C., 1987. Optimal extrapolation of derivatives. *Metrika* **34**, 45–60.

Asymptotic Permutation Tests in Factorial Designs - Part I

Edgar Brunner
University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany
ebrunne1@gwdg.de

The analysis of factorial designs in the case of unequal sample sizes and variances is a challenging problem, in particular if the normal distribution is not assumed. For large samples, it is known that the Wald-type statistic (WTS) is asymptotically χ^2-distributed under the respective null hypotheses. For small samples, however, the WTS may become extremely liberal, even in the case of the normal distribution. Since this problem is known for a long time, there is an abundance of references related to procedures for testing treatment effects and interactions in completely randomized heteroscedastic one-, two-, and higher-way layouts and also category effects in hierarchical designs. For normal distributions, two easy to implement procedures are the ANOVA- or Welch-James-type statistic (Brunner et al., 1997; Johansen, 1980) which are quite good approximations, see Richter and Payton (2003) and Vallejo et al. (2010) for simulation results. However, in comparison to the WTS, their actual significance levels are even asymptotically unknown. Therefore it is the aim of this talk to present a procedure which is applicable in general unbalanced, heteroscedastic factorial designs. Moreover, the class of error distribution is quite general since only some regularity assumptions on the moments are needed. In particular, we present an asymptotically exact permutation test (Pauly et al., 2013). For the ease of convenience, the problem will be presented in a two-way layout while the results will be stated in general factorial designs. It is shown by simulations that this test keeps the pre-assigned level quite satisfactorily, even in the case of very small sample sizes and different shapes of error distributions in the single samples. The simulation results are not

only presented for exchangeable settings but also for various symmetric and skewed distributions with different combinations of variances and sample sizes (positive and negative pairing). Furthermore, we compare the quality of the approximation with that of some competitors such as the WTS, the ANOVA-type statistic and the classical F-test. The suggested permutation procedure is counter-intuitive (Huang et al., 2006). The theoretical details, however, why it works will be explained in the subsequent talk by Markus Pauly.

Keywords: Analysis of variance, permutation tests, heteroscedastic designs.

References

Brunner, E., Dette, H., Munk, A. (1997): Box-Type Approximations in Nonparametric Factorial Designs, *Journal of the American Statistical Association* 92, pp. 1494-1502.

Huang, Y., Xu, H., Calian, V., and Hsu, J.C. (2006): To permute or not to permute, *Bioinformatics* 22, pp. 2244-2248.

Johansen, S. (1980): The Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression, *Biometrika* 67, pp. 85-92.

Pauly, M., Brunner, E., and Konietschke, F. (2013): Asymptotic Permutation Tests in General Factorial Designs, Preprint.

Richter, S.J. and Payton, M.E. (2003): Performing Two-Way Analysis of Variance Under Variance Heterogeneity, *Journal of Modern Applied Statistical Methods* 2, pp. 152-160.

Vallejo, G., Fernández, M.P., and Livacic-Rojas, P. E. (2010): Analysis of unbalanced factorial designs with heteroscedastic data, *Journal of Statistical Computation and Simulation*, 80, pp. 75-88.

Continuous endpoints in the design of Bayesian two–stage studies

Pierpaolo Brutti, Fulvio De Santis, Stefania Gubbiotti, Valeria Sambucini
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
pierpaolo.brutti@uniroma1.it

Phase II trials are usually designed as single–arm two–stage studies focused on a binary endpoint that is obtained by dichotomizing a continuous variable of clinical interest. However, moving from a continuous to a binary outcome inevitably causes loss of information.

For this reason here we directly consider the unthresholded original variable and we introduce a two–stage design under a Bayesian predictive framework. The same idea was previously proposed for instance by Whitehead *et al.* (2009) and Wason *et al.* (2012) in a frequentist context.

More formally, let X be the variable of interest with unknown mean θ representing treatment efficacy (higher values = higher efficacy) and assume that the trial is considered successful if there is substantial evidence that $\theta > \theta^\star$, where $\theta^\star$ denotes a pre-specified clinically relevant target. The general scheme we propose is an extension of the standard two-stage design for binary observations originally due to Simon (1989) and has the following structure: the trial starts by collecting n_1 measures of X on n_1 patients enrolled in the first stage. If the observed sample mean, $\bar{x}_{n_1}$, is below a suitable threshold r_1, the trial stops for lack of efficacy; otherwise, the trial continues to the second stage. At the second stage we accrue n_2 additional patients and consider the overall observed sample mean $\bar{x}_n$ obtained using all the $n = n_1 + n_2$ measures of X. Then if $\bar{x}_n < r$, where r is a further threshold, the treatment is declared not promising; otherwise the treatment is considered worthy of further evaluation in phase III trials.

The Bayesian predictive approach we develop adjusts the two–stage design proposed by Sambucini (2008) to the continuous setup. In essence,

in both stages we choose the optimal sample sizes in order to control the predictive probabilities of obtaining a large posterior probability that θ exceeds the target $\theta^\star$, under the assumption that the treatment is actually effective. Our proposal relies on the distinction between analysis and design priors adopted, for instance, in the simulation-based approach for sample size determination of Wang and Gelfand (2002). In addition, because of the complex interaction between the predictive and the posterior distributions, all the quantities of interest are evaluated by suitable simulation techniques.

Keywords: Analysis and Design priors, Bayesian predictive approach, Phase II clinical trials, Two–stage design.

References

Sambucini, V. (2008): A Bayesian predictive two-stage design for phase II clinical trials. *Stat Med*, Vol. 27, N. 8, pp. 1199-1224.

Simon, R. (1989): Optimal two-stage designs for phase II clinical trials, *Control Clin Trials*, Vol. 10, pp. 1-10.

Wang F. and Gelfand A.E. (2002): A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models, *Stat Sci*, Vol. 17, N. 2, 193-208.

Wason, J.M., Mander, A.P. and Thompson, S.G. (2012): Optimal multistage designs for randomised clinical trials with continuous outcomes, *Stat Med*, Vol. 31, N. 4, pp. 301-312.

Whitehead J., Valdès-Màrquez E. and Lissmats A. (2009): A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain, *Pharm Stat*, Vol. 8, N. 2, pp. 125-135.

Sensitivity analysis and optimal control of some applied probability models

Ekaterina Bulinskaya
Lomonosov Moscow State University, Russia
ebulinsk@yandex.ru

It is well known that in order to investigate some real process or system one has to construct an appropriate mathematical model. Before using the model it is necessary to carry out the sensitivity analysis, that is, to be sure in the model stability to small parameters fluctuations and underlying processes perturbations, see, e.g. (Bulinskaya, 2007).

To illustrate the procedure we consider a periodic-review inventory model with several suppliers generalizing those introduced in (Caliskan-Demirag et al., 2012), (Huggins, Olsen, 2010) and (Papachristos, Katsaros, 2008).

To simplify the presentation suppose here that there are two suppliers and the second one is unreliable. That means, the order is delivered immediately with probability p and with one-period delay with probability $q = 1 - p$. Moreover, we assume that the demand is described by a sequence of i.i.d. r.v.'s and there exist the replenishment orders constraints. Below we treat the budget restriction, namely, the money amount available for orders at both suppliers is bounded by a fixed quantity A. Let x be the initial inventory level. Denote by c_i the unit order cost at the i-th supplier, $a_i = A/c_i$, $i = 1, 2$, and r the unit shortage penalty. Put also $\Delta^0 = \{0 \le c_2 \le pr\}$ and $\Delta^k = \{r(p + \sum_{i=1}^{k-1} \alpha^i) < c_2 \le r(p + \sum_{i=1}^{k} \alpha^i)\}$ for $k \ge 1$ where α is discount factor. Our aim is to choose the order quantities minimizing the n step discounted costs. The following theorem shows that the optimal order policy is characterized by a sequence of critical levels.

Theorem. *Let the order costs satisfy the following relations* $c_2 < c_1 - qr$ *and* $\{(c_1, c_2) \in \Delta^k\}$, $k \ge 0$. *Then it is optimal to order nothing at the first supplier for any* n *and at the second supplier for* $n \le k$.

There also exists an increasing sequence $\{u_n\}_{n\geq k+1}$ *such that at the second supplier the optimal order size is* a_2 *for the* $x < u_n - a_2$*, it is equal to* $u_n - x$ *for* $x \in [u_n - a_2, u_n)$ *and one orders nothing for* $x \geq u_n$.

The other constraints and relations between c_i, $i = 1, 2$, leading to different order policies are considered as well. The stability conditions are established.

Keywords: decisions under uncertainty, optimal control, sensitivity analysis.

Acknowledgements: This work was supported by RFBR grant 13-01-00653.

References

Bulinskaya E.V. (2007): Sensitivity analysis of some applied models, *Pliska Studia Mathematica Bulgarica*, Vol. 18, pp. 57-90.

Caliskan-Demirag O., Chen Y. and Yang Yi. (2012): Ordering policies for periodic-review inventory systems with quantity-dependent fixed costs, *Operations Research*, Vol. 60, pp. 785-796.

Huggins E.L., Olsen T-L. (2010): Inventory control with generalized expediting, *Operations Research*, Vol. 58, pp. 1414-1426.

Papachristos S., Katsaros A. (2008): A periodic-review inventory model in a fluctuating environment. *IIE Transactions*, Vol. 40, pp. 356-366.

Effective classification of branching processes with several points of catalysis

Ekaterina Vl. Bulinskaya
Lomonosov Moscow State University
bulinskaya@yandex.ru

We introduce and study the model of *generalized catalytic branching process* (GCBP) which is a system of particles moving in space and branching only in the presence of catalysts. More exactly, let at the initial time there be a particle which moves on some finite or countable set S according to a continuous-time Markov process with infinitesimal generator A. When this particle hits a finite set $W = \{w_1, \dots, w_N\} \subset S$ of catalysts, say at site w_k, it spends there time having the exponential distribution with parameter 1. Afterwards the particle either branches or leaves site w_k with probability α_k and $1 - \alpha_k$ ($0 < \alpha_k < 1$), respectively. If particle branches (at site w_k), it may produce a random non-negative integer number ξ_k of offsprings. It is assumed that all newly born particles behave as independent copies of their parent.

The particular case of GCBP was considered in Doering, Roberts (2013) when W consists of a single catalyst. The main tool for the moment analysis of the process was the spine technique, that is "many-to-few lemma", and renewal theory. Another case, for $S = \mathbb{Z}^d$, $d \in \mathbb{N}$, and the Markov chain being a symmetric, homogeneous and irreducible random walk with a finite variance of jump sizes, was investigated by Yarovaya (2012). There the necessary and sufficient conditions for exponential growth of the mean particles numbers were established due to application of the spectral theory to evolution operators.

To implement the moment analysis of GCBP we employ other methods. To this end we involve the hitting times with taboo (see, e.g., Bulinskaya (2013)) and introduce an auxiliary Bellman-Harris process with $N(N+1)$ types of particles extending the approach proposed by Topchii, Vatutin (2013). Then we use the criticality conditions for a

multi-type Bellman-Harris process and the theorems on the long-time behavior of its moments which can be found in Mode (1971). So, this technique allows us to generalize the results by Doering, Roberts (2013) as well as by Yarovaya (2012).

Keywords: catalytic branching process, moment analysis, hitting times with taboo, multi-type Bellman-Harris process, criticality conditions.

Acknowledgements: This work is partially supported by Dmitry Zimin Foundation "Dynasty".

References

Bulinskaya E.Vl. (2013): Local Particles Numbers in Critical Branching Random Walk, *Journal of Theoretical Probability*, DOI 10.1007/s10959-012-0441-4, arXiv:1203.2362 [math.PR].

Doering L., Roberts M. (2013): Catalytic Branching Processes via Spine Techniques and Renewal Theory. In: Donati-Martin C., Lejay A., Rouault A. (Eds.) *Séminaire de Probabilités XLV*, Lecture Notes in Mathematics, Springer-Verlag, Berlin-Heidelberg, arXiv:1106.5428 [math.PR].

Mode C.J. (1971): *Multi-type Branching Processes. Theory and Applications*, Elsevier Publishing Co. Ltd., London.

Topchii V.A., Vatutin V.A. (2013): Catalytic Branching Random Walk in $\mathbb{Z}^d$ with branching at the origin only, *Siberian Advances in Mathematics*, Vol. 23, N. 1, pp. 28-72.

Yarovaya E.B. (2012): Spectral properties of evolutionary operators in branching random walk models, *Mathematical Notes*, Vol. 92, N. 1-2, pp. 115-131.

Development of MDR method

Alexander Bulinski
Lomonosov Moscow State University
bulinski@yandex.ru

The problems concerning the high dimensional data analysis are discussed. We consider a binary response variable Y which depends on some factors $X_1, \ldots, X_n$. In a number of medical and biological studies, e.g., in genetics, such Y can describe the state of a patient health. For example $Y = 1$ and $Y = -1$ mean "sick" and "healthy", respectively. One can assume that there are genetic and non-genetic risk factors provoking specified complex diseases such as diabetes, hypertension, myocardial infarction and others.

Many researchers share the paradigm that the impact of any single factor can be rather small (non-dangerous) whereas certain combinations of these factors can lead to significant effect. Moreover, usually one assumes that the response variable depends only on some part of factors. A challenging problem in modern genetics is to identify the collection of factors responsible for increasing the risk of specified complex disease. The progress in the human genome reading (especially the micro-chip techniques) permitted to collect the genetic datasets for analysis by means of various complementary statistical tools. The theoretical contributions are provided along with various simulation procedures. The review of investigations in genome-wide association studies (GWAS) during the last five years is given, e.g., in Visscher et al. (2012).

Here we concentrate on the multifactor dimensionality reduction (MDR) method introduced by M.D.Ritchie et al. (2001) and its further development. Following Bulinski (2012) we study the estimate of prediction error of Y by means of a function of discrete random variables $X_1, \ldots, X_n$. To this end we employ the penalty function and use the K-cross validation procedure. In this way it is possible to justify the choice of significant collection of factors. We also tackle the applica-

tions of this approach to analysis of risks of complex diseases started in Bulinski et al. (2012).

Keywords: Binary response variable, selection of significant factors, penalty function and prediction error, cross-validation, new versions of MDR method.

Acknowledgements: This work is supported by RFBR grant 13-01-00612.

References

Bulinski, A.V.(2012): To the foundations of the dimensionality reduction method for explanatory variables, *Zapiski Nauchnyh Seminarov POMI*, Vol. 408, pp. 84–101 (in Russian; English translation: *Journal of Mathematical Sciences*).

Bulinski A., Butkovsky O., Sadovnichy V., Shashkin A., Yaskov P., Balatskiy A., Samokhodskaya L., Tkachuk V. (2012): Statistical methods of SNP data analysis and applications, *Open Journal of Statistics*, Vol. 2, N. 1, pp. 73–87.

Ritchie M.D., Hahn L.W., Roodi N., Bailey R.L., Dupont W.D., Parl F.F., Moore J.H. (2001): Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *The American Journal of Human Genetics*, Vol. 69, N. 1, pp. 138–147.

Visscher P.M., Brown M.A., McCarthy M.I., Yang J. (2012): Five years of GWAS discovery. *The American Journal of Human Genetics*, Vol. 90, N. 1, pp. 7-24.

Enhancement of the Acceleration Oriented Kinetic Model for the Vehicular Traffic Flow

A.V. Burmistrov
Institute of Computational Mathematics and Mathematical Geophysics
SB RAS, prospect Akademika Lavrentjeva, 6, Novosibirsk, Russia, 630090
Novosibirsk State University, Pirogova str., 2, Novosibirsk, Russia, 630090
burm@osmf.sscc.ru

We develop our algorithms in the frame of the kinetic VTF model suggested in [Waldeer, 2004]. A distinctive feature of this model consists in introducing of the acceleration variable into the set of phase coordinates along with the velocity coordinate of the car. Such a modification of the phase space allowed to describe not only a constrained traffic but also a higher car density regimes.

For a single car with acceleration a and velocity v the probability density $f(\cdot)$ solves the integro-differential equation of Boltzmann type:

$$\left(\frac{\partial}{\partial t} + a\frac{\partial}{\partial v}\right) f(a,v,t) = \int\limits_{\bar{a},\bar{v},a'} \big[\Sigma(a|a',v,\bar{a},\bar{v}) f(a',v,t) - \Sigma(a'|a,v,\bar{a},\bar{v}) f(a,v,t)\big]\, f(\bar{a},\bar{v},t)\, \mathtt{d}\bar{a}\, \mathtt{d}\bar{v}\, \mathtt{d}a', \tag{1}$$

with the initial distribution $f(a,v,0) = f_0(a,v)$. Here $\bar{a}$ and $\bar{v}$ are the acceleration and the velocity of the leading car, which interacts with the current car situated straight behind it. The function $\Sigma(\cdot)$ is a weighted interaction rate function. As the car acceleration a is added to the phase coordinates, there are only acceleration jumps (no velocity jumps, as in other kinetic models) produced by the pairwise interactions in the system. Moreover, after the interaction takes place, the leader does not change its acceleration. Therefore the function $\Sigma(\cdot)$ is not symmetric.

In previous works [Burmistrov & Korotchenko, 2011, 2012] we succeeded to construct the basic integral equation of the second kind $F = \mathbf{K}F + F_0$. Its solution F is closely connected with the solution

$f(a, v, t)$ to the equation (1), while the kernel $\mathbf{K}$ describes the evolution of the N-particle system of vehicles. The integral equation enables us to use well-developed techniques of the Monte Carlo simulation for estimating the functionals of solution to the equation (1), as well as to perform parametric analysis [Burmistrov & Korotchenko, 2012].

In this work we are going to take into consideration such aspects as variety of vehicle classes (trucks and cars), diversity of driver behaviors (conservative or more aggressive), multilane traffic with overtaking and vehicle grouping on the road. It will result in more realistic interaction profiles for the model under study.

Keywords: Monte Carlo Simulation, N-Particle System, Markov Chain, Integral Equation of the Second Kind.

Acknowledgements: This work was partly supported by the Russian Foundation for Basic Research (grants 11-01-00252, 12-01-31134).

References

Waldeer K. T. (2004): A Vehicular Traffic Flow Model Based on a Stochastic Acceleration Process, *Transport Theory and Statistical Physics*, Vol. 33, N. 1, pp. 7-30.

Burmistrov A.V., Korotchenko M.A. (2011): Statistical Modeling Method for Kinetic Traffic Flow Model with Acceleration Variable. *Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" – AMSA-2011, Novosibirsk, Russia, 20-22 September, 2011.* Novosibirsk: Publishing House of NSTU, pp. 411-419.

Burmistrov A.V., Korotchenko M.A. (2012): Weight Monte Carlo Method Applied to Acceleration Oriented Traffic Flow Model. In: L. Plaskota and H. Wozniakowski (Eds.) *Monte Carlo and Quasi-Monte Carlo Methods 2010*, Springer Proceedings in Mathematics & Statistics, Springer-Verlag, Berlin-Heidelberg, pp. 277-291.

Exponential inequalities for the distribution tails of multiple stochastic integrals with Gaussian integrators

Alexander Bystrov
Novosibirsk State University
bystrov@ngs.ru

Let $\xi(t)$ be a centered Gaussian process on $[0,1]$ with the covariance function $\Phi(t,s)$. Denote

$$\Delta\Phi(t,s) = \mathbf{E}\,(\xi(t+\delta)-\xi(t))\,(\xi(s+\delta)-\xi(s)),\ \ t,s,t+\delta,s+\delta \in [0,1], \delta > 0.$$

We consider the case when

$$\Delta\Phi(t,s) = \begin{cases} \delta g_1(t) + o(\delta),\ t = s, \\ \delta^2 g_2(t,s) + o(\delta^2),\ \ t \neq s, \end{cases} \tag{1}$$

as $\delta \to 0$, where the functions $g_1(t), g_2(t,s)$ are assumed to be bounded. We study a multiple stochastic integral of the form

$$I(f) := \int_{[0,1]^m} f(t_1,...,t_m) d\xi(t_1)...d\xi(t_m)$$

which was defined in (Borisov, Bystrov, 2005).

Theorem 1. *Let f be a bounded function and $\Phi(t,s)$ meets* (1)*. Then*

$$\mathbf{P}(|I(f)| > x) \leq C_1(m) \exp\left\{ - \left(\frac{x}{C_2(m,f,g_1,g_2)} \right)^{1/d} \right\}. \tag{2}$$

To prove inequality (2), we use the following version of Chebyshev's power inequality:

$$\mathbf{P}(|I(f_N)| > x) \leq \inf_k \; x^{-2k}\, \mathbf{E}|I(f_N)|^{2k},$$

where $\{f_N\}$ is a sequence of step functions converging to f in certain kernel space. The main problem here is to obtain a suitable upper bound for the even moment on the right-hand side of this inequality.

We also study a subclass of the integrating processes $\xi(t)$ when more exact (and in some sense optimal) inequality takes place:

$$\mathbf{P}(|I(f)| > x) \leq C_1(m) \exp\left\{ - \left(\frac{x}{C_2(m, f, g_1, g_2)} \right)^{2/d} \right\}. \qquad (3)$$

The subclass includes processes which arise as limit elements for sequences of standard empirical processes in case of weakly dependent observations. In this case $I(f)$ is the weak limit for the sequences of normalized degenerate V-statistics with kernel f (Borisov, Bystrov, 2006)

Keywords: Stochastic integrals, exponential inequalities, Gaussian processes, weak dependence, V-statistics.

References

Borisov, I. S. and Bystrov, A. A. (2005): Constructing a stochastic integral of a non-random function without orthogonality of the noise., *Theory Probab. Appl.*, N. 50, pp. 52-80.

Borisov, I. S. and Bystrov, A. A. (2006):Stochastic integrals and asymptotic analysis of canonical von Mises statistics based on dependent observations., *IMS Lecture Notes Monograph Series*, N. 51, pp. 1-17.

Ranking of Multivariate Populations in Case of Very Small Sample Sizes

Eleonora Carrozzo
Department of Management and Engineering, University of Padova, Italy
eleonora.carrozzo@unipd.it

Livio Corain
Department of Management and Engineering, University of Padova, Italy
livio.corain@unipd.it

When the response variable is multivariate in nature, the need to define an appropriate ranking related to populations of interest such as products, services, teaching courses, degree programs, and so on is very common in both experimental and observational studies in many areas of applied research. Recently Arboretti et al. (2010) proposed the application of the NonParametric Combination (NPC) methodology (Pesarin and Salmaso, 2010) to develop a nonparametric solution for this kind of problems which has proved to be particularly effective when the underlying data generation mechanism is non-normal in nature. The purpose of this work is to extend and validate the proposal of Arboretti et al. (2010) in case of very small sample sizes coming up to consider also the non replicated design, provided that the variances/covariances can be assumed as known. In order to validate our proposal we performed a comparative simulation study where we consider as benchmark an heuristic method proposed by literature (Musci et al., 2011). As confirmed by the simulation study and by the application to a real case study in the field of developing new products for laundry industry, we can state that the proposed ranking method for multivariate populations is certainly a valid and effective solution also in case of very small sample sizes.

Keywords: global ranking, permutation tests, nonparametric combination.

References

Arboretti Giancristofaro R., Corain L., Gomiero D., Mattiello F. (2010): Nonparametric Multivariate Ranking Methods for Global Performance Indexes, *Quaderni di Statistica*, Vol. 12, pp. 79-106.

Musci R., Cordellina A., Crestana A. (2011): A Novel Process for Ranking Products in Detergent Tests: GPS-Tools. Proceedings of the *41st CED - Comite Espanol De La Detergencia Annual Meeting: "Surfactants, detergents and cosmetics. From science to implementation"*, March 6-7, 2011, Barcelona, Spain, p. 4.

Pesarin F., Salmaso L. (2012): *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.

Partially Adaptive Estimation of an Ordered Response Model Using a Mixture of Normals

Steven B. Caudill
Rhodes College
caudills@rhodes.edu

In the usual ordered probit model the error variance is fixed and equal to one, and flexible probability estimates are achieved by adjusting the cutoffs, which are estimated. In our approach, the interval cutoffs are fixed and flexibility is incorporated by allowing the error variances in the two regimes in the mixture model to differ. In our approach $\delta_0 = -\infty, \delta_1 = 0,$ and $\delta_K = +\infty$. This yields the following partition and relationship between latent y and observed y

$$\begin{gathered} -\infty \leq 0 \leq 1 \leq 2 \leq 3 \leq \cdots \delta_{K-1} \leq +\infty \\ y = 1 \text{ if } y_i^* \leq 0 \\ y = 2 \text{ if } 0 \leq y_i^* \leq 1 \\ y = 3 \text{ if } 1 \leq y_i^* \leq 2 \\ y = 4 \text{ if } 2 \leq y_i^* \leq 3 \\ etc. \end{gathered}$$

The density function upon which our probabilities are based is given by

$$f(y_i^*; \lambda, \mu_{i1}, \sigma_1^2, \mu_{i2}, \sigma_2^2) = \lambda\phi(y_i^*; \mu_{i1}, \sigma_1^2) + (1-\lambda)\phi(y_i^*; \mu_{i2}, \sigma_2^2) \quad (1)$$

Using this density, probabilities based on a mixture of two normal densities is given by

$$P_{ik} = prob(y_i = k) = F[k] - F[k-1] \text{ for } k = 1 \ldots K \quad (2)$$

In practice, the method has proven to be extremely flexible. The model estimated is an immediate application of the EM algorithm given in Caudill and Long (2010). Preliminary results using data from Dustmann and van Soest (2004) on the English language proficiency of Indian men indicates that the partially adaptive model performs better (has a higher maximized value of the likelihood function) than the usual ordered probit model or the generalized ordered probit model.

Keywords: Normal mixture, EM algorithm, ordered probit.

References

Caudill S.B., Long J.E. (2010): Do Former Athletes Make Better Managers? Evidence from a Partially Adaptive Grouped-Data Regression Model, *Empirical Economics*, Vol. 39, N. pp. 275-90.

Dustmann C., Van Soest A. (2004): An Analysis of Speaking Fluency of Immigrants Using Ordered Response Models With Classification Errors, *Journal of Business & Economic Statistics*, Vol. 22, N. pp. 312-321.

Sparse factor models for high-dimensional interaction networks

David Causeur
Agrocampus, IRMAR, UMR 6625 CNRS
david.causeur@agrocampus-ouest.fr

Analysis of data generated by high-throughput technologies has received an increased scrutiny in the statistical literature, especially motivated by emerging challenges in systems biology, neuroscience or astrology. Microarray technologies for genome analysis or brain imaging and electroencephalography share the common goal of providing a detailed overview of complex systems on a large scale. Statistical analysis of the resulting data usually aims at identifying key components of the whole system essentially by large-scale significance, regression or supervised classification analysis. However, usual issues such as the control of the error rates in multiple testing or model selection in classification turns out to be challenging in high dimensional situations. Some papers (Leek and Storey, 2007 and 2008, Friguet *et al.*, 2009, Causeur *et al.*, 2012) have especially pointed out the negative impact of dependence among tests on the consistency of the ranking which results from multiple testing procedures in high dimension. These papers essentially show that unmodeled heterogeneity factors can result in an unexpected dependence across data, which generates a high variability in the actual False Discovery Proportion and more generally affects the efficiency of the classical simultaneous testing methods. Linear modelling of latent effects therefore appears as a general and flexible framework for dependence in high-dimensional data.

Models for interaction network among the components of a complex system are often used to give more insight to a list of selected features of a complex system. They often reveal some key components which changes lead to variations of other connected components. This suggests that it is crucial to account for the system-wide dependence

structure to select these regulators. A sparse factor model is proposed to identify a low-dimensional linear kernel which captures data dependence. ℓ_1-penalized estimation algorithms are presented and strategies for module detection in both relevance and Graphical Gaussian Models for networks are deduced. The properties are illustrated by issues in statistical genomics (see Blum *et al*, 2010).

Keywords: High dimension, Factor model, Graphical Gaussian Model, Interaction network, LASSO.

References

Blum, Y., Le Mignon, G., Lagarrigue, S. and Causeur, D. (2010). A factor model to analyse heterogeneity in gene expressions. *BMC Bioinformatics*. 11:368.

Causeur, D., Chu, M.-C., Hsieh, S. and Sheu, C.-F. (2012) A factor-adjusted multiple testing procedure for ERP data analysis. *Behavior Research Methods*. **44** (3), 635–643.

Friguet, C., Kloareg, M. and Causeur, D. (2009) A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104, 1406–1415.

J.T. Leek and J. Storey. (2007) Capturing heterogeneity in gene expression studies by Surrogate Variable Analysis. *PLoS Genetics*, 3, e161.

J.T. Leek and J. Storey. A general framework for multiple testing dependence. (2008) *Proceedings of the National Academy of Sciences*, 105, 18718–18723.

A statistical approach to the H index

Paola Cerchiello
University of Pavia
paola.cerchiello@unipv.it

Paolo Giudici
University of Pavia
giudici@unipv.it

The measurement of quality of academic research is a rather controversial issue. Recently, in 2005, Hirsch has proposed a measure that has the advantage of summarizing in a single summary statistics all the information that is contained in the citation counts of each author. From that seminal paper, a huge amount of research has been lavished, focusing, on one hand on the development of correction factors to the H index and, on the other hand, on the pros and cons of such measure proposing several possible alternatives. In the present work, we propose an exact, rather than asymptotic, statistical approach and, to achieve this objective, we work directly on the two basic components of the H index: the number of produced papers and the related citation counts vector. Such quantities will be modelled by means of a compound stochastic distribution, that exploits, rather than eliminate, the variability present in both the production and the impact dimensions of a scientist's work.

Our proposal is evaluated on a database of homogeneous scientists made up of 131 full professors of statistics employed in Italian universities. These scientists form a cohort of people that has grown their careers under similar conditions: both in terms of academic rules (they belong to the same country) and in terms of research modus operandi (they belong to the same scientific community). Such database has been collected by a public organization named VIA-Academy (www.via- academy.org) that aims at improving the quality of Italian scientists by providing open feedbacks on their research quality on a bibliometric basis. We have cleaned and refined the data, and added for each scientists, her/his cita-

tion counts vector. The refinement has involved a long activity of disambiguation (from homonimies and wrong affiliations) that was carried out by employing the well known Publish or Perish. On the basis of our novel approach we show interesting results in terms of alternative measures of quality that can be complementary to the H index.

Keywords: H index, convolution, extreme values distributions

References

Ball, P. (2005): Index aims for fair ranking of scientists. *Nature 436 (7053): 900*.

Beirlant, J. and Einmahl, J. H. J. (2010): Asymptotics for the Hirsch index. *Scandinavian Journal of Statistics*, VOL. 37, pp. 355-364.

Cerchiello, P. and Giudici, P. (2012): On the distribution of functionals of discrete ordinal variables. *Statistics and Probability Letters*, VOL. 82, pp. 2044-2049.

Hirsch, J. E. (2005): An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, pp. 16569-16572.

Pratelli, L., Baccini, A., Barabesi, L. and Marcheselli, M. (2012): Statistical Analysis of the Hirsch Index. *Scandinavian Journal of Statistics*, VOL. 39, pp. 681-694.

Fixed-Width Confidence Intervals and Asymptotic Expansion of Percentiles for the Standardised Version of Sample Location Statistic

Bhargab Chattopadhyay
Department of Mathematical Sciences, The University of Texas at Dallas
bhargab@utdallas.edu

Nitis Mukhopadhyay
Department of Statistics, University of Connecticut
nitis.mukhopadhyay@uconn.edu

A parametric approach for constructing fixed-width confidence interval for the location parameter when samples come from a location scale family will be presented. We revisit Mukhopadhyay (1982)'s two-stage procedure in such situation and will propose a modified two-stage procedure for finding fixed-width ($= 2d$) confidence interval of a location parameter with a pre-assigned confidence coefficient ($>= 1 - \alpha$). In addition a method to compute the asymptotic expansion of the percentile point of the standardised version of the sample location will be presented which will be helpful in achieving the second-order efficiency of the modied two-stage procedure. These percentile points can be used for proposing tests of hypotheses or confidence intervals of the location parameter when samples arrive from a distribution with unknown location and scale parameter.

As an illustration we have asymptotically expressed the percentile point $b_{m,\alpha}$; of a test pivot based on ginis mean dierence (GMD). Based on large-scale simulations, approximations, and data analyses, we report that our methodology can be used in case when observations arrive from Normal distribution.

Keywords: Location-scale family; Taylor expansion; Asymptotic Efficiency;

Asymptotic Consistency.

References

Chattopadhyay, B., Mukhopadhyay, N. (2013): Two-Stage Fixed-Width Confidence Intervals for a Normal Mean in the Presence of Suspect Outliers, *Sequential Analysis*, in press.

Mukhopadhyay, N. (1982): Stein's two-stage procedure and exact consistency, *Scandinavian Actuarial Journal*,1982 (2),110-122.

Nonparametric testing goodness-of-fit of a regression reliability function using the Beran estimator

Ekaterina Chimitova, Victor Demin
Novosibirsk State Technical University, Novosibirsk, Russia
ekaterina.chimitova@gmail.com

The Beran estimator (Beran (1981)) is one of the most popular nonparametric estimators for a conditional reliability function under the given value of the observed covariate. The statistical properties of the Beran estimator were studied in Dabrowska (1992), Gonzalez and Cadarso (1994), McKeague and Utikal(1990), Van Keilegom, Akritas and Veraverbeke (2001).

Let us denote survival times or failure times which depend on the scalar covariate x as T_x. The reliability function is defined as $S(t|x) = P(T_x \geq t) = 1 - F(t|x)$. Let $\mathbf{Y} = \{(Y_1, x_1), (Y_2, x_2), ..., (Y_n, x_n)\}$ be the sample of observations, where n is the sample size, x_i is the value of a covariate for the i-th object, Y_i is the survival time.

The Beran estimator is defined as

$$\tilde{S}_{h_n}(t|x) = \prod_{Y_{(i)} \leq t} \left\{ 1 - \frac{W_{n(i)}(x; h_n)}{1 - \sum_{j=1}^{i-1} W_{n(j)}(x; h_n)} \right\},$$

where x is the value of a covariate for which the reliability function has been estimated; $W_{n(i)}(x; h_n), i = 1, ..., n$ are the Nadaraya-Watson weights

$$W_{n(i)}(x; h_n) = K\left(\frac{x - x_i}{h_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right),$$

where $K\left(\frac{x-x_i}{h_n}\right)$ is the kernel function, h_n is the smoothing parameter.

In this paper we propose the method of choosing an optimal smoothing parameter for the Beran estimator. This method is based on the minimization of the observed deviation of lifetimes from a nonparametric estimator of the inversed reliability function obtained by kernel smoothing. By means of the Monte-Carlo simulations it has been shown that the method results in more precise estimates than when using a fixed smoothing parameter. We propose a goodness-of-fit test for parametric and semiparametric reliability regression models which is based on the distance between the Beran nonparametric estimator and the tested conditional reliability function for given values of the covariate. The power of the proposed test has been investigated for various pairs of competing hypotheses.
Keywords: conditional reliability function, Beran's estimator, kernel smoothing, goodness-of-fit tests.

References

Beran R. (1981). Nonparametic regression with randomly censored survival data. *Technical report*. Department of Statistics, University of California, Berkeley.

Dabrowska D.M. (1992). Nonparametric quantile regression with censored data. *Sankhya Ser. A. 54*, 252-259.

Gonzalez M.W., Cadarso S.C. (1994) Asymptotic properties of a generalized Kaplan-Meier estimator with some application, *J. Nonparametric Statistics, 4*, 65-78.

McKeague I.W., Utikal K.J.(1990) Inference for a nonlinear counting process regression model. *Ann. Statist., 18.*, 1172-1187.

Van Keilegom I., Akritas M.G., Veraverbeke N. (2001) Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics & Data Analysis Vol. 35*, 487-500.

Group-Sequential Response-Adaptive Designs

Steve Coad
Queen Mary, University of London
d.s.coad@qmul.ac.uk

Suppose that two treatments are being compared in a clinical trial. Then, as the trial progresses, one of the treatments may look more promising and it would be desirable to allocate a higher proportion of patients to this treatment. A response-adaptive randomization rule can be used to reduce the number of patients on the inferior treatment. The simplest such rules may be represented as urn models, in which balls of different types are added or removed from the urn according to previous assignments and responses. Alternatively, sequential maximum likelihood estimation rules can be used in which optimal treatment assignment probabilities are derived and the unknown parameters are replaced by their current maximum likelihood estimates.

Although most of the existing work deals with response-adaptive randomization in the context of a fixed trial size, it is often more efficient to conduct a trial group sequentially. For normal data with known variances, Jennison and Turnbull (2001) showed that response-adaptive randomization can be incorporated into a general family of group sequential tests without affecting the error probabilities. This is achieved when the group sizes do not depend on the estimated mean responses at the previous stage in any other way but through their difference. Morgan and Coad (2007) considered binary response trials, and showed that the drop-the-loser rule is the most effective allocation rule.

Zhu and Hu (2010) proposed a general group-sequential response-adaptive procedure and proved that the sequential test statistics asymptotically follow the canonical joint distribution in Jennison and Turnbull (2001). An error spending function is used to obtain the appropriate stopping boundaries at the different interim analyses, thus allowing un-

equal information levels. Since the sequential maximum likelihood estimation rules considered are not optimal, we use simulation to study the finite-sample performance of the efficient randomized-adaptive designs of Hu, Zhang and He (2009), which attain the Cramér-Rao lower bound for the variance of the allocation proportion.

Keywords: asymptotically best rule, error spending function, interim analysis, sequential maximum likelihood estimation, variability.

Acknowledgements: This is joint work with Hsiao Yin Liu at Queen Mary, University of London, who is supported by a Scholarship from the Ministry of Education in Taiwan.

References

Hu F., Zhang L.-X., He X. (2009): Efficient Randomized-Adaptive Designs, *The Annals of Statistics*, Vol. 37, N. 5A, pp. 2543-2560.

Jennison C., Turnbull B.W. (2001): Group Sequential Tests with Outcome-Dependent Treatment Assignment, *Sequential Analysis*, Vol. 20, N. 4, pp. 209-234.

Morgan C.C., Coad D.S. (2007): A Comparison of Adaptive Allocation Rules for Group-Sequential Binary Response Clinical Trials, *Statistics in Medicine*, Vol. 26, N. 9, pp. 1937-1954.

Zhu H., Hu F. (2010): Sequential Monitoring of Response-Adaptive Randomized Clinical Trials, *The Annals of Statistics*, Vol. 38, N. 4, pp. 2218-2241.

A Further Study of the Randomized Play-the-Leader Design

Steve Coad
Queen Mary, University of London
d.s.coad@qmul.ac.uk

Nancy Flournoy
University of Missouri
flournoyn@missouri.edu

Caterina May
Università del Piemonte Orientale
caterina.may@unipmn.it

A response-adaptive design may be considered asymptotically optimal when it assigns patients to the best treatment with an allocation proportion converging to one, so that each treatment is assigned infinitely often. This property is ethically very desirable in a clinical trial. Both the rule generated by a two-color, randomly reinforced urn and the play-the-leader (PL) rule of Durham and Yu (1990) have this property.

Since the seminal work of Durham and Yu (1990), the randomly reinforced urn designs have been widely studied in the literature, as reviewed in Flournoy, May and Secchi (2012). For example, Flournoy, May, Moler and Plo (2010) have compared their performance with other response-adaptive designs by simulation.

To our knowledge, the randomized PL rule has not received any further attention. Durham and Yu (1990) proved that, for dichotomous responses, the mean allocation proportion for the best treatment grows at an exponential rate or faster. In this sense, the randomized PL rule is superior to the randomly reinforced urn design. However, no results were provided concerning finite-sample properties.

The aim of the present work is to investigate further the performance of PL designs. The whole distribution and the trajectories of treatment

allocations are investigated by simulation. Particular attention is paid to the presence of extreme trajectories, and hence early sequences are analyzed. Moreover, a comparison between the performances of the two treatment allocation rules described here is a focus of the study.

Keywords: efficiency, ethical alocation, randomly reinforced urn, response-adaptive design, simulation.

Acknowledgements: This work was initiated while the authors were Visiting Fellows at the Design and Analysis of Experiments programme at the Isaac Newton Institute for Mathematical Sciences in Cambridge during July-December 2011.

References

Durham S.D., Yu K.F. (1990): Randomized Play-the Leader Rules for Sequential Sampling from Two Populations, *Probability in Engineering and Information Science*, Vol. 4, pp. 355-367.

Flournoy N., May C., Moler J.A., Plo F. (2010): On Testing Hypotheses in Response-Adaptive Designs Targeting the Best Treatment. In: Giovagnoli A., Atkinson A.C., Torsney B., May C. (Eds.) *mODa 9 - Advances in Model-Oriented Design and Analysis*, Physica-Verlag HD, Berlin, pp. 81-88.

Flournoy N., May C., Secchi P. (2012): Asymptotically Optimal Response-Adaptive Designs for Allocating the Best Treatment: An Overview. *International Statistical Review*, Vol. 80, N. 2, pp. 293-305.

Monte Carlo Sampling Using Parallel Processing for Multiple Testing in Genetic Association Studies

Chris Corcoran
Utah State University
chris.corcoran@usu.edu

Pralay Senchaudhuri, William Welbourn
Cytel Software Corporation, Myriad Genetics
pralay@cytel.com, bill.welbourn@aggiemail.usu.edu

In this paper we introduce a parallel processing approach for Monte Carlo simulation in large-scale genetic association studies. The mapping of the human genome and the rapid advancement of genotyping technology has led to an extraordinary increase in genetic association studies with complex disease. The focus of these studies has evolved dramatically over the past two decades, from investigations of relatively few targeted candidate genes with hypothesized biological effects, to so-called genome-wide hypothesis-free scans (to identify or implicate new genes) involving hundreds of thousands or even millions of genetic markers from across the human genome. Marker panels of 1-2M SNPs are now common for genome-wide studies, and developing technologies (such as exome or whole-genome sequencing) will allow routine comparisons over marker sets that are orders of magnitude larger. With so many hypothesis tests, the need to preserve the rate of false positive findings presents some critical statistical and computational difficulties. Existing methods and their implementations often perform poorly under common conditions. For example, investigators generally apply a Bonferroni-type correction to control the nominal overall false positive rate. However, this presumes independence between tests, which can lead to significant conservatism where dependence exists. An alternative method utilizes the joint distribution of markers under the complete null hypothesis (i.e., no association with any marker). As opposed to adjusting p-values according to the same minimum distribution, the p-value single-step method applies the distribution of the minimum p-value to

the observed minimum, followed by the same adjustment to the minimum of the remaining p-values, and so on. This is the so-called minP adjustment (Westfall and Young, 1993), which reduces conservatism in large part by accounting for the dependence between p-values across all tests (Dudoit et al, 2003). The desired distribution is obtained generally through Monte Carlo simulation of the exact joint permutation distribution of the test statistics. Although such an approach is straightforward in principle, it has been impractical or infeasible for most genome-wide studies (e.g., Han et al, 2009; Gao et al, 2008). To significantly accelerate the required resampling for permutation-based minP, we propose a parallel processing algorithm, which in application has reduced computing times to a fraction of those observed using conventional approaches and software.

Keywords: genetic association, multiple testing, Monte Carlo, parallel processing, permutation test.

Acknowledgements: This work was supported by NIH grant R43 HG004027.

References

Dudoit S., Shaffer J., and Boldrick J. (2003): Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science*, Vol. 18, N. 1, pp. 71-103.

Gao, X., Starmer J., and Martin E.R. (2008): A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*, Vol. 32, N. 4, pp. 361-9.

Han B., Kang H.M., and Eskin E. (2009): Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, Vol. 5, N. 4, p. e1000456.

Westfall P.H., and Young S.S. (1993): *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*, John Wiley & Sons, New York.

Asymptotic Results for Randomly Reinforced Urn Models and their Application to Adaptive Designs

Irene Crimaldi
IMT Institute for Advanced Studies Lucca
irene.crimaldi@imtlucca.it

Urn models are a very popular topic because of their applications in various fields: sequential clinical trials, biology, economics and computer science. A large number of "replacement policies" has been considered and studied by many authors, from different points of view and by means of different methods. We will focus on some central limit theorems and some related statistical tools.

The main central limit theorem will be stated for an arbitrary sequence of real random variables. Indeed, although this result has been thought for urn problems, it deals with the general problem of the rate of convergence of predictive distributions and empirical distributions for dependent data.

Keywords: adaptive design, central limit theorem, empirical and predictive distribution, stable convergence, urn model.

Acknowledgements: This talk is based on some works which were partially supported by MIUR (PRIN 2008), by INdAM (GNAMPA 2009) and by CNR PNR project "CRISIS Lab".

References

Caldarelli G., Chessa A., Crimaldi I., Pammolli F. (2012): Weighted Networks as Randomly Reinforced Urn Processes, submitted to *Physical Review E*.

Berti P., Crimaldi I., Pratelli L., Rigo P. (2011): A central limit theorem and its applications to multicolor randomly reinforced urns, *J. Appl. Probab.*, Vol. 48, N. 2, pp. 527-546.

Bassetti F., Crimaldi I., Leisen F. (2010): Conditionally identically distributed species sampling sequences, *Adv. Appl. Probab.*, Vol. 42, N. 2, pp. 433-459.

Berti P., Crimaldi I., Pratelli L., Rigo P. (2010): Central limit theorems for multicolor urns with dominated colors, *Stochastic Processes and their Applications*, Vol. 120, N. 8, pp. 1473-1491.

Berti P., Crimaldi I., Pratelli L., Rigo P. (2009): Rate of convergence of predictive distributions for dependent data, *Bernoulli*, Vol. 15, N. 4, pp. 1351-1367.

Crimaldi I. (2009): An almost sure conditional convergence result and an application to a generalized Pólya urn, *Internat. Math. F.*, Vol. 4, N. 23, pp. 1139-1156.

Crimaldi I., Leisen F. (2008): Asymptotic results for a generalized Pólya urn with "multi-updating" and applications to clinical trials, *Communications in Statistics - Theory and Methods*, Vol. 37, N. 17, pp. 2777-2794.

Crimaldi I., Letta G., Pratelli L. (2007): A strong form of stable convergence. In Donati-Martin C., Émery M., Rouault A., Stricker C. (Eds.) *Séminaire de Probabilités XL*, Lecture Notes in Mathematics, Vol. 1899, Springer Verlag, pp. 203-225.

Comparison for alternative imputation methods for ordinal data

Federica Cugnata
DEMM, University of Milan
federica.cugnata@unimi.it

Silvia Salini
DEMM, University of Milan
silvia.salini@unimi.it

Most of the literature on missing data has focused on the case of quantitative data. Less attention has been devoted to the treatment of missing imputation methods for ordinal data, although ordinal variables occur in many fields. Existing methods are generally an adaptation of techniques originally designed for quantitative variables (Ferrari et al. 2011).
In this paper we propose to use the CUB models to impute missing data in presence of ordinal variables. In CUB models, answers of ordinal response items of a questionnaire are interpreted as the result of a cognitive process, where the judgement is intrinsically continuous but it is expressed in a discrete way within a prefixed scale of m categories. The rationale of this approach stems from the interpretation of the final choices of respondents as result of two components, a personal *feeling* and some intrinsic *uncertainty* in choosing the ordinal value of the response (Iannario and Piccolo, 2012). The first component is expressed by a shifted Binomial random variable. The second component is expressed by a Uniform random variable. The two components are linearly combined in a mixture distribution. The acronym CUB stands for a Combination of Uniform and (shifted) Binomial random variables.
If CUB is the real model that have generated the ordinal data observed, it is justified our approach to approximate unobserved values assuming the same model.

The first step will be to consider a benchmarking dataset and to compare our approach to some general methods of missing imputation (Mattei et al. 2012). The second step consists in running a simulation study designed by varying the `CUB` parameters in which `CUB` as well as other methods of multiple imputation are considered and compared. Finally a real dataset of customer satisfaction surveys coming from the airline industry will be considered. A large number of respondents are available, near 40000, but if one would consider only complete-case analysis (CCA) or available-case analysis (ACA) the number of units drastically decrease. This is what often happens in customer satisfaction surveys. Imputation of missing data is almost always necessary in this context.

References

Iannario M. and Piccolo D. (2012). *CUB Models: Statistical Methods and Empirical Evidence*, in Modern Analysis of Customer Satisfaction Surveys, Kenett R.S. and Salini S. Editors, John Wiley and Sons, Chichester: UK.

Ferrari P., Annoni P., Barbiero A. and Manzi G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics and Data Analysis*, 55:2410-2420.

Mattei A., Mealli F. and Rubin D.B. (2012). *Missing data and imputation methods*, in Modern Analysis of Customer Satisfaction Surveys, Kenett R.S. and Salini S. Editors, John Wiley and Sons, Chichester: UK.

Real time detection of trend-cycle turning points

Estela Bee Dagum
Department of Statistics, University of Bologna
estela.beedagum@unibo.it

Silvia Bianconcini
Department of Statistics, University of Bologna
silvia.bianconcini@unibo.it

A common feature of industrialized economies is that economic activity moves between periods of expansion, in which there is broad economic growth, and periods of recession, in which there is broad economic contraction. Understanding these phases has been the focus of much macroeconomic research over the past century, and particularly the identification and prediction of turning points. Chronologies of business cycle turning points are currently maintained in the United States by the National Bureau of Economic Research (NBER), and in Europe by the Centre for Economic Policy Research (CEPR). The identification of a new turning point in the economy requires additional data, which is only available after the turning point is reached. As a consequence, these turning points can only be identified with a time lag.

For real time analysis, official statistical agencies analyze final trend-cycle estimates, generally derived using asymmetric moving average techniques. But the use of these nonparametric asymmetric filters introduces revisions as new observations are added to the series, and delays in detecting true turning points. In this paper, we consider a reproducing kernel representation of commonly applied nonparametric trend-cycle predictors (Dagum and Bianconcini, 2008 and 2013) to derive asymmetric filters that monotonically converge to the corresponding symmetric one. We consider three specific criteria of bandwidth selection, namely:

1. minimization of the transfer function which implies an optimal

compromise between reducing revisions and phase shift;

2. minimization of the gain function which implies revisions reduction, and

3. minimization of the phase shift function which implies reduction of the time lag to detect a true turning point.

Hence, we can obtain a family of real time trend-cycle predictors with the important properties of either minimization of revision, or fast detection of turning points or an optimal compromise between these two. The behavior of the proposed procedure is illustrated with real series mainly used to identify and predict true macroeconomic turning points.

Keywords: Asymmetric filters, reproducing kernel Hilbert space, bandwidth selection, transfer function.

References

Dagum, E.B. and Bianconcini, S. (2008), The Henderson Smoother in Reproducing Kernel Hilbert Space, *Journal of Business and Economic Statistics*, 26 (4), 536-545.

Dagum, E.B. and Bianconcini, S. (2013), A unified probabilistic view of nonparametric predictors via reproducing kernel Hilbert spaces, *Econometric Reviews*, forthcoming.

Henderson, R. (1916), Note on Graduation by Adjusted Average, *Transaction of Actuarial Society of America*, 17, 43-48.

Musgrave, J. (1964), *A set of end weights to end all end weights*, Working paper. Washington DC: Bureau of Census.

Joint prior distributions for variance components in Bayesian analysis of normal hierarchical models

Haydar Demirhan
Department of Statistics, Hacettepe University, Türkiye
demirhanhaydar@hotmail.com

Zeynep Kalaylioglu
Department of Statistics, METU, Türkiye
kzeynep@metu.edu.tr

With the advances in computational techniques and increase in computing power, more emphasis is placed on simulation assessment of the performances of statistical methods and practical use of simulation based techniques such as Markov chain Monte Carlo methods has flourished. This has enabled developments in Bayesian methods to analyze various complex models such as multilevel models where the variance of a response variable has multiple components. As there is usually no sufficient prior knowledge regarding the components of the variance, researchers have been after determining a diffuse reference prior for the variance components. So far the priors in the literature assumed a priori independence between the variance components, e.g. Lambert et al. (2005), Browne and Draper (2006), Gelman (2006), Polson and Scott (2012). Motivated by the facts that i. the variance components are intrinsically linked, and ii. they have interactive effect on the parameter draws in Gibbs sampling, we model the variance components a priori jointly in a multivariate fashion paying special attention to generalized multivariate log-gamma distribution (G-MVLG) (Demirhan and Hamurkaroglu (2011)). We use a random coefficient normal hierarchical model, an important class of multilevel models, to illustrate our approach. Extensive Monte Carlo simulation study is conducted to assess and compare bias and efficiency of the Bayesian estimates and their sensitivity to vari-

ance hyperparameters under independent and joint variance priors. Our simulation results show that multivariate modeling of variance components in a multilevel model leads to i. better estimation properties for the response and random coefficient model parameters and ii. posterior random coefficient estimates that are insensitive to variance hyperparameters. The contribution of simulation in our study is threefold: 1. bias and efficiency properties are obtained by Monte Carlo simulation, ii. a simulation based technique utilizing directional derivatives is developed to investigate the sensitivity of posterior outcome, iii. Gibbs sampling is used for posterior distributions.

Keywords: multilevel models, random coefficient, variance components, directional derivative, sensitivity.

References

Browne W.J., Draper D. (2006): A comparison of Bayesian and likelihood-based methods for fitting multilevel models, *Bayesian Analysis*, Vol. 1, N. 3, pp. 473-514.

Demirhan H., Hamurkaroglu (2011): On a multivariate log-gamma distribution and the use of the distribution in the Bayesian analysis. *Journal of Statistical Planning and Inference*, Vol. 141, N.3, pp.1141-1152.

Gelman A. (2006): Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, Vol. 1, N. 3, pp. 515-533.

Lambert P.C., Sutton A.J., Burton P.R., Abrams K.R., Jones A.R. (2005): How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS, *Statistics in Medicine*, Vol. 24, N. 15, pp. 2401-2428.

Polson N.G., Scott J.G. (2012): On the half-cauchy prior for a global scale parameter, *Bayesian Analysis*, Vol. 7, N. 4, pp. 887-902.

Challenges in random variate generation

Luc Devroye
McGill University, Montreal, Canada
lucdevroye@gmail.com

At the heart of most simulations are algorithms for generating random variables. Even today, there are distributions for which we can only simulate in an approximative manner, if we can simulate from them at all. Much progress has been made over the past two decades by refining the rejection method on the one hand, and by developing novel methods such as coupling from the past (CFTP) on the other hand. Still, for indirectly specified distributions such as probability laws that are solutions of so-called distributional identities, or distributions given by Levy-Khinchine measures, challenges abound. The talk surveys some of the work that remains to be done.

A copula-based approach for discovering inter-cluster dependence relationships

F. Marta L. Di Lascio
School of Economics and Management, Free University of Bozen-Bolzano, Italy
marta.dilascio@unibz.it

Simone Giannerini
Department of Statistical Sciences, University of Bologna, Italy
simone.giannerini@unibo.it

In this work we focus on clustering dependent data according to the multivariate structure of the generating process (GP hereafter). Di Lascio and Giannerini (2012) proposed an algorithm based on copula function (Sklar, 1959), called CoClust, that accomplishes this task. The CoClust is based on the assumption that the clustering is generated by a multivariate probability model. In particular, each cluster is represented by a (marginal) univariate density function while the whole clustering is modeled through a joint density function defined via copula whose dimension is equal to the number of clusters. Therefore, the interest is on the inter-cluster dependence relationship rather than on the intra-cluster relationship: observations in different clusters are dependent while observations in the same cluster are independent. The CoClust showed a very good performance in many different scenarios. However, it has some drawbacks: in particular, it allocates all the observations and it has an high computational burden. In this work we propose a modified version of the CoClust that overcomes these problems. Specifically, the new algorithm $i)$ is able to discard observations irrelevant to the clustering, $ii)$ is able to recognize different dependence structures within a single data set and $iii)$ is computationally feasible. The algorithm selects the observations candidate to allocation through pairwise dependence measures and classifies them through a criterion based on the loglikelihood of a copula fit. Our approach does not require either to choose

a starting classification or to set a priori the number of clusters. Also, it uses a semi-parametric approach in which the margins are estimated through the empirical distribution function while the copula model is estimated by using the loglikelihood function. We have tested the algorithm in a large simulation study where we vary the kind of data GP, copula model, kind of margins, sample size and number of clusters. Furthermore, the new features of the algorithm are assessed and compared to the original CoClust; specifically, we test its capability of identifying different clustering structures, i.e. data coming from different data GPs, as well as of distinguishing dependent from independent data in a given dataset. Finally, we provide some examples on real data.

Keywords: CoClust algorithm, Copula function, Dependence structure.

Acknowledgements: The first author acknowledges the support of the School of Economics and Management, Free University of Bozen-Bolzano, Italy.

References

Di Lascio F.M.L., Giannerini S., (2012): A Copula-Based Algorithm for Discovering Patterns of Dependent Observations, *Journal of Classification*, Vol. 29, N. 1, pp. 50-75.

Di Lascio F.M.L., Giannerini S., (2013): The CoClust algorithm to discover gene expression relationships. Working paper.

Sklar A., (1959): Fonctions de répartition à n dimensions et leures marges, *Publications de l'Institut de Statistique de L'Université de Paris*, Vol. 8, pp. 229-231.

Parallel Monte Carlo method for American option pricing

A.V. Dmitriev, S.M. Ermakov
Saint Petersburg State University
alx.dmitriev@gmail.com, sergej.ermakov@gmail.com

Parallel version of the Monte Carlo method is considered for the American option pricing problem. In general the price of American option can be found from the Black-Scholes equation with a moving boundary (cf. e.g. Duffy, 2006). Analytical solutions of the equation are seldom available and hence such derivatives must be priced by numerical technics. Penalty method (Zvan et al., 1998), (Nielsen et al., 2002) allows to remove the difficulties associated with a moving boundary. Then the problem can be solved numerically on a fixed domain.

The finite difference method allows to bring the problem to solving system of equations where unknown variables are function values in lattice nodes. In the case of the American option problem this system is the system of nonlinear equations. After linearization the problem of finding the American option price is reduced to sequential solving of linear equations.

The Monte Carlo algorithm which belongs to class of parametrically splitting algorithms (Ermakov, 2010) is proposed for solving derived problem. In terms of parametrically splitting algorithms a parametric set here is the set of random numbers which are used in calculations of Monte Carlo estimators and which should be split among processors. In order to substantiate the method it is necessary to investigate the question of stochastic stability of the Monte Carlo algorithm. Stochastic stability is understood as boundedness of the correlation matrix of the component estimators when number of time steps increases. Sufficient conditions for stochastic stability of the algorithm were obtained. Numerical experiments which demonstrate the possibility of effective parallelization of the algorithm calculations were performed for different

values of parameters.

Keywords: Monte Carlo methods, statistical modeling, American options, penalty method.

Acknowledgements: This work was supported by the RFBR grant N^{o} 11-01-00769a

References

Duffy D., 2006: *Finite difference methods in financial engineering: a partial differential equation approach.* John Wiley & Sons Ltd. 423 p.

Zvan R., Forsyth P.A., Vetzal K.R., 1998: Penalty methods for American options with stochastic volatility. *Journal of Computational and Applied Mathematics* no.218, p. 91-199.

Nielson B.F, Skavhaug O., Tvelto A., 2002: Penalty and front-fixing methods for the numerical solution of American option problems. *J. Comp. Finan.* no. 4.

Ermakov S.M., 2010: Parametrically splitting algorithms. *Vestnik St. Petersburg University: Mathematics*, Volume 43, Issue 4, Allerton Press, Inc. pp 211-216.

Two-stage optimal designs in nonlinear mixed effect models: application to pharmacokinetics in children

Cyrielle Dumont
UMR 738, INSERM, University Paris Diderot, Paris, France; Department of Clinical Pharmacokinetics, Institut de Recherches Internationales Servier, Suresnes, France
cyrielle.dumont@inserm.fr

Marylore Chenel, France Mentré
Department of Clinical Pharmacokinetics, Institut de Recherches Internationales Servier, Suresnes, France and UMR 738, INSERM, University Paris Diderot, Paris, France
marylore.chenel@fr.netgrs.com, france.mentre@inserm.fr

Nonlinear mixed effect models (NLMEM) are used in population pharmacokinetics (PK) to analyse concentrations of patients during drug development, particularly for pediatric studies. Approaches based on the Fisher information matrix (M_F) can be used to optimise their design (Mentré, 2001; Tod, 2008). A first-order linearization of the model was proposed to evaluate M_F for these models (Mentré, 1997) and is implemented in the R function PFIM (Bazzoli, 2010). Local optimal design needs some *a priori* values of the parameters which might be difficult to guess. Therefore adaptive designs, among which two-stage designs, are useful to provide some flexibility and were applied in pharmacometrics (Foo, 2012; Zamuner, 2010). However, articles in other contexts (Federov, 2012) discussed that two-stage designs could be more efficient than fully adaptive designs. Moreover, two-stage designs are easier to implement in clinical practice. We implemented in a working version of PFIM the optimisation of the determinant of M_F for two-stage designs in NLMEM. We evaluated, with a simulation approach, for a small group of children, the impact of one-stage and two-stage designs on the precision of parameter estimation when the 'true' PK parameters are different than the *a priori* ones.

Keywords: Adaptive design, Design optimisation, Nonlinear mixed effect models, PFIM, Population pharmacokinetics.

References

Mentré F., Dubruc C., Thénot J.P. (2001): Population pharmacokinetic analysis and optimization of the experimental design for Mizolastine solution in children, *Journal of Pharmacokinetics and Pharmacodynamics*, Vol. 28, N. 3, pp. 299-319.

Tod M., Jullien V., Pons G. (2008): Facilitation of drug evaluation in children by population methods and modelling, *Clinical Pharmacokinetics*, Vol. 47, N. 4, pp. 231-243.

Mentré F., Mallet A., Baccar D. (1997): Optimal design in random-effects regression models, *Biometrika*, Vol. 84, N. 2, pp. 429-442.

Bazzoli C., Retout S., Mentré F. (2010) Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0, *Computer Methods and Programs in Biomedicine*, Vol. 98, N. 1, pp. 55-65.

Foo L.K., Duffull S. (2012): Adaptive optimal design for bridging studies with an application to population pharmacokinetic studies, *Pharmaceutical Research*, Vol. 29, N. 6, pp. 1530-1543.

Zamuner S., Di Iorio V.L., Nyberg J., Gunn R.N., Cunningham V.J., Gomeni R., Hooker A.C. (2010): Adaptive-Optimal Design in PET Occupancy Studies, *Clinical Pharmacology & Therapeutics*, Vol. 87, N. 5, pp. 563-571.

Federov V., Wu Y., Zhang R. (2012): Optimal dose-finding designs with correlated continuous and discrete responses, *Statistics in Medicine*, Vol. 31, N. 3, pp. 217-234.

Invariant dependence structures

Fabrizio Durante
School of Economics and Management
Free University of Bozen-Bolzano, Italy
fabrizio.durante@unibz.it

In general it is quite difficult to find an explicit model to real data. Therefore one is often interested in asymptotic behaviour and theorems in order to approximate the true model by some limiting results (e.g., central limit theorem, large deviations, etc.).

In univariate extreme-value theory, for instance, it is well known that, under suitable assumptions, and for a sufficiently large threshold u, the conditional excess distribution function of a random variable X can be approximated by a Generalized Pareto Distribution (a result known as the "Pickands–Balkema–de Haan Theorem"). In the multivariate case, investigations along the same lines have analysed the asymptotic behaviour of the multivariate distribution function $F_{\mathbf{u}}$ of the random vector $(X_1, \ldots, X_d)$ given that $X_1 > u_1, \ldots, X_d > u_d$ for sufficiently large $\mathbf{u} \in \mathbb{R}$. In view of Sklar's Theorem, such a conditional distribution $F_{\mathbf{u}}$ can be described in terms of:

- the asymptotic behaviour of the marginals,
- the asymptotic behaviour of the dependence structure, i.e. the behaviour of its copula.

Motivated by these investigations, here we aim at considering copulas that are invariant under univariate truncation (with respect to the i–th coordinate), i.e. those copulas C such that, if a copula C is associated with a random vector $\mathbf{X}$ and $u \in \mathbb{R}$ is a given threshold, then C is also the copula of $[\mathbf{X} \mid X_i > u]$. In fact, such copulas can be used in the approximation of the dependence structure of the conditional distribution function of $[\mathbf{X} \mid X_i > u]$ for sufficiently large $u > 0$.

In particular, we present the family of bivariate copulas that are invariant under univariate truncation. This family can capture non–exchangeable dependence structures and can be easily simulated. Moreover, it presents strong probabilistic similarities with the class of Archimedean copulas from a theoretical and practical point of view. Related inference methods will be also discussed.

Finally, by using a result by Jaworski (2013), it is showed how such family can be used in the characterization of high-dimensional copulas that are invariant under univariate contagion.

Possible applications to problems arising in survival analysis and financial contagion are also addressed.

Keywords: Copula, Dependence, Random Generation.

Acknowledgements: This work was supported by School of Economics and Management, Free University of Bozen–Bolzano, via the project "Risk and Dependence".

References

Durante F., Jaworski P. (2012): Invariant dependence structure under univariate truncation, *Statistics*, Vol. 46 , pp. 263–267.

Durante F., Jaworski P., Mesiar R. (2011): Invariant dependence structures and Archimedean copulas, *Statistics & Probability Letters*, Vol. 81, pp. 1995–2003.

Jaworski P. (2013): Invariant dependence structure under univariate truncation: the high-dimensional case, *Statistics*, in press.

Spatial sampling design in the presence of sampling errors

Evangelos Evangelou
University of Bath
e.evangelou@bath.ac.uk

Zhengyuan Zhu
Iowa State University
zhuz@iastate.edu

A sampling design scheme for spatial models for the prediction of the underlying Gaussian random field will be presented. The choice of the sampled (gauged) sites is put into an optimal design of experiments context, i.e. the sites are selected to optimise a design criterion from all sites within the region of interest. Caselton and Zidek (1984) and Shewry and Wynn (1984) are among the first to apply Lindley's information measure (Lindley, 1956) as a criterion for spatial sampling. In this talk we assume that observations are sampled with error, while interest lies in predicting the random field *without* the error term. Furthermore, the error variance is allowed to be different at each location (cf. Stein, 1995).

In order to reduce the uncertainty in prediction, every location is sampled more than once. On the other hand, obtaining the optimal design in this case is computationally harder. To that end, we present a hybrid algorithm by combining simulated annealing nested within an exchange algorithm.

Computational studies are presented, which show that under some circumstances, non-symmetric designs are superior to symmetric ones. Evidently, sampling error should be accounted for in the design of a sampling scheme.

Keywords: Geostatistics; information; measurement error; sampling design.

References

Caselton, W.F. and Zidek, J.V. (1984): Optimal monitoring network designs, *Statistics & Probability Letters*, Vol. 2, N. 4, pp. 223-227.

Lindley, D.V. (1956): On a measure of the information provided by an experiment, *The Annals of Mathematical Statistics*, Vol. 27, N. 4, pp. 986-1005.

Shewry, M.C. and Wynn, H.P. (1987): Maximum entropy sampling, *Journal of Applied Statistics*, Vol. 14, N. 2, pp. 165-170.

Stein, M.L. (1995): Locally lattice sampling designs for isotropic random fields, *The Annals of Statistics*, Vol. 23, N. 6, pp. 1991-2012.

An algorithm to simulate VMA processes having a spectrum with fixed condition number

Matteo Farné
University of Bologna
matteo.farne2@unibo.it

This paper proposes a method to simulate multivariate time series with suitable properties for the computation of spectral estimates and the assessment of their performances. In particular, this method was developed to obtain p-dimensional time series with a variable spectrum ($|f| \leq 1/2$) having fixed condition number c (intended as the ratio between the maximum eigenvalue λ_p and the minimum eigenvalue λ_1) across frequencies.

First of all, a procedure to generate $p-$dimensional diagonal matrices with trace 1 having fixed condition number is proposed. This goal is achieved through the solution of a linear equation system setting the ratio $\frac{\lambda_p}{\lambda_1}$ to c and imposing the equidistance between conseutive diagonal elements. Using the obtained matrices as covariance matrices for a zero-mean multivariate normal distribution, it is possible to obtain multivariate normal time series with a constant spectrum and fixed condition number across frequencies. Setting different values for the condition number c, it is possible to control the spread among the variances of the different component series.

Then, the described method is extended in order to generate VMA time series of any order having fixed condition number across frequencies. Specifying scalar matrices as the coefficients of the VMA process at any lag, we can compute the transfer function matrix and thus the multivariate spectrum using the diagonal matrices obtained at the previous stage. The condition number c controls the spread among the variances and the spectra of the component series, while the parameters

controlling the scalar matrices rule the dynamics of the spectrum across frequencies.

To conclude, a detailed analysis of the main features of the generated series is conducted with the help of specific examples, specific error measurements are defined to quantify their distance to the target of the generated series. Finally, a brief discussion about the properties of the present method with reference to Kolmorogov asymptotic estimation framework is provided, assessing the performance of a new multivariate spectral estimator.

Keywords: multivariate time series, spectral analysis, well-conditioning.

References

Boehm, H., von Sachs, R. (2009). Shrinkage estimation in the frequency domain of multivariate time series. J. of Multivariate Analysis 100:913-935.

Ledoit O., Wolf. M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88:365-411

Priestley, M.B (1981), *Spectral analysis and time series*, Vol.II, Chap. 9, Academic Press, London

Complex areal sampling strategies for estimating forest cover and deforestation at large scale

Lorenzo Fattorini
University of Siena
lorenzo.fattorini@unisi.it

Maria Chiara Pagliarella
University of Siena
mc.pagliarella@unisi.it

A post-Kyoto protocol, the Reduction of Emissions from Deforestation and forest Degradation (REDD) project was proposed and initiated in 2005. In this framework the monitoring of forest cover at large scale by statistically sound methodologies is a key pre-requisite. Forest cover is usually estimated at large scale by spatial sampling strategies, in which the study region is partitioned into N polygons of equal size (e.g. quadrats). Then, a sample of n units is selected, aerial photos of the sampled units are provided and visually interpreted to determine the forest cover within. Usually, the scheme adopted to select units are familiar schemes such as stratified sampling or cluster sampling. The purpose of this paper is to investigate the use of a complex spatial schemes proposed by Fattorini (2006) to account for the presence of spatial autocorrelation among the units. Indeed, adjacent units are often more alike than units that are far apart, thus giving a poor contribution to the sample information. Fattorini (2006) suggest modifying the simple random sampling without replacement in such a way that, at each drawing, the probabilities of selecting those units that are adjacent to the previously selected ones are reduced or increased according to a prefixed factor $\beta \geq 0$. Thus, at the first drawing, each unit has the same probability $\tau_1(j) = 1/N (1, \ldots, N)$ of being selected. Subsequently, conditional on the first $i-1$ selected units $j_i, \ldots, j_{i-1}$, the remaining $N - i + 1$

units are selected at drawing $i (i = 2, \ldots, n)$ with probability

$$\tau_i(j|j_i, \ldots, j_{i-1}) = \begin{cases} \frac{\beta}{N-i+1+(\beta-1)N(C)} & \text{if} \quad j \in C, \\ \frac{1}{N-i+1+(\beta-1)N(C)} & \text{otherwise} \end{cases} \tag{1}$$

where C denotes the set of units which are contiguous to at least one of the $i-1$ units $j_i, \ldots, j_{i-1}$ and $N(C)$ denotes the cardinality of the set. While the sampling scheme is easy to implement, the Horvitz-Thompson (HT) estimator is inapplicable even for moderate values of N and n, because the computation of inclusion probabilities involves enumerating all the possible samples and all the orderings in which the units enter the sample. Accordingly, the inclusion probabilities may be estimated by simulation using an appropriate number of replications of the sampling scheme as suggested by Fattorini (2006, 2009). As the number of simulated samples increases, the empirical HT estimator has the same performance of the HT estimator which would be obtained using the true inclusion probabilities. In order to check the validity of this strategy, a simulation study was performed. Negligible bias is involved using the estimated inclusion probabilities instead of the true ones, along with gains in efficiency relative to the use of familiar designs.

Keywords: forest monitoring, spatial sampling, Horvitz-Thompson estimation, inclusion probabilities, simulation.

References

Fattorini L. (2006): Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, Vol. 93, N. 2, pp. 269-278.

Fattorini L. (2009): An adaptive algorithm for estimating inclusion probabilities and performing the Horvitz-Thompson criterion in complex designs. *Computational Statistics*, Vol. 24, N. 4, pp. 623-639.

Double-barrier first-passage times of jump-diffusion processes

Lexuri Fernández
Ekonomi Analisiaren Oinarriak II Saila,
University of the Basque Country UPV/EHU,
lexuri.fernandez@ehu.es

Peter Hieber, Matthias Scherer
Lehrstuhl für Finanzmathematik (M13),
Technische Universität München,
hieber@tum.de, scherer@tum.de

Required in a wide range of applications first-passage time problems have attracted considerable interest over the past decades. Since analytical solutions often do not exist, one strand of research focuses on fast and accurate numerical techniques. Some authors rely on Monte-Carlo simulations. However, the standard Monte-Carlo simulation on a discrete grid exhibits two disadvantages: first, even for 1000 discretization intervals per unit of time, we obtain a significant discretization bias. Second, computation time increases rapidly if one has to simulate on a fine grid. Several authors focused on unbiased simulation schemes. Metwally and Atiya (2002) provide an unbiased, fast, and accurate alternative based on the so-called "Brownian bridge technique". This simulation technique has various applications in finance (see, e.g., Ruf and Scherer (2011), Hieber and Scherer (2010), Henriksen (2011)).

We show how the standard Brownian bridge technique can be adapted to a large variety of exotic double barrier products. Those products are very flexible and thus allow investors to adapt to their specific hedging needs or speculative views. However, those contracts can hardly be traded if there is no fast and reliable pricing technique. To provide this flexibility for the Brownian bridge technique, we extend the existing algorithms and (1) allow to price double barrier derivatives that trigger different events depending on which barrier was hit first and (2) allow to

evaluate payoff streams that depend on the first-passage time. Furthermore, (3) we show that time dependent barriers can easily be treated. Finally, we discuss the implementation and show that – in contrast to most alternative techniques – the Brownian bridge algorithms are easy to understand and implement.

Keywords: Double-barrier problem, first-exit time, first-passage time, Brownian bridge, barrier options.

Acknowledgements: Lexuri Fernández is supported by a FPI-UPV/EHU grant. Peter Hieber acknowledges funding by the TUM graduate school.

References

Henriksen, P. N. (2011): Pricing barrier options by a regime switching model, *Journal of Quantitative Finance*, Vol. 11, No. 8, pp. 1221- 1231.

Hieber, P. and Scherer, M. (2010). Efficiently pricing barrier options in a Markov-switching framework, *Journal of Computational and Applied Mathematics* , Vol. 235, pp. 679-685.

Metwally, S. and Atiya, A. (2002): Using Brownian bridge for fast simulation of jump-diffusion processes and barrier options, *Journal of Derivatives*, Vol. 10, pp. 43-54.

Ruf, J. and Scherer, M. (2011): Pricing corporate bonds in an arbitrary jump-diffusion model based on an improved Brownian-bridge algorithm, *Journal of Computational Finance*, Vol. 14, No. 3, pp. 127- 145.

Laws of large numbers for random variables with arbitrarily different and finite expectations via regression method

Silvano Fiorin
University of Padua
fiorin@stat.unipd.it

The convergence is studied for the sequence $\frac{1}{n}\sum_{i=1}^{n} Y_i$ when each Y_i is a real random variable and the expectations $E(Y_i)$ define a sequence of finite and possible different values; the usual technique of taking the differences $(Y_i - E(Y_i))$ is avoided because the convergence of $\frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i))$ to zero, in the general case, gives no informations about the asymptotic behaviour of $\frac{1}{n}\sum_{i=1}^{n} Y_i$. The adopted method is based on the possibility of embedding the probability distributions of each Y_i as a conditional probability distribution of a suitable pruduct type measure. Thus it is constructed a product space $\mathcal{X} \times \mathbb{R}^1$ with a product type probability measure where the *marginal* measure $P_{\mathcal{X}}$ is assigned on the Borel σ-field $\mathcal{B}_{\mathcal{X}}$ defined over the metric space $\mathcal{X}$ obtained by the closure of the set of the probability distribution functions $F_{Y_i}(y) = P(Y_i \leq y)$ for each r.v. Y_i. Moreover a class of conditional probability measures $P(x, B)$ is defined over the usual Borel σ-field $\mathcal{B}^1$ of $\mathbb{R}^1$ such that the below properties are satisfied:

i) $P(x, \cdot)$ is a probability measure over $\mathcal{B}^1$ having x as its probability distribution function, for each $x \in \mathcal{X}$

ii) $P(\cdot, B)$ is a $\mathcal{B}_{\mathcal{X}}$-measurable function for each fixed $B \in \mathcal{B}^1$

Thus, by the product measure theorem, the joint probability measure $P_{\mathcal{X}\times\mathbb{R}^1}$ can be derived over σ-field $\mathcal{B}_{\mathcal{X}} \times \mathcal{B}^1$ and the *marginal* random variable Y is considered i.e. the map $Y(x, y) = y, \forall (x, y) \in \mathcal{X} \times \mathbb{R}^1$. The expectation $E(Y) = \int_{\mathcal{X}} \left(\int_{\mathbb{R}^1} y dP(x, \cdot) \right) dP_{\mathcal{X}}(x)$ which, by Fubini theorem, can be written as integral of conditional expectations is

really a relevant element of our analysis. In fact a wide class of laws of large numbers is constructed such that the sequence $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is almost surely convergent to $E(Y)$. Furthermore, depending the possible limit $E(Y)$ on the marginal probability measure $P_{\mathcal{X}}$, the strict connection is investigated between $P_{\mathcal{X}}$ and the permutations (or rearrangements) for the r.v. Y_i's. In the literature the role played by rearrangements of Y_i's in the strong laws of large numbers was studied in Chobanyan *et al.* (2004) using the results on convergent rearrangements for Fourier series. Nevertheless the strategy and results here proposed differ consistently with respect to the analysis available in the literature. For instance the possibility of characterizing the limit as the expectation $E(Y)$ of a marginal r.v. Y is a property of the approach here suggested. Finally some comments about the assumptions here adopted; all the proofs are given in case of:

i) independent but not identically distributed r.v. Y_i's;

ii) pairwise uncorrelated r.v. Y_i's, using theorems 5.1.1 and 5.1.2 in Chung (2001) or alternatively theorems 3.1.1 and 3.1.2 in Chandra (2012).

Keywords: Law of large numbers, regression function, rearrangements of terms of a series, non-stationary processes.

References

Chandra T.K. (2012): *Laws of large numbers*, Narosa publishing house, New Delhi.

Chobanyan S., Levental S., Mandrekar V. (2004): Prokhorov blocks and strong law of large numbers under rearrangements, *Journal of Theoretical Probability*, Vol. 17, N. 3, pp. 647-672.

Chung T.K. (2001): *A course in probability theory*, Academic Presse, San Diego.

Algebraic characterization of saturated designs

Roberto Fontana
Politecnico di Torino
roberto.fontana@polito.it

Fabio Rapallo
Università del Piemonte Orientale
fabio.rapallo@unipmn.it

Maria Piera Rogantin
Università di Genova
rogantin@dima.unige.it

Saturated fractions play an important role in Design of Experiments. Given a model, saturated fractions are fractions of a factorial design with as much points as the number of parameters of the model. Using Algebraic Statistics, we characterize saturated fractions of a design in terms of the circuits of the design matrix and we define a criterion to actually check whether a given fraction is saturated or not. Algebraic Statistics is a discipline encompassing the application of Combinatorics and Polynomial algebra to Statistics. Its most prominent results concern essentially the analysis of contingency tables (Drton *et al*, 2009) and Design of Experiments (Pistone *et al*, 2001). Some connections between such two fields of applications are discussed in Fontana *et al* (2012).

Our approach is based on two main ingredients. First, we identify a factorial design with a contingency table whose entries are the indicator function of the fraction, i.e., they are equal to 1 for the fraction points and 0 otherwise. This implies that a fraction can also be considered as a subset of cells of the table. Second, we apply tools from Algebraic Statistics to characterize the saturated fractions. Most of the relevant combinatorial tools are summarized in Ohsugi (2012). The definition of circuits is the core of our algorithm, and it has already been considered in the framework of contingency tables in Kuhnt *et al* (2013) for the

definition of robust procedures for outliers detection, but limited to two-way tables.

From the point of view of computations, the circuits do not depend on the particular fraction to be analyzed. Thus, our theory is particularly useful in the context of simulated designs where one needs to generate several saturated fractions. Indeed the computation of the determinant of the design matrix is not needed to check design singularity.

Keywords: Circuits, Estimability, Linear models, Markov moves.

References

Drton M., Sturmfels B., Sullivant S. (2009): *Lecture on Algebraic Statistics*, Birkhauser, Basel.

Fontana R., Rapallo F., Rogantin M.P. (2012): Markov Bases for Sudoku Grids. In: Di Ciaccio A., Coli M., Angulo Ibanez J.M. (Eds.) *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics, Springer-Verlag, Berlin, pp. 305-315.

Kuhnt S., Rapallo F., Rehage A. (2013): Outlier Detection in Contingency Tables based on Minimal Patterns, *Statistics and Computing*, In press.

Ohsugi H. (2012): A Dictionary of Gröbner Bases of Toric Ideals. In: Hibi T. (Eds.) *Harmony of Gröbner Bases and the Modern Industrial Society*, World Scientific, Singapore, pp. 253-281.

Pistone G., Riccomagno E., Wynn H.P. (2001): *Algebraic Statistics. Computational Commutative Algebra in Statistics*, Chapman&Hall/CRC, Boca Raton.

Simulations and Computations of Weak Dependence Structures by using Copulas

Enrico Foscolo
Free University of Bozen-Bolzano, School of Economics and Management
enrico.foscolo@unibz.it

High-dimensional copulas have been recognized as a standard tool for constructing flexible multivariate models; e.g., see Joe (1997), Salvadori et al. (2007), Jaworski et al. (2010), and the references therein.

Recently, Durante et al. (2012) have provided a method to construct a class of n-dimensional copulas defined as follows

$$\tilde{\mathbf{C}}(\mathbf{u}) = \mathbf{D}(u_1, \ldots, u_{n-1})\, u_n + \mathbf{A}(u_1, \ldots, u_{n-1})\, f(u_n) \qquad (1)$$

for all $\mathbf{u} \in \mathbb{I}^n$. This class is obtained from a $(n-1)$-dimensional copula $\mathbf{D}$ and some suitable auxiliary functions $\mathbf{A}$ and f. Conditions under which functions of type (1) have been fully characterized in terms of properties of the auxiliary functions have been given. This class of copulas models weak dependence structures between random variables, and it generalizes various families of copulas, including Farlie-Gumbel-Morgenstern distribution and copulas with quadratic sections in one variable. As a further feature, they allow to describe a non-exchangeable behavior among the components of a random vector.

Here we focus on the manner to construct and to simulate 3-dimensional copulas of type (1) starting with an absolutely continuous 2-dimensional copula $\mathbf{D}$; shortly, $\tilde{\mathbf{C}}_3$. Specifically, we provide a standard way for defining an appropriate parametric function $\mathbf{A}\colon \mathbb{I}^2 \to \mathbb{R}$ such that the resulting function $\tilde{\mathbf{C}}_3$ is an absolutely continuous 3-dimensional copula for a wide class of absolutely continuous functions $f\colon \mathbb{I} \to \mathbb{R}$.

Since $\tilde{\mathbf{C}}_3$ has been proved to be absolutely continuous, the simulation procedure can be based on the conditional distribution method.

As an application, we show our methods at work both with weak dependent continuous and non-continuous random variables, and we un-

derlie some of its computational features.

Keywords: Copula, Weak Dependence, Sampling Algorithm, Farlie-Gumbel-Morgenstern distribution.

Acknowledgements: The author acknowledges the support of Free University of Bozen-Bolzano via the project *Handling High-Dimensional Systems in Economics*.

References

Durante F., Foscolo E., Rodríguez-Lallena J. A., Úbeda-Flores M. (2012): A Method for Constructing Higher-Dimensional Copulas, *Statistics*, Vol. 46, N. 3, pp. 387-404.

Jaworski P., Durante F., Härdle W., Rychlik T. (Eds.) (2010): *Copula Theory and its Applications*, Springer, Berlin.

Joe H. (1997): *Multivariate models and dependence concepts*, Chapman & Hall, London.

Salvadori G., De Michele C., Kottegoda N. T., Rosso R. (2007): *Extremes in Nature. An Approach Using Copulas*, Springer, Dordrecht.

Bayesian Random Item Effects Modeling: Analyzing Longitudinal Survey Data

Jean Paul Fox
Department of Research Methodology, Measurement, and Data Analysis
University of Twente, the Netherlands
g.j.a.fox@utwente.nl

To analyse complex clustered item response data, Bayesian hierarchical IRT models have been developed. Particularly well-known is the multilevel IRT (MLIRT) model (e.g., Fox, 2010; Fox and Glas, 2001), which defines a multilevel modelling structure on the person parameters. The MLIRT model has been extended to accommodate random effects item parameters to model jointly group differences in the person and item parameters (e.g, De Jong, Steenkamp and Fox, 2007; Fox, 2010; Verhagen and Fox, 2013). The group-specific item parameters are assumed to be normally distributed around the general item parameter for each item. In this double random effects structure, group means are defined around a common mean, which defines the group-specific parameters on a common scale. The model makes it possible to estimate one common latent scale for the person parameters across countries, taking variations in country-specific item parameters into account, which enables the meaningful comparison of individuals and countries.

It will be shown that this hierarchical IRT model is also suitable for analysing longitudinal or repeated measurements data. The model combines a longitudinal multilevel structure on the person parameter with random occasion-specific item parameters. The random item effects parameters are implemented to accommodate item characteristic changes over time that can occur in repeated measurement settings. Linear or non-linear time effects and time-varying covariates can be incorporated on different levels to explain growth in the person parameters or variations in item parameters. The model supports measurement time-invariance testing without the need for anchor items.

The comprehensive model is typically meant to analyze longitudinal survey data, where questionnaires are repeatedly used to measure changes in attitude, performance, or quality of life. Main interest is focused on studying individual latent growth given repeated measurements and categorical outcomes, without assuming time-invariant measurement characteristics. In a medical setting, examples of applications to longitudinal questionnaires will be given.

Keywords: Bayesian Statistical Methods, Item Response theory, MCMC, Time-variance item functioning.

References

De Jong, M.G., Steenkamp, J.B.E.M., Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260-278.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications.* New York: Springer.

Fox, J.-P., Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Verhagen, J. and Fox, J.-P. (2013). Bayesian Tests of Measurement Invariance. *British Journal of Mathematical and Statistical Psychology.* (online doi: 10.1111/j.2044-8317.2012.02059.x.).

Maximum Likelihood Based Sequential Designs for Logistic Binary Response Models

Fritjof Freise
Otto-von-Guericke University Magdeburg
fritjof.freise@ovgu.de

One of the problems optimal design in nonlinear models has to face is, that the information matrices and hence the resulting designs are depending on the actual value of the unknown parameter β. In this work sequential methods based on maximum likelihood estimation are considered.

In each step observations are taken, the parameter is estimated and new design points are determined using these estimates instead of the true value of the parameter. Even though the dependencies of the sequences of estimators and designs can be quite complicated, convergence results can be achieved using an approach motivated by an article of Ying and Wu (1997):

Given an initial design securing the existence of the maximum likelihood estimator, it is possible to derive an essentially recursive formulation for the sequence of the estimators. Convergence of the estimator is shown using methods from stochastic approximation literature.

The results are presented for the logistic binary response model and illustrated by simulations.

Keywords: maximum likelihood, sequential design, stochastic approximation.

References

Ying T., Wu C.F.J. (1997): An asymptotic theory of sequential designs based on maximum likelihood recursion, *Statistica Sinica*, Vol. 7, pp. 75-91.

Fixed design regression estimation based on real and artificial data

Dmytro Furer and Michael Kohler
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany
furer@mathematik.tu-darmstadt.de,
kohler@mathematik.tu-darmstadt.de

Adam Krzyżak
Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8
krzyzak@cs.concordia.ca

In this article we study fixed design regression estimation based on real and artificial data, where the artificial data comes from previously undertaken similar experiments. A least squares estimate is introduced which gives different weights to the real and the artificial data. It is investigated under which condition the rate of convergence of this estimate is better than the rate of convergence of an ordinary least squares estimate applied to the real data only. The results are illustrated using simulated and real data.

Keywords: Fixed design regression, nonparametric estimation, L_2 error, rate of convergence.

Acknowledgements: This work was supported by NSERC grant.

References

Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, no 3, pp. 1371–1385.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, no 1, pp. 71–82.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. 2nd edition, Marcel Dekker, New York.

Fromkorth, A. and Kohler, M. (2011). Analysis of least squares regression estimates in case of additional errors in the variables. *Journal of Statistical Planning and Inference*, **141**, issue 1, pp. 172-188.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer-Verlag, New York.

Kohler, M. (2006). Nonparametric regression with additional measurement errors in the dependent variable. *Journal of Statistical Planning and Inference*, **136**, issue 10, pp. 3339-3361.

Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, issue 7, pp. 3054–3058.

Manson, S. S. (1965). Fatigue: A complex subject - some simple approximation. *Experimental Mechanics*, **5**, no 7, pp. 193-226.

Analysis of a Finite Capacity M/G/1 Queue with Batch Arrivals and Threshold Overload Control

Gaidamaka Yu., Samuylov K., Sopin Ed.
Peoples' Friendship University of Russia
{ygaidamaka, ksam, esopin}@sci.pfu.edu.ru

Shorgin S.
Institute of Informatics Problems of RAS
sshorgin@ipiran.ru

The hysteretic load control mechanism (Gaidamaka 2012) is used to prevent overload in SIP-server networks (Gurbani 2012). According to figure 1 the system operates in normal (s=0), overload (s=1), and discard (s=2) modes depending on three queue length thresholds – onset threshold L, abatement threshold H and discard threshold R.

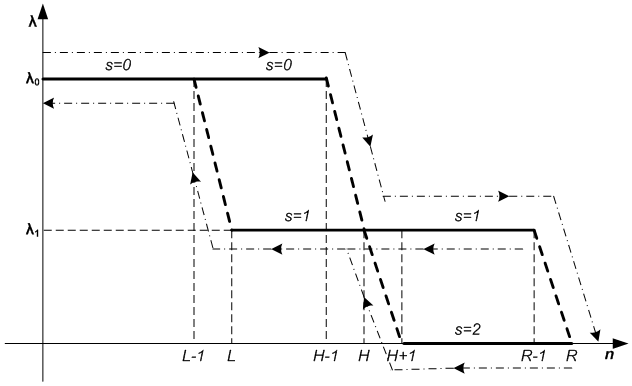

Figure 1. Hysteretic load control

Batch arrivals of SIP-messages should be taken into account, so we describe SIP-server operation in terms of the $M^{[X]}|G|1|\langle L,H\rangle|\langle H,R\rangle$

queuing system. Similar models without batch arrival were analyzed for example in (Roughan 2000, Sopin 2012). We obtained the system of equations for the steady state probability distribution and formulas for SIP-server quality of service parameters: the probabilities that the system is in overload and discard modes, the average control cycle time, and the average time spent in overload and discard modes.

Keywords: SIP-server, overload control, queuing system, batch arrival, finite queue, hysteretic control, threshold.

Acknowledgements: This work was supported in part by the Russian Foundation for Basic Research (grant 12-07-00108).

References

Gurbani V., Hilt V., Schulzrinne H. (2012) Session Initiation Protocol (SIP) Overload Control, *draft-ietf-soc-overload-control-10*.

Abaev P.O., Gaidamaka Y.V., Pechinkin A.V., Razumchik R.V., Shorgin S.Y. (2012): Simulation of overload control in SIP server networks, *Proceedings of the 26th European Conference on Modelling and Simulation, ECMS*, Germany, Koblenz, pp. 533-539.

Gaidamaka Y., Samouylov K., Sopin E. (2012): Analysis of M/G/1 queue with hysteretic load control, *XXX International Seminar on Stability Problems for Stochastic Models and VI International Workshop "Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems". Book of Abstracts*, pp. 87-90.

Roughan M., Pearce C.E.M. (2000): A martingale analysis of hysteretic overload control, *Advances in Performance Analysis*, N. 1, pp. 1-30.

A covariate-adjusted response adaptive design based on the Klein urn

Arkaitz Galbete, José Moler
Public University of Navarra
arkaitz.galbete@unavarra.es, jmoler@unavarra.es

Fernando Plo
University of Zaragoza
fplo@unizar.es

In a response-adaptive allocation rule, experimental units are allocated depending on previous allocations and responses. The main target is to take advantage of the information that the experiment is giving in order to reduce the number of experimental units allocated in inferior treatments, so that, in the context of clinical trials, less patients will be allocated in treatments with a bad performance.

A design that reduces predictability of future allocations while maintaining balance among prognostic factors was presented in Pocock and Simon (1975). This design is covariate-adaptive, but it does not use the previous responses of patients in the next allocation. In Atkinson (1982), previous responses are used for the next allocation. A linear regression model is proposed to explain the response of patients under classical assumptions of homocedasticity and incorrelation of errors. By using theory of optimal designs the procedure looks for the minimization of the variance of the updated OLS estimator.

There is a growing interest in covariate-adjusted response-adaptive (CARA) designs, were patients are allocated depending on the previous allocations, covariates and responses and, also, on the current patient's covariate. In Zhang et al. (2007) a general framework for CARA designs is presented and asymptotic results are obtained for this type of designs.

We explore the characteristics of a CARA design based in the Klein

urn, where two treatments are compared. The Klein urn design is an allocation rule that has been proved competitive with other good response-adaptive designs, see Galbete et al. (2013). It has the advantage that its stochastic structure is easy to handle, and therefore their performance characteristics can be fully studied analytically and not only via simulation. We also study the effects of the covariates in the response to treatments by means of adaptive-regression techniques.

Keywords: Randomization based inference, covariate-adjusted response-adaptive designs, clinical trials.

Acknowledgements: This work was partially supported by the project MTM2010-15972.

References

Atkinson, A. C. (1982): Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69, 1, 61–67.

Galbete, A., Moler, J. A. and Plo, F. (2013): A response-driven adaptive design based on the Klein urn. *Submitted to Methodology and Computing in Applied Probability*.

Pocok, S.J. and Simon, R. (1975): Sequential treatment assignment with balancing for prognostic factors. *Biometrics*, **31**, 103–115.

Zhang, L.X., Hu, F., Cheung, S. H. and Chan, W.S. (2007): Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, Vol 35, 3, 1166-1182.

Randomization-based inference (RBI) in clinical trials

Arkaitz Galbete, José A. Moler and Henar Urmeneta
Public University of Navarra
arkaitz.galbete@unavarra.es, jmoler@unavarra.es,
henar@unavarra.es

Fernando Plo
University of Zaragoza
fplo@unizar.es

Nowadays, in the context of clinical trials, the random allocation of subjects to treatments is firmly established. The main function of the randomization process is to avoid several sources of bias that can interfere the conclusions of the study.

When the probability distribution of the allocation rule depends on the previous allocations the randomization procedure is said adaptive and when it also depends on the previous responses, it is called response-adaptive. The appropriate design to randomize patients depends on the goals of the trial and there is not an optimal choice. For wide families of response-adaptive designs, some studies have been carried out in order to study the degree of compromise between ethical and inferential criteria, see, for instance, Hu and Rosenberger (2006).

The use of randomization based inference (RBI) instead of population models is a controversial issue for clinical trials, see, for instance, chapter 7 in Rosenberger and Lachin (2002). When the population model is not acceptable, RBI is a promising alternative. However, the particular randomization procedure used must be established in the preliminar planning of the trial and it will influence the final inferential conclusions , see, for instance, Cook and DeMets (2008).

In this work we focus on the use of RBI when a response-adaptive randomization scheme has been used to allocate patients. One of the main drawbacks of RBI is the computational cost of obtaining the exact

distribution of the test statistic. Here we present an algorithm that alleviate the difficulty to obtain exact p-values for some test-statistics. This algorithm is helpful when a small to moderate number of patients has been allocated with a response-adaptive design. We also obtain asymptotic properties of these test-statistics that are useful for large sample.

Keywords: Randomization based inference, response-adaptive designs, clinical trials.

Acknowledgements: This work was partially supported by the project MTM2010-15972.

References

Cook, T. D. and DeMets, D. L. (2008): *Introduction to Statistical Methods for Clinical Trials*, Chapman and Hall, New York.

Hu, F. and Rosenberger, W. F. (2006): The theory of response-adaptive randomization in clinical trials. Wiley, New York.

Rosenberger, W. F. and Lachin, J. M. (2002): Randomization in Clinical Trials. Theory and practice. Wiley, New York.

Measures of dependence for infinite variance distributions

Bernard GAREL
National Polytechnic Institute of Toulouse
garel@enseeiht.fr

In this paper, we introduce the signed symmetric covariation coefficient and give its main properties. Links with the generalized association parameter, put forward by Paulauskas and now called alpha-covariance, are addressed. We also propose estimation procedures for these two coefficients. Then the signed symmetric autocovariation function is defined for a stationary stable process. A characterization of the MA processes with this function is established.

Let X_1 and X_2 be jointly $S\alpha S$ and let $\mathbf{\Gamma}$ be the spectral measure of the random vector (X_1, X_2). The covariation of X_1 on X_2 is the real number defined by

$$[X_1, X_2]_\alpha = \int_{S_2} s_1 s_2^{\langle \alpha-1 \rangle} \mathbf{\Gamma}(d\mathbf{s}), \tag{1}$$

where for real numbers s and a: if $a \neq 0$, $s^{\langle a \rangle} = |s|^a \mathrm{sign}(s)$ and if $a = 0$, $s^{\langle a \rangle} = \mathrm{sign}(s)$. The covariation norm is defined by

$$\|X_1\|_\alpha = ([X_1, X_1]_\alpha)^{1/\alpha}. \tag{2}$$

Let (X_1, X_2) be a bivariate $S\alpha S$ random vector with $\alpha > 1$. The signed symmetric covariation coefficient between X_1 and X_2 is the quantity:

$$\mathrm{scov}(X_1, X_2) = \kappa_{(X_1,X_2)} \left| \frac{[X_1, X_2]_\alpha [X_2, X_1]_\alpha}{\|X_1\|_\alpha^\alpha \|X_2\|_\alpha^\alpha} \right|^{\frac{1}{2}}, \tag{3}$$

where $\kappa_{(X_1,X_2)}$ is a sign which depends on the different signs of the covariations. This coefficient exhibit desirable properties as does the ordinary Pearson correlation coefficient. More, it is easy to estimate. It coincides with the alpha-correlation introduced by Paulauskas in the sub-Gaussian case.

Keywords: Signed symmetric covariation coefficient, alpha-covariance, stable distributions, spectral measure.

References

Garel B., d'Estampes L., Tjostheim D. (2004): Revealing some unexpected dependence properties of linear combinations of stable random variables using symmetric covariation, *Communications in Statistics: Theory and Methods*, 33 (4), 769-786.

Garel B, Kodia B. (2009): Signed symmetric covariation for alpha-stable dependence modeling, *C.R. Acad. Sci. Paris* Ser I, 347-352.

Kodia B., Garel B. (2013): Estimation and comparison of signed symmetric covariation coefficient and generalized association parameter, *To appear in Communications in Statistics, Theory and Methods.*

An alternative for the computation of IMSE optimal designs of experiments

Bertrand GAUTHIER, Luc PRONZATO and João RENDAS
Université de Nice-Sophia Antipolis, I3S - UMR7271 - UNS CNRS
bgauthie@i3s.unice.fr, pronzato@i3s.unice.fr,
rendas@i3s.unice.fr

This work adresses the problem of designing experiments (*i.e.* choosing sampling points) in the framework of kernel-based interpolation models.

The integrated mean squared error (IMSE) criterion is a classical tool for evaluating the overall performance of interpolators (see for instance [Sacks et al., 1989]). For a fixed class of models and a given design size, it is therefore natural to try to choose sampling points such that the resulting interpolation minimises the IMSE-criterion among all possible samplings. In such case, one speaks about *IMSE-optimal design of experiments*.

However, in practice IMSE-optimal designs are relatively hard to compute. The evaluation of the IMSE criterion is indeed quite numerically expensive (it requires the computation of the integral of the mean-squared prediction error over the whole space) and the global optimization is often made difficult due to the presence of many local minima. The present work aims at investigating alternative ways to compute IMSE-optimal designs.

Let $\mathcal{X}$ be a measurable set and let μ be a σ-finite measure on $\mathcal{X}$. We denote by $L^2(\mathcal{X}, \mu)$ the Hilbert space of square-integrable (with respect to μ) real-valued functions on $\mathcal{X}$. We consider a real random field $(Z_x)_{x \in \mathcal{X}}$ indexed by $\mathcal{X}$. We assume that Z is centered, Gaussian, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with values in $L^2(\mathcal{X}, \mu)$.

The choice of the IMSE criterion for learning such a random field Z (that is, for the prediction of Z over the entire $\mathcal{X}$ from observations) naturally leads to the definition of an integral operator T on $L^2(\mathcal{X}, \mu)$,

with,

$$\forall f \in L^2(\mathcal{X}, \mu), \forall x \in \mathcal{X}, T[f](x) = \int_{\mathcal{X}} f(t)K(x,t)d\mu(t) \qquad (1)$$

(see for instance [Gauthier and Bay, 2012] for the interest of such operators when dealing with kernel-based interpolation models). The present work aims at exploiting the spectral decomposition of T for constructing IMSE-optimal designs.

More precisely, we describe an alternative approach to the determination of IMSE-optimal designs which does not require the explicit computation of the IMSE integral for each design tested (this therefore offers advantages in terms of numerical cost). We also study the impact of spectral truncations and quadrature rules on the IMSE criterion. Finally, we introduce a new *spectral criterion* for designing experiments for centered second-order random field models. Numerical examples illustrate the interest and behavior of our approach.

Keywords: random field models, kernel-based interpolation, IMSE-optimal designs, spectral decomposition of integral operators, spectral criterion.

Acknowledgements: This work was supported by the the project ANR-2011-IS01-001-01 DESIRE (DESIgns for spatial Random fiElds), joint with the Statistics Departement of the JKU Universität, Linz (Au).

References

Sacks, J. and Welch, W.J. and Mitchell, T.J. and Wynn, H.P. (1989): Design and analysis of computer experiments, *Statistical science*, Vol. 4, N. 4, pp. 409-423.

B. Gauthier and X. Bay (2012): Spectral approach for kernel-based interpolation, *Annales de la faculté des sciences de Toulouse*, Vol. 21, N. 3, pp. 439-479.

Design of Experiments using R

Albrecht Gebhardt
Institute of Statistics, University Klagenfurt, Austria
albrecht.gebhardt@uni-klu.ac.at

The programming language S has a long history and increased its popularity amongst statisticians espacially with the advent of its free software dialect R during the last decade. One key advantage of R is its extensibility by means of addon packages which resulted in more than 4000 packages available at present. Only a few of these packages are dedicated to design of experiments leaving several methods unimplemented.

An attempt to fill some of these gaps regarding optimal design is made in Rasch, Pilz, Verdooren, Gebhardt (2011) with its accompanying OPDOE library. While first versions of that library focussed on getting the implementation of the algorithms done a new version with a somewhat simplified interface and improvements will be presented.

The functions in the OPDOE library cover several topics of experimental design, including simple statistical tests, regression models, tests in analysis of variance models and sequential testing. The capabilities of the presented R library will be shown by a collection of examples covering these topics.

Keywords: design of experiments, linear model, analysis of variance, R language

References

D. Rasch, J. Pilz, L.R. Verdooren, A. Gebhardt (2011): *Optimal Experimental Design with R*, CRC Press, Boca Raton.

R Development Core Team (2009): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Hierarchical Fractional Factorial Designs for Model Identification and Discrimination

Subir Ghosh
University of California, Riverside, USA
subir.ghosh@ucr.edu

Two fractional factorial designs $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ with m factors and n_1 and n_2 runs, respectively are called "Hierarchical Designs" (HDs) if the runs in $\boldsymbol{T}_1$ are included in the runs of $\boldsymbol{T}_2$, $n_1 < n_2$. The design $\boldsymbol{T}_2$ is then obtained from the design $\boldsymbol{T}_1$ by augmenting to the n_1 runs in $\boldsymbol{T}_1$ an additional $(n_2 - n_1)$ runs and equivalently the design $\boldsymbol{T}_1$ is obtained from the design $\boldsymbol{T}_2$ by deleting $(n_2 - n_1)$ runs. The n_u observations for $\boldsymbol{T}_u$ are elements of the vector $\boldsymbol{Y}_{\boldsymbol{T}_u}$, $u = 1, 2$. The (p x 1) vector $\boldsymbol{\beta}$ consists of the p unknown parameters representing the factorial effects of interest including the general mean, σ^2 is an unknown parameter, $\boldsymbol{X}_{\boldsymbol{T}_u}$, $u = 1, 2$ are the (n_u x p) design matrices, and $\boldsymbol{I}_{n_u}$ are the n_u x n_u identity matrices in the models below:

$$E\left(\boldsymbol{Y}_{\boldsymbol{T}_u}\right) = \boldsymbol{X}_{\boldsymbol{T}_u}\boldsymbol{\beta}, \;\; Var\left(\boldsymbol{Y}_{\boldsymbol{T}_u}\right) = \sigma^2\boldsymbol{I}_{n_u}, \;\; p \leq n_1 < n_2, \;\; u = 1, 2. \tag{1}$$

For the (($n_u - p$) x n_u) matrices $\boldsymbol{Z}_{\boldsymbol{T}_u}$ with ranks $(n_u - p)$, $u = 1, 2$, satisfying $\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{X}_{\boldsymbol{T}_u} = \boldsymbol{0}$ with $r = p \leq n_1 < n_2$, Rank ($\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Z}'_{\boldsymbol{T}_u}$) = Rank ($\boldsymbol{Z}_{\boldsymbol{T}_u}$) = $(n_u - p)$ and

$$E\left(\left(\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Z}'_{\boldsymbol{T}_u}\right)^{-\frac{1}{2}}\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Y}_{\boldsymbol{T}_u}\right) = \boldsymbol{0},$$

$$Var\left(\left(\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Z}'_{\boldsymbol{T}_u}\right)^{-\frac{1}{2}}\boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Y}_{\boldsymbol{T}_u}\right) = \sigma^2\boldsymbol{I}_{n_u}. \tag{2}$$

The least squares estimators of $\boldsymbol{\beta}$ and their variances for the models in (1) are

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{T}_u} = \left(\boldsymbol{X}'_{\boldsymbol{T}_u}\boldsymbol{X}_{\boldsymbol{T}_u}\right)^{-1}\boldsymbol{X}'_{\boldsymbol{T}_u}\boldsymbol{Y}_{\boldsymbol{T}_u}, \text{Var}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{T}_u}\right) = \sigma^2\left(\boldsymbol{X}'_{\boldsymbol{T}_u}\boldsymbol{X}_{\boldsymbol{T}_u}\right)^{-1}. \tag{3}$$

More interestingly

$$\text{Cov}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{T}_u}, \boldsymbol{Z}_{\boldsymbol{T}_u}\boldsymbol{Y}_{\boldsymbol{T}_u}\right) = \mathbf{0}. \tag{4}$$

A class of s possible models is considered for describing the n_u observations in the vector $\boldsymbol{Y}_{\boldsymbol{T}_u}$. The μ and p_1 elements in $\boldsymbol{\beta}_1^*$ are common parameters in the s models. In any two models w and w', the p_2 parameters in $\boldsymbol{\beta}_2^{(w)}$ and the p_2 parameters in $\boldsymbol{\beta}_2^{(w')}$ are not all identical: some may be identical and the others are different, $p = 1 + p_1 + p_2$. The s models are then

$$\begin{aligned} E\left(\boldsymbol{Y}_{\boldsymbol{T}_u}\right) &= \mu \boldsymbol{j}_{n_u} + \boldsymbol{X}^*_{1\boldsymbol{T}_u}\boldsymbol{\beta}_1^* + \boldsymbol{X}^{(w)}_{2\boldsymbol{T}_u}\boldsymbol{\beta}_2^{(w)} \\ Var\left(\boldsymbol{Y}_{\boldsymbol{T}_u}\right) &= \sigma^2 \boldsymbol{I}_{n_u}, \;\; p \leq n_1 < n_2, \;\; u = 1, 2. \end{aligned} \tag{5}$$

We consider the problem of model identification and discrimination of these s models. We present the optimum fractional factorial HDs for this purpose.

Keywords: Factorial Experiments, Linear Models, Model Identification, Model Discrimination, Optimum Designs.

References

Atkinson, A. C. and Fedorov, V. V. (1975): The design of experiments for discriminating between two rival models. *Biometrika,* Vol. 62, 57−70.

Ghosh, S., Tian, Y. (2006): Optimum two level fractional factorial plans for model identification, and discrimination. *Journal of Multivariate Analysis,* Vol. 97, 1437−1450.

Srivastava, J. N. (1975): Designs for searching non-negligible effects, in: Srivastava, J. N. (Ed.), *A Survey of Statistical Design and Linear Models,* North−Holland, Amsterdam, pp. 507−520.

Designing Surveillance Strategies for Optimal Control of Epidemics Using Outcome-Based Utilities

Gavin J. Gibson
Heriot-Watt University, Edinburgh, UK
g.j.gibson@hw.ac.uk

In many studies of design the object of the exercise is to provide information on model parameter values. When gathering data on the spread of infectious diseases on the other hand, the object may be to implement control measures using the acquired data and optimal observation strategies are therefore those which maximise the impact of the control in terms of disease reduction. One example would be in the study of arboreal pathogens when infected trees are removed from the population as soon they are detected.

This talk will explore approaches to optimal design where the utility functions to be maximised are based on the subsequent dynamics of the system, rather than measures on the information gained about model parameters. In particular we will explore how so called non-centered parameterisations of epidemic systems might be exploited in order to simplify the task of identifying optimal strategies, when the number of surveillance and control parameters is large.

Keywords: Optimal control of epidemics, outcome-based utilities, non-centered parameterisations.

Unit roots in presence of (double) threshold processes

Francesco Giordano
Di.S.E.S., University of Salerno
giordano@unisa.it

Marcella Niglio, Cosimo Damiano Vitale
Di.S.E.S., University of Salerno
mniglio@unisa.it, cvitale@unisa.it

In time series literature, the study of unit roots in presence of the so called *Threshold Autoregressive Processes* (Tong, 1990), has been differently faced (see among the others, Caner and Hansen (2001), Bec, Salem and Carrasco (2004), Kapetanios and Shin (2006)). In our contribution we introduce a two regimes threshold model characterized by a double threshold variable:

$$\nabla X_t = \rho_1 X_{t-1} I_{t-d} + \rho_2 X_{t-1}(1 - I_{t-d}) + e_t, \tag{1}$$

with $\nabla X_t = X_t - X_{t-1}$, $\rho_j = \phi_1^{(j)} - 1$, for $j = 1, 2$ the indicator function $I_{t-d} = 1$ if $X_{t-d} \geq r_1$ and $I_{t-d} = 0$ otherwise.
where the second threshold is related to the nonlinear structure of the error term e_t that has the following structure

$$e_t = e_{1,t} I'_{t-1} + e_{2,t}(1 - I'_{t-1}) \tag{2}$$

with indicator function I'_{t-1}:

$$I'_{t-1} = \begin{cases} 1 & \text{if} \quad \nabla X_{t-1} \geq r'_1 \\ 0 & \text{if} \quad \nabla X_{t-1} < r'_1. \end{cases}$$

In model (1)-(2) $\{e_{i,t}\}$ is a sequence of i.i.d. random variables with $E(e_{i,t}) = c_i$, $-\infty < c_i < \infty$ and $Var(e_{i,t}) = \sigma_i^2 < \infty$, for $i = 1, 2$.
Following Seo (2008) we employ and evaluate the performance of an

Augmented Dickey and Fuller type test to investigate the presence of unit roots in the proposed model. Instead of the sampling distribution of the test is approximated in Seo (2008) by using a residual block bootstrap approach (Paparoditis and Politis, 2003), a non trivial problem has been left open: it is related to the block length that, as expected, affects the inferential issues. The selection of the block length has been faced in our contribution whose results have been discussed through a wide Monte Carlo study.

Keywords: Threshold model, unit roots, block bootstrap.

References

Bec F., Salem M.B., Carrasco M. (2004): Tests for Unit-Root versus Threshold Specification With an Application to the Purchasing Power Parity Relationship, *Journal of Business & Economic Statistics*, Vol. 22, pp. 382-395.

Caner M., Hansen B. (2001): Threshold Autoregression with a Unit Root. *Econometrica*, Vol. 69, pp. 1555-1596.

Kapetanios G., Shin Y. (2008): Unit root tests in three-regime SETAR model. *The Econometrics Journal* Vol. 9, pp. 252278.

Paparodotis E., Politis D.N. (2003): Residual-based block bootstrap for unit root testing, *Econometrica*, Vol. 71, N. 3, pp. 813-855.

Seo H.M. (2008): Unit root test in a threshold autoregression: asymptotic theory and residual-based block bootstrap. *Econometric Theory*, Vol. 24, pp. 1699-1716.

Tong H. (1990): *Non-Linear Time Series. A Dynamical System Approach*, Clarendon Press, Oxford.

Simulation in clinical trials: some design problems

Alessandra Giovagnoli
ex-Department of Statistical Sciences
alessandra.giovagnoli@unibo.it

Clinical trials are usually very lengthy and/or very costly, frequently fail - in the sense that they turn out to be pointless or inconclusive - and especially there is always the possibility of adverse effects for the patients or healthy volunteers entering the trial. Since the well-known FDA's 2004 Critical Path Initiative, Modelling and Simulation (M&S) has rooted itself firmly in the clinical research area, as evidently shown in the medical and pharmaceutical literature, see for instance Kimko and Peck (2011).

It has been suggested (Padilla *et al.*, 2011) to divide the whole M&S area into: M&S Theory, defining the academic foundations of the discipline, M&S Engineering, looking for general methods that can be applied in various problem domains, and M&S Applications, solving real world problems. This is a particularly insightful distinction for clinical trial simulation (CTS) too, and the theoretical aspects should not be neglected. Nowadays a large amount of CTS software is available on the web, either commercially or freely downloadable, but it is legitimate to ask ourselves how trustworthy and useful it is.

The topic of this presentation is to further the discussion of experimental design problems in CTS, started in Giovagnoli and Zagoraiou (2012). The simulator of a clinical trial will very likely include a stochastic component so the rationale for using standard statistical tools, in particular, standard experimental design theory, is restored; however, planning a simulated experiment is different from designing a real one, due to differences in the endpoints, aims, precision required etc. In simulations, we would normally experiment on a wider design space and/or increase the number of factors of interest and the levels that are simul-

taneously tried. One may also wonder about the role of randomization. Most of all, adaptive design deserves special attention, since by their very nature simulated experiments are mostly adaptive. It would be interesting to see if a combined approach of optimal design methods and simulation brings useful results.

Keywords: Experimental Design, Clinical Trial Simulation.

References

Kimko H.C., Peck C. (eds.) (2011) *Clinical Trial Simulations: Applications and Trends* Volume I: Series Advances in the Pharmaceutical Sciences Series, Springer Publishing.

Padilla J., Diallo S.Y., Tolk A. (2011): Do We Need M&S Science? *SCS M&S Magazine* Vol. 4, pp.161-166

Giovagnoli A., Zagoraiou M. (2012): Simulation of Clinical Trials: a review with emphasis on the design issues *STATISTICA*, Vol LXXII N.1, pp.63-80.

Estimation of complex item response theory models using Bayesian simulation-based techniques

Cees A. W. Glas
Department of Research Methodology, Measurement, and Data Analysis
University of Twente, the Netherlands
c.a.w.glas@utwente.nl

In the last ten years, the family of item response theory (IRT) models has expanded tremendously (see, for instance, van der Linden & Hambleton, 1997; de Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004). In this presentation, estimation of complex IRT models will be discussed. First, it will be shown that complex IRT model consist of an IRT measurement model and an additional structural model. Examples discussed are a multidimensional IRT model (Bguin, & Glas, 2001), multilevel IRT models (Fox & Glas, 2001), and generalizability theory IRT models (Briggs & Wilson, 2007). Two approaches to estimating such models are discussed. The first one is a two-step procedure, where the IRT measurement model is validated in a first step, followed by a second step where the structural model is estimated conditional on the results of the first phase. It is argued, that for the second phase, Bayesian simulation-based methods such as Markov chain Monte Carlo are much more convenient than traditional likelihood-based methods. The second estimation procedure is an analogous Bayesian procedure where both the measurement model and the structural model are estimated concurrently. The presentation will be completed by several applications from the field of educational measurement.

Keywords: Bayesian Statistical Methods, Item Response theory, MCMC, multilevel models.

References

Béguin, A.A., Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 471-488.

Briggs, D.C., Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44,* 131-155.

De Boeck, P., Wilson, M. (Eds.) (2004). *Explanatory Item Response Models: A generalized linear and nonlinear approach.* New York NJ: Springer.

Fox, J.P., Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized latent variable modeling.* London: Chapman & Hall.

van der Linden, W. J., Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory.* New York, NJ: Springer Verlag.

Potential advantages and disadvantages of stratification in methods of randomization

Aenne Glass
Institute for Biostatistics and Informatics in Medicine and Ageing Research
University Medicine Rostock
aenne.glass@uni-rostock.de

Guenther Kundt
Institute for Biostatistics and Informatics in Medicine and Ageing Research
University Medicine Rostock
guenther.kundt@uni-rostock.de

Clinical trials are common to evaluate the effectiveness and safety of a new medication to diagnose or treat a disease. To prevent imbalance between treatment groups for known prognostic factors, patients are randomized in strata. This may help to prevent type I and type II errors via reduction of variance. On the other hand, stratification involves administrative efforts. Cost-benefit-questions have to be clarified.
To investigate a shortlist of potential advantages and disadvantages of stratification, first, complete randomization was considered to quantify its risk of imbalance (Kernan, 1999), and second, the stratified case was compared vs. the unstratified for restricted randomization from different angles (Green, 1978). Determining the probability of

i) observing prognostic imbalance of at least 10% between two treatment groups, caused by complete (unstratified) randomization,

ii) observing a (clinically or statistically relevant) difference between the endpoints of two treatments, when both treatments were equally effective (type I error),

iii) failing to observe a difference that truly exists (type II error),

by simulation studies, a quantification of the influence of stratification on particular trial characteristics can consequently be given.

We show simulation results indicating that stratification is helpful in restricted randomization procedures for superiority trials with less than 500 patients. The risk of randomization-associated prognostic imbalance can amount to 0.16-0.60 for 50% factor prevalence, depending on trial size. We show further that the risk of imbalance multiplies for smaller trials or/and according to a more prevalent prognostic factor. In large superiority trials relevant imbalance has not been observed, independently of any factor prevalence, and hence, it is not necessary to stratify here. We show decreased probabilities of type I error from expected 0.05 unstratified to 0.036-0.001 stratified, and decreased type II error. Keeping the number of strata small so that the complexity of randomization scheme remained manageable no disadvantage of stratification could be detected so far.

Keywords: randomized clinical trials, stratified randomization, imbalance

References

Kernan, W. N., C. M. Viscoli, R. W. Makuch, L. M. Brass and R. I. Horwitz (1999): Stratified Randomization for Clinical Trials, *Journal of Clinical Epidemiology*, Vol. 52, N. 1, pp. 19-26.

Green, S. B. (1978): The effect of stratified randomization on size and power of statistical tests in clinical trials, *Journal of Chronic Diseases*, Vol. 31, N. 6-7, pp. 445-454.

Application of nonparametric goodness-of-fit tests for composite hypotheses

Alisa A. Gorbunova, Boris Yu. Lemeshko, Stanislav B. Lemeshko and Andrey P. Rogozhnikov
Novosibirsk State Technical University, Novosibirsk, Russia
lemeshko@fpm.ami.nstu.ru

In this paper we consider the problem of testing composite hypotheses of the form H_0: $F(x) \in \{F(x,\theta), \theta \in \Theta\}$, when the estimate $\hat{\theta}$ of scalar or vector parameter of the distribution is calculated using the same sample. In this case the conditional distribution $G(S\,|H_0\,)$ of nonparametric test statistics is affected by a number of factors: the form of the tested distribution $F(x,\theta)$; the type and the number of the parameters estimated; and sometimes the value of the parameter and the estimation method.

To solve the problem of testing composite hypotheses using the Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests different approaches in the studies of various authors were used. In our studies (Lemeshko et al. (2009, 2009a, 2010, 2011)) we used simulation methods to construct models of statistics distributions and tables of percentage points. When statistics distributions of nonparametric goodness-of-fit tests depend on the value of the distribution parameter (for example, in the case of the inverse Gaussian, generalized Weibull distributions), the distribution models were approximated for some integer values of the corresponding parameters in papers of Lemeshko et al. (2009, 2010, 2011).

While testing the composite hypotheses parameters are estimated during the analysis. Therefore, there is no way to find preliminarily the required distribution to test the hypothesis of a goodness-of-fit for the specific value of the parameter $\hat{\theta}$. In this case we propose to find the distribution of the test statistic and to calculate the p-value in the interactive way (while testing the corresponding composite hypothesis).

An implemented interactive technique enables the correct application of the criteria even in the cases when the tests statistics distribution for true H_0 is unknown. The software developed allows us to use methods of parallel computing to accelerate calculations and to use all available computing resources. The software makes it possible to test composite hypotheses for various parametric models of probability distribution using nonparametric Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling, Kuiper, Watson and three different Zhang tests.
Keywords: goodness-of-fit testing, composite hypothesis, Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling, Zhang tests.

References

Lemeshko B.Yu., Lemeshko S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part I. *Measurement Techniques*. Vol. 52, 6. - P.555-565.

Lemeshko B.Yu., Lemeshko S.B. (2009a). Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques*. Vol. 52, 8. - P.799-812.

Lemeshko B.Yu., Lemeshko S.B. and Postovalov S.N. (2010). Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses. *Communications in Statistics - Theory and Methods*, 2010. Vol. 39, No. 3. - P. 460-471.

Lemeshko B.Yu., Lemeshko S.B. (2011). Models of Statistic Distributions of Nonparametric Goodness-of-Fit Tests in Composite Hypotheses Testing for Double Exponential Law Cases. *Communications in Statistics - Theory and Methods*, Vol. 40, No. 16. - P. 2879-2892.

Indirect inference and data cloning for non-hierarchical mixed effects logiit models

Anna Gottard, Giorgio Calzolari
Dipartimento di Statistica, Informatica, Applicazioni (DISIA)
gottard@disia.unifi.it, calzolari@disia.unifi.it

Common statistical applications involve inference for statistical models in the presence of unobserved relevant factors, acting on data within a clustered structure. Non-hierarchical data are sometimes of interest. An interesting class of models in this framework is Multiple Membership models (Hill and Goldstein, 1998), an extension of hierarchical multilevel models. These models allow a statistical unit to belong to more than one cluster. Multiple membership models assign random effects for each element of the grouping, supposing that, conditionally on a latent variable, units are *iid*. An interesting case consists Multiple membership logit (MML) models. The basic problem in these models is that a multi-dimensional integration has to be computed in order to obtain maximum likelihood estimates and the likelihood function is therefore intractable. Ignoring the multiple membership clustering might bring to distort results, while composite likelihoods such as quasi-likelihood or partial-likelihood have been shown to provide seriously biased and inconsistent estimators in the case of binary responses (Rodriguez and Goldman, 1995).

At the moment, the most preferable procedure for MML model estimation is based on Bayesian paradigm and MCMC methods. However, some researchers could prefer to avoid Bayesian inference as unable of making an explicit choice of a priori distributions or for preferring frequentist inference. To solve this inferential issue in a non-Bayesian framework, we are here proposing two different approaches: data cloning and indirect inference. Data cloning (Lele et al., 2007) is a novel approach to compute maximum likelihood estimates together their standard errors, as the collapsing points of posterior distributions

computed on cloned data. Indirect inference (Gourieroux *et al.*, 1993) is a class of estimators. including the generalized and simulated methods of moments. It is based on a simulation estimation procedure utilizing an auxiliary model for estimating the parameters of the model of interest. In this work propose a data cloning estimator together with an indirect estimator for MML models. In particular, we show how both estimates and standard errors can be easily derived with both the approaches. The two proposed estimators are compared by means of a Monte Carlo study. Simulations show a negligible loss of efficiency for the indirect inference estimator, compensated by a relevant computational gain.

Keywords: Data cloning, Indirect inference, Intractable likelihoods, Multiple membership logit models, Non-hierarchical data.

References

Gourieroux C., Monfort A., Renault E. (1993): Indirect inference. *Journal of Applied Econometrics* Vol. 8(S1), pp. S85-S118.

Hill P., Goldstein H. (1998): Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral statistics*, Vol. 23, N. 2, pp. 117-128.

Lele S., Nadeem K., Schmuland B. (2010): Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, Vol. 105, N. 492, pp. 1617-1625.

Rodriguez G., Goldman N. (1995): An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 158, N. 1, pp.73-89.

Outliers in Multivariate GARCH Models

Aurea Grané
Universidad Carlos III de Madrid
aurea.grane@uc3m.es

Helena Veiga, Belén Martín-Barragán
Universidad Carlos III de Madrid
mhveiga@est-econ.uc3m.es, bmbarrag@est-econ.uc3m.es

In multivariate time series, correct estimation of the correlations among the series plays a key role in portfolio selection. In the univariate case, presence of outliers in financial data is known to lead parameter estimation biases, invalid inferences and poor volatility forecasts. This work analyses the impact of outliers in multivariate time series. We found that the impact in volatility follows a similar pattern to that in univariate time series, but, more interesting, our multivariate approach allows to analyse the impact on correlations. In Grané and Veiga (2009) a general outlier detection wavelet-based method was proposed, which was proven to be very effective and much more reliable than other alternatives in the literature. Our proposal is to extend this procedure to the context of Multivariate GARCH models by considering random-projections of multivariate residuals. The models under study are the Diagonal BEKK (described in Engle and Kroner 1995), CCC (Bollerslev 1990) and DCC (Engle 2002) models, often used in empirical applications. The effectiveness of this new procedure is evaluated through an intensive Monte Carlo study considering isolated and patches of additive level outliers (ALOs) and additive volatility outliers (AVOs).

To illustrate the performance of our proposal, in Table 1 we show the percentage of correct detection of ALOs and the average number of false ALOs (false positives) in 1000 replications of simulated series of size n from a D-BEKK model with multivariate Gaussian errors. Outliers were randomly placed along the bivariate series. As preliminary results, we observe that the percentage of correct detections is very high when

the size of the outliers is moderate ($\omega = 10\sigma_y$). The average of false positives is quite small, meaning that the detection method is reliable.

Table 1. First results for the D-BEKK model

	n	% correct detections	false positives
1 outlier of size $\boldsymbol{\omega} = 5\boldsymbol{\sigma_y}$	1000	43.8	3.5
	3000	38.7	3.8
	5000	36.1	3.6
1 outlier of size $\boldsymbol{\omega} = 10\boldsymbol{\sigma_y}$	1000	99.1	3.6
	3000	99.3	3.1
	5000	99.3	3.9
3 outliers of size $\boldsymbol{\omega} = 5\boldsymbol{\sigma_y}$	1000	36.7	1.0
	3000	36.5	1.2
	5000	36.1	1.1
3 outliers of size $\boldsymbol{\omega} = 10\boldsymbol{\sigma_y}$	1000	96.5	0.5
	3000	97.8	1.1
	5000	97.8	1.3

Keywords: Multivariate GARCH Models, Outliers, Wavelets.
Acknowledgements: This work was supported by MTM2010-17323, ECO2012-32401, ECO2011-25706 and MTM2012-36163-C06-03

References

Bollerslev, T. (1990): Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model, *Review of Economics and Statistics*, Vol. 42, pp. 498-505.

Grané A., Veiga H. (2009): Wavelet-based detection of outliers in financial time series, *Computational Statistics and Data Analysis*, Vol. 54, N. 11, pp. 2580-2593.

Engle, R. and K. Kroner (1995): Multivariate simultaneous generalized ARCH, *Econometric Theory*, Vol. 11, pp. 122-150.

Engle, R. (2002): Dynamic conditional correlationa simple class of multivariate GARCH models, *Journal of Business and Economic Statistics*, Vol. 20, pp. 339-350.

On a Bahadur-Kiefer representation of von Mises statistic type for intermediate sample quantiles

Nadezhda Gribkova
St.Petersburg State University, Mathematics and Mechanics Faculty, Russia
nv.gribkova@gmail.com

Roelof Helmers
Centre for Mathematics and Computer Science, Amsterdam, the Netherlands
R.Helmers@cwi.nl

Consider a sequence $X_1, X_2, \dots$ of i.i.d. real-valued random variables with distribution function (*df*) F, $X_{1:n} \le \cdots \le X_{n:n}$ – the order statistics based on the sample $X_1, \dots, X_n$, $n \in \mathbb{N}$. Let $F^{-1}(u) = \inf\{x : F(x) \ge u\}$, $u \in (0,1)$, denote the left-continuous inverse of F, and F_n, F_n^{-1} – the empirical *df* and its inverse respectively, put $f = F'$ to be a density of F, when it exists.

Let k_n be a sequences of integers, such that $k_n \to \infty$, whereas $p_n := k_n/n \to 0$, as $n \to \infty$. Let $\xi_{p_n} = F^{-1}(p_n)$, $\xi_{p_n\, n:n} = F_n^{-1}(p_n)$ denote p_n-th population and empirical quantile respectively.

Let $SRV_\rho^{-\infty}$ be a class of regularly varying in $-\infty$ functions such that $g \in SRV_\rho^{-\infty}$ if and only if:
(i) $g(x) = \pm|x|^\rho\, L(x)$, for $|x| > x_0$, with some $x_0 < 0$, $\rho \in \mathbb{R}$, and $L(x)$ is a positive slowly varying function at $-\infty$;
(ii) $\Big|\, g(x + \triangle x) - g(x)\Big| = O\Big(|\, g(x)|\, \Big|\frac{\triangle x}{x}\Big|^{1/2}\Big)$, when $\triangle x = o(|x|)$, as $x \to -\infty$. Here is one of our main results.

THEOREM 1. *Let $k_n \to \infty$, $p_n \to 0$, as $n \to \infty$, and suppose that F^{-1} is differentiable in $(0, \varepsilon)$ for some $\varepsilon > 0$ and that $f \in SRV_\rho^{-\infty}$ with $\rho = -(1+\gamma)$, $\gamma > 0$. Let G be some function differentiable in $F^{-1}((0,\varepsilon))$, and $g = G' \in SRV_\rho^{-\infty}$ with some $\rho \in \mathbb{R}$. Then*

$$\int_{\xi_{p_n\, n:n}}^{\xi_{p_n}} (G(x) - G(\xi_{p_n}))\, d\, F_n(x) = -\frac{1}{2}[F_n(\xi_{p_n}) - p_n]^2 \frac{g}{f}(\xi_{p_n}) + R_n,$$

where $\boldsymbol{P}\big(|R_n| > A\, p_n^{3/4}(\log k_n/n)^{5/4}\frac{|g|}{f}(\xi_{p_n})\big) = O(k_n^{-c})$ *for each* $c > 0$*, and* $A > 0$ *is some constants, which depends only on* c*.*

Moreover, if, in addition, $k_n^{-1}\log n \to 0$*, as* $n \to \infty$*, then* $\boldsymbol{P}\big(|R_n| > A\, p_n^{3/4}(\log n/n)^{5/4}\frac{|g|}{f}(\xi_{p_n})\big) = O(n^{-c})$.

Theorem 1 provides a Bahadur – Kiefer type representation for the sum of order statistics lying between the intermediate population p_n-quantile and the corresponding sample quantile by a von Mises type statistic approximation, especially useful in establishing second order approximations for (slightly) trimmed means (cf. Gribkova & Helmers (2007, 2013)).

Keywords: Bahadur – Kiefer type representation, intermediate sample quantiles, Bahadur – Kiefer processes, quantile processes, von Mises statistic type approximation.

Acknowledgements: The work of the first author was partially supported by the Russian Foundation for Basic Research (grant RFBR no. SS-1216.2012.1)

References

Gribkova, N.V., Helmers, R. (2007): On the Edgeworth Expansion and the M out of N Bootstrap Accuracy for a Studentized Trimmed Mean, *Math. Methods Statist.*, Vol. 16, pp. 142-176.

Gribkova, N., Helmers, R. (2012): On a Bahadur-Kiefer Representation of von Mises Statistic Type for Intermediate Sample Quantiles, *Probab. Math. Statist.*, Vol. 32, N. 2, pp.255-279.

Gribkova, N.V., Helmers, R. (2013): Second Order Approximations for Slightly Trimmed Means, *Theory Probab. Appl.*, Vol.58 (to appear); arXiv:1104.3347v1 [math.PR].

Modeling the Optimal Investment Strategies in Sparre Andersen Risk Model

Alexander Gromov
Lomonosov Moscow State University
gromovaleksandr@gmail.com

We consider a special case of Sparre Andersen risk model in which interclaim times have Erlang(2) distribution. Let T_i be the occurrence time of the i-th claim, N_t number of claims in time interval $[0, t]$, Y_i the amount of the i-th claim and c the premium intensity of the insurer. It is supposed that $Y_1, Y_2, \ldots$ are positive r.v.'s with absolute continuous distribution function Q and $(T_i - T_{i-1}) \sim Erlang(2, \beta)$, $\beta > 0$. We also assume that there is a risky asset and the insurer has the possibility to invest money in this asset. The price Z_t of this asset is modeled by geometric Brownian motion

$$dZ_t = Z_t(\mu dt + \sigma dW_t).$$

The insurer dynamically chooses the amount A_t of capital invested into the risky asset at time t. Moreover, we consider the investment strategies $A = (A_t)_{t \geq 0}$ adapted to filtration generated by Brownian motion W_t. Using some strategy A the capital of the insurer R_t^A satisfies the following stochastic differential equation:

$$dR_t^A = (c + \mu A_t)dt + \sigma A_t dW_t - dU_t,\ U_t := \sum_{i=1}^{N_t} Y_i,$$

where $R_0^A = s > 0$ is the initial capital. Let $\tau^A := \inf\{t > 0 : R_t^A < 0\}$ be the ruin time and $\delta^A(s) := P(\tau^A = \infty | R_0^A = s)$ the ultimate survival probability of the insurer under investment strategy A. Our goal is to maximize $\delta^A(s)$ over all admissible investment strategies, i.e. find optimal survival probability $\delta(s) := \sup_A \delta^A(s)$ and optimal strategy A^* such that $\delta(s) = \delta^{A^*}(s)$. We assume that function $\delta(s) \in$

$C^4[0,\infty)$ and prove that the optimal survival probability $\delta(s)$ satisfies the Hamilton–Jacobi–Bellmann equation

$$\sup_{A\geq 0}\{\beta^2 E\delta(s-Y) - \left(\beta - (c+\mu A)\frac{d}{ds} - \frac{\sigma^2 A^2}{2}\frac{d^2}{ds^2}\right)^2 \delta(s)\} = 0.$$

From this equation we deduce that the optimal A^*, which maximizes the left-hand side of the equation could be obtained from the following cubic function for each $s \geq 0$

$$\sigma^4\delta^{(4)}(s)A^3 + 3\sigma^2\mu\delta'''(s)A^2 + (2\mu^2\delta''(s) - 2\sigma^2\delta''(s) + 2\sigma^2 c\delta'''(s))A + 2\mu c\delta''(s) - 2\beta\mu\delta'(s) = 0$$

Though it is complicated to find the explicit formulae for the optimal survival probability in this case, it is convenient to compute the solution and optimal strategy for particular values of the parameters, using approximation algorithms similar to those provided in (Gerber, Shiu, 2004) and (Willmot, Woo, 2006). In this paper we discuss particular cases of exponentially and Pareto distributed claims.

Keywords: Erlang(2) risk model, ruin probability, investment, HJB equation.

References

Gerber H.U., Shiu S.W. (2004): The Time Value of Ruin in a Sparre Andersen Risk Model, *North American Actuarial Journal*, Vol. 2, N. 9, pp. 99–118.

Willmot G.E., Woo J.H. (2006): On the Class Of Erlang Mixtures with Risk Theoretic Applications, *North American Actuarial Journal*, Vol. 2, N. 11, pp. 99–118.

Modeling longitudinal data with finite mixtures of regression models

Bettina Grün
Johannes Kepler Universität Linz
Bettina.Gruen@jku.at

Finite mixtures of regression models are useful for longitudinal data where the development of a variable over time is to be described and heterogeneity with respect to this evolvement over time is suspected. In the following we consider finite mixtures of linear mixed-effects models for potentially censored data (Grün and Hornik, 2012) and finite mixtures of linear additive models (Grün, Scharl and Leisch, 2011). Both models are estimated using a variant of the EM algorithm.

Finite mixtures of linear mixed-effects models are fitted to account for individual-specific effects in longitudinal data. In this case the missing data in the EM algorithm are the component memberships (as usually for finite mixture models) and the random effects. If the data also contains censored observations, these unobserved values are also added to the missing data. An efficient implementation of the E-step is then possible by determining the required moments of the truncated multivariate normal distribution using the method proposed in Tallis (1961).

B-splines are used in finite mixtures for longitudinal data to allow a flexible modeling of the functional development over time. However, the degrees of freedom for the spline bases are often restricted to be the same over all components. In addition the optimal number is determined by comparing the models fitted with all different degrees of freedom under consideration which is computationally expensive. Linear additive models have the advantage that the number of degrees of freedom only need to be sufficiently large without the exact number being crucial. The suitable smoothness of the function is automatically determined by regularizing the coefficients of the spline regressors. We fit finite mixtures of linear additive models using a variant of the EM

algorithm where different suitable smoothness levels are determined for each component.

We describe the two different models as well as their estimation with variants of the EM algorithm. Their application is illustrated and the R package **flexmix** (Grün and Leisch, 2007) is presented which allows to fit these models.

Keywords: finite mixture, longitudinal data, regression model, mixed-effects model.

Acknowledgements: This work was supported by the Austrian Science Fund (FWF): Elise-Richter grant V170-N18.

References

Grün B., Hornik K. (2012): Modelling Human Immunodeficiency Virus Ribonucleic Acid Levels with Finite Mixtures for Censored Longitudinal Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 61, N. 2, pp. 201–218.

Grün B., Leisch F. (2007): Fitting Finite Mixtures of Generalized Linear Regressions in R. *Computational Statistics & Data Analysis*, Vol. 51, N. 11, pp. 5247–5252.

Grün B., Scharl T., Leisch F. (2011): Modelling Time Course Gene Expression Data with Finite Mixtures of Linear Additive Models. *Bioinformatics*, Vol. 28, N. 2, pp. 222–228.

Tallis G. M. (1961): The Moment Generating Function of the Truncated Multi-Normal Distribution. *Journal of the Royal Statistical Society B*, Vol. 23, N. 1, pp. 223–229.

A Bayesian nonparametric mixture model for cluster analysis

Alessandra Guglielmi
Politecnico di Milano, Dipartimento di Matemaica
alessandra.guglielmi@polimi.it

Raffaele Argiento[1], Andrea Cremaschi[2]
[1]CNR-IMATI, Milano, [2]University of Kent
raffele@mi.imati.cnr.it, ac429@kent.ac.uk

We introduce a new model for cluster analysis in a Bayesian nonparametric context. Typically, clustering means partitioning a set of n objects (i.e. data) into k groups, even if the common features of the objects in each group are unknown or unobservable (i.e. latent).

Here we propose a Bayesian nonparametric model, that combines two ingredients: Dirichlet process mixture (DPM) models of Gaussian distributions, and a heuristic clustering procedure, called DBSCAN. The DBSCAN algorithm (Ester et al., 1996) is a density-based clustering technique, where the word *density* refers to the spatial disposition of the data points, that are *dense* when forming a group; two data points are in the same cluster if their distance is smaller than some threshold. On the other hand, it is well-known that DPM models are convenient in order to assign a prior directly on the partition of the data, representing the natural parameter in the cluster analysis context. Moreover, the number of clusters is not fixed a priori, but it is estimated as a feature of the partition of the observations. However, here, instead of considering the prior on the random partition ρ of the data induced from the DPM, we consider a deterministic transformation of ρ as a new parameter. The Bayesian cluster estimate will be given in terms of this new random partition, and will result from the minimization of the posterior expectation of a loss function.

To summarize, our model is based on the slackness of the natural

clustering rule of DPM models of parametric densities, when we mean that two observations X_i and X_j are in the same cluster if, and only if, the latent parameters θ_i and θ_j are equal. We say instead that two observations share the same cluster if the distance between the densities corresponding to their latent parameters is smaller than a threshold ϵ. We complete the definition in order to provide an equivalence relation among data labels. The resulting new random partition parameter ρ_ϵ is coarser than the original ρ. Of course, since this procedure depends on the value of the threshold, we suggest a strategy to fix it.

In addition, we discuss implementation and applications of the model to a simulated bivariate dataset from a mixture of two densities with a curved cluster, and to a dataset consisting of gene expression profiles measured at different times, known in literature as Yeast cell cycle data. In particular, we implemented a MCMC algorithm that extends the well-known Gibbs sampler algorithms for DPM models; the code was written in C. Comparison with more standard clustering algorithm will also be given. In both cases, the cluster estimates from our model turn out to be more effective.

Keywords: Bayesian Nonparametrics, Dirichlet process mixture models, Cluster analysis, DBSCAN.

References

Ester, M., Kriegel, H. P. , Xu, X. (1996): Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification. In: *Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, Lecture Notes in Computer Science*, Springer, pp. 67–82.

A Comparison of Different Permutation Approaches to Testing Effects in Unbalanced Two-Level ANOVA Designs

Sonja Hahn
Department of Psychology, University of Jena
hahn_sonja@uni-jena.de

Luigi Salmaso
Department of Management and Engineering, University of Padova
luigi.salmaso@unipd.it

Analysis of Variance (ANOVA) is a procedure used to compare the means of various samples. Qualitative variables called factors group the different samples. Each factor consists of two or more levels. Parametric ANOVA approaches assume normally distributed error terms within the samples. Permutation tests like Synchronized Permutations (Pesarin and Salmaso, 2010) do not impose this assumption. These procedures rearrange the data to obtain an empirical distribution of the test statistic. A variety of permutation approaches for ANOVA designs have been developed. They differ in various aspects. The units that are actually permuted (e.g. raw data vs. residuals), if there are restrictions in the permutation mechanism and in the definition of the test statistic (see e.g. Kherad-Pajouh and Renaud, 2010).

In many real applications the sample sizes in an ANOVA differ. This is called an unbalanced design. There is a broad literature about unbalanced designs in parametric testing. For permutation tests this topic recieved some attention recently (see e.g. Kherad-Pajouh and Renaud, 2010). The present paper extents the Synchronized Permutation approach to unbalanced two-way two-level ANOVA designs. It further compares it to different other permutation approaches and a parametric ANOVA by a simulation study.

The simulation study investigates the behavior of the different pro-

cedures for various types of unbalanced designs, different error term distributions (normal distribution, Cauchy distribution, exponential distribution) and different combinations of active effects. The main outcomes are the adherence to the nominal significance level as well as power.

The Synchronized Permutation approach yields comparable results to the best performing competing permutation approaches. In comparison with parametric ANOVA these permutation approaches adhere to the nominal significance level also for non normal error term distributions.

We discuss the benefits and limits of the different permutation approaches with regard to the results of the simulation study and additional considerations.

Keywords: Synchronized Permutations, Conditional Testing Procedures, Nonparametric Statistics, Unbalanced ANOVA.

References

Kherad-Pajouh, S., Renaud, D. (2010): An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA, *Computational Statistics and Data Analysis*, Vol. 54, N. 7, pp. 1881-1893

Pesarin, F., Salmaso, L. (2010): *Permutation Tests for Complex Data: Theory, Application and Software*, Wiley & Sons, New York.

Likelihood-Free Simulation-Based Optimal Design

Markus Hainy, Werner G. Müller, Helga Wagner
Johannes Kepler University Linz, Austria
markus.hainy@jku.at, werner.mueller@jku.at,
helga.wagner@jku.at

Simulation-based optimal design techniques are a convenient tool for solving a particular class of optimal design problems. The goal is to find the optimal configuration of factor settings with respect to an expected utility criterion. This criterion depends on the specified probability model for the data and on the assumed prior distribution for the model parameters. We develop new simulation-based optimal design methods which incorporate likelihood-free approaches and utilize them in novel applications.

Most simulation-based design strategies solve the intractable expected utility integral at a specific design point by using Monte Carlo simulations from the probability model. Optimizing the criterion over the design points is carried out in a separate step. Müller (1999) introduces an MCMC algorithm which simultaneously addresses the simulation as well as the optimization problem. In principle, the optimal design can be found by detecting the mode of the sampled design points. Several improvements have been suggested to facilitate this task for multidimensional design problems (see e.g. Amzal, Bois, Parent, and Robert, 2006).

We aim to extend the simulation-based design methodology to design problems where the likelihood of the probability model is of an unknown analytical form but it is possible to simulate from the probability model. We further assume that prior observations are available. In such a setting it is seems natural to employ approximate Bayesian computing (ABC) techniques in order to be able to simulate from the conditional probability model. In particular, we consider a spatial extreme value ex-

ample where it is not possible to obtain the analytical representation of the generalized extreme value distribution for dimensions greater than two (cf. Erhardt and Smith, 2012). We investigate the benefits and the limitations of our design methodology for this problem.

Keywords: Simulation based optimal design, approximate Bayesian computing, Markov chain Monte Carlo.

References

Amzal B., Bois F.Y., Parent E., Robert C.P. (2006): Bayesian-Optimal Design via Interacting Particle Systems, *Journal of the American Statistical Association*, Vol. 101, N. 474, pp. 773-785.

Erhardt R.J., Smith R.L. (2012): Approximate Bayesian Computing for Spatial Extremes, *Computational Statistics and Data Analysis*, Vol. 56, N. 6, pp. 1468-1481.

Müller P. (1999): Simulation Based Optimal Design. In: Bernardo J.M., Berger J.O., Dawid A.P., Smith, A.F.M. (Eds.) *Bayesian Statistics 6*, Oxford University Press, New York, pp. 459-474.

Dynamic Structured Copula Models

Wolfgang Härdle, Ostap Okhrin
C.A.S.E. - Centre for Applied Statistics and Economics,
Humboldt-Universität zu Berlin, D-10178 Berlin, Germany
haerdle@wiwi.hu-berlin.de, ostap.okhrin@wiwi.hu-berlin.de

Yarema Okhrin
Dep. of Statistics, University of Augsburg, D-86135 Augsburg, Germany
yarema.okhrin@wiwi.uni-augsburg.de

The key difference between univariate and multivariate time series analysis is the fact that the future dynamics is affected not only by the univariate past but also by cross-sectional dependencies. These dependencies are not constant and vary in time. The most straightforward and therefore best established approach to modelling such dependencies is via the correlation (or covariance) matrix. Time varying conditional volatilities are modelled using, e.g., GARCH-type processes. These models still assume that the parameters for the process are constant over the entire estimation period.

Another disadvantage of covariance-based dependency modelling is the fact that it fails to capture important types of data features. First, covariances are measures of linear dependence and therefore fail to represent nonlinear relationships. Secondly, elliptical distributions postulate symmetric dependency. Thirdly, the covariance matrix fails to fit the heavy tails typical of asset returns. An approach which partially solves these problems is based on copulae.

Time-varying copulae were considered recently by Patton (2004), Rodriguez (2007) and others. In contrast to those papers, Giacomini, Härdle and Spokoiny (2009) used a novel method based on local adaptive estimation. The idea of that approach is to determine a period of homogeneity wherein the parameter of a low-dimensional Archimedean copula can be approximated by a constant.

The online instantaneous selection of high dimensional dependency

structures via multivariate copulae is still an open problem. Here we tackle this problem via multivariate hierarchical Archimedean copulae. A detailed analysis of this copula class is given in Okhrin, Okhrin and Schmid (2013). Unlike simple Archimedean copulae, the HAC is characterised not only by its parameters, but also by its structure. The time-varying dependency therefore affects its structure and parameters simultaneously. The proposed technique allows us to determine the periods with local constant structure and parameters. It is based on the selection of an appropriate interval out of a set of candidate intervals. This procedure requires the calculation of a sequence of critical values (by simulations) that are used in testing local homogeneity. Local homogeneity is checked via a test against a change point alternative. To assess the performance of the methodology developed, we perform extensive simulations and empirical studies.

Keywords: copula, multivariate distribution, Archimedean copula, adaptive estimation.

References

Giacomini, E., Härdle, W. K. and Spokoiny, V. (2009): Inhomogeneous dependence modeling with time-varying copulae, *Journal of Business and Economic Statistics* 27(2), pp. 224-234.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013): On the structure and estimation of hierarchical Archimedean copulas, *Journal of Econometrics* 173, pp. 189-204.

Patton, A. J. (2004): On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* 2, pp. 130-168.

Rodriguez, J. C. (2007): Measuring financial contagion: A copula approach, *Journal of Empirical Finance 14*, pp. 401-423.

Time change related to a delayed reflection

B.P. Harlamov
IPME, RAS, St.Petersburg, Russia
b.p.harlamov@gmail.com

This work is devoted to a semi-Markov approach to the problem of reflection of a diffusion process from a boundary of its range of values. The semi-Markov approach consists in treating the process with the help of a family of the first exit points from neighborhoods of such arising points. This approach has some advantage comparative to the classic approach. For example, it gives a more economic and informative way for computer modeling the Wiener process. The main attainment of this approach is that it calls attention to a class of continuous semi-Markov processes which are only defined to have the Markov property with respect to such first exit times (Harlamov, 2008).

Every continuous strictly Markov process is semi-Markov, but not vice versa. There exist transformations of continuous strictly Markov processes which lose the Markov property, but safe the semi-Markov one. The delayed reflection is such a transformation. Truncation of an one-dimensional continuous strictly Markov process on the boundary of some interval gives a simple example of such a reflection. In this case the delay is represented by intervals which fall on time when non-truncated process goes behind the boundary.

The semi-Markov approach permits to describe all the class of possible reflections of the Markov diffusion process from a boundary of its range of values (Harlamov, 2007). For example, if we consider a diffusion process on the positive half-line, reflected from zero, then for any $r > 0$ family of distributions of the first exit time from the one-sided neighborhood of zero $[0, r)$ dives all the variants of semi-Markov reflections. Laplace transformation of this distribution is being found as a solution of an ordinary differential equation of Bernoulli type with coefficients, determined by the original Markov process. All the set of possible reflections follows from the set of arbitrary constants of this equation (which depend on the parameter of Laplace transformation).

This set of solutions has an edge point, where the arbitrary constant is equal to zero identically. This solution corresponds to the instantaneous reflection.

In the present work a time change, which transforms the process with instantaneous reflection into the process with some positive meaning of the arbitrary constant, is treated. A representation of the time change in terms of the original locally Markov process is obtained. Laplace transformation of distribution of the difference $\sigma_r - \sigma_r^o$ is derived, where σ_r is the first exit time from $[0, r)$ for the process with delayed reflection, and σ_r^o is that for the corresponding process with instantaneous reflection. It is proved that if the reflected process preserves the global Markov property then this difference is distributed exponentially. Application of the semi-Markov model of delayed reflection in chromatography is discussed in work (Harlamov, 2012).

Keywords: diffusion, Markov, continuous semi-Markov process

Acknowledgements: This work is supported by grant RFBR 12-01-00457-a

References

Harlamov B.P. (2007) Diffusion process with delay on edges of a segment. *Zapiski nauchnyh seminarov POMI,* Vol. 351, pp. 284–297 (in Russian).

Harlamov B.P. (2008): *Continuous semi-Markov processes.* ISTE & Wiley, London.

Harlamov B.P. (2012) Stochastic model of gas capillary chromatography. *Communication in Statistics - Simulation and Computation.* Vol.41, Issue 7, pp. 1023–1031.

Multiplicative Methods of Computing D-Optimal Stratified Experimental Designs

Radoslav Harman
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
harman@fmph.uniba.sk

Suppose that we intend to perform an experiment consisting of a series of independent trials. Let $\mathcal{X}$ be a finite design space, and, for each design point $x \in \mathcal{X}$, let the real-valued observation $y(x)$ satisfy the linear regression model $y(x) = f'(x)\beta + \epsilon(x)$, where $f(x)$ is a known regression vector, β is an unknown parameter, and $\epsilon(x)$ is a random error with $E(\epsilon(x)) = 0$, $Var(\epsilon(x)) \equiv \sigma^2 < \infty$.

An approximate experimental design of the linear regression model is a probability measure ξ on $\mathcal{X}$. For any $x \in \mathcal{X}$ the value $\xi(x)$ represents the proportion of the trials to be performed under the experimental conditions x.

Let $\mathcal{X}_1, ..., \mathcal{X}_k$ be a decomposition of $\mathcal{X}$ into non-overlapping partitions. Let $s_1, ..., s_k$ be given positive constants summing to 1. A design ξ on $\mathcal{X}$ will be called stratified, if it allocates the proportion s_j of trials to the partition $\mathcal{X}_j$ for all $j = 1, ..., k$, i.e., $\xi(\mathcal{X}_1) = s_1$, ..., $\xi(\mathcal{X}_k) = s_k$. The most important special cases of stratified designs correspond to marginally restricted designs, where k is equal to the number of levels of one factor (e.g., Cook and Thibodeau 1980). A stratified design ξ^* is called D-optimal if it maximizes the determinant of the information matrix for β.

The aim of the contribution is to propose two multiplicative methods for computing D-optimal stratified designs. The first one is a generalization of the multiplicative re-normalization heuristic suggested in Martín-Martín (2006), which turns out to be rapid in most test problems, although its theoretical convergence properties are unclear. The second proposed algorithm, which we call barycentric, usually requires more

iterations to compute a design of given efficiency, but we can guarantee its monotonicity and convergence. The barycentric algorithm is derived by a systematic approach from the multiplicative algorithm for the so-called generalized problem of D-optimality, or $D_{\mathcal{H}}$-optimality (Harman and Trnovská 2009). Importantly, the deletion rules developed using the $D_{\mathcal{H}}$-optimality approach can be used with the re-normalization heuristic as well as the barycentric algorithm, making them significantly more efficient. Both methods are very simple to implement in matrix-based programming languages such as R or Matlab.

Keywords: stratified design, marginal restrictions, D-optimal design, multiplicative algorithm, barycentric algorithm.

Acknowledgements: This work was supported by the VEGA grant 1/0163/13.

References

Cook R.D., Thibodeau L.A. (1980): Marginally Restricted D-Optimal Designs, *Journal of the American Statistical Association*, Vol. 75, N. 370, pp. 366-371.

Martín-Martín R., Torsney B., López-Fidalgo J. (2007): Construction of marginally and conditionally restricted designs using multiplicative algorithms, *Computational Statistics & Data Analysis*, Vol. 51, Issue 12, pp. 5547-5561.

Harman R., Trnovská M. (2009): Approximate D-optimal designs of experiments on the convex hull of a finite set of information matrices, *Mathematica Slovaca*, Vol. 56, Issue 6, pp. 693-704.

Theory and algorithm for clustering rows of a two-way contingency table

Chihiro Hirotsu
Meisei University
Hirotsu@ge.meisei-u.ac.jp

Shoichi Yamamoto
RPM Co., Ltd.
s-yamamoto@rpmedical.co.jp

An overall goodness of fit chi-square test for independence is a well known approach to a contingency table. It cannot, however, give any detailed information on the association between the rows and columns. Therefore several multiple comparison approaches have been proposed but the method based on one degree of freedom chi-squared statistic is less informative and the result of the analysis is often difficult to interpret since the degrees of freedom for interaction is usually so large. Therefore the row- and/or column-wise multiple comparisons have been proposed in Hirotsu (1983) and verified to be useful in several occasions as compared with other multiple comparison approaches, see Greenacre (1988) and Hirotsu (1991, 1993). The multiple comparison procedure proposed is essentially the Scheffe type based on the generalized squared distances among rows and the reference distribution is that of the largest root of the Wishart matrix. It is usually easy to obtain and interpret those significant clusters when the number of rows is small, say, up to 10. However, if it is more than 10, we need some stopping rule working automatically for obtaining significant classification of the reasonable number of clusters. One of the purposes of the present paper is therefore to propose such a stopping rule.

An interesting extension of the method is to the one-way layout with the ordered categorical responses. In this case the procedure is essentially unchanged excepting the definition of the squared distance

reflecting the natural ordering and the related asymptotic distribution. Then the reference distribution becomes that of the largest root of a non-orthogonal Wishart matrix which is very difficult to handle. The usual normal approximation is quite unsatisfactory especially when the first and the second largest roots are close each other. Therefore a - approximation has been proposed as a more reasonable one in Hirotsu (2009). The proposed algorithm includes also the calculation of those reference distributions. Essentially the same approach can be applied also to the two-way ANOVA model and gives an interesting procedure for the profile analysis of repeated measures.

Keywords: block interaction model, moderately large table, non-orthogonal Wishart distribution, ordered categories, row-wise multiple comparisons.

References

Greenacre, M.J. (1988): Clustering the rows and columns of a contingency table, *J. Classification*, 5, pp. 39-51.

Hirotsu, C. (1983): Defining the pattern of association in two-way contingency tables, *Biometrika*, 70, pp. 579-589

Hirotsu, C. (1991): An approach to comparing treatments based on repeated measures, *Biometrika*, 78, pp. 583-594.

Hirotsu, C. (1993): Beyond analysis of variance techniques : Some applications in clinical trials, *International. Statistical. Review*, 61, pp. 183-201.

Hirotsu, C. (2009): Clustering rows and/or columns of a two-way contingency table and a related distribution theory, *Computational Statistics and Data Analysis*, 53, pp. 4508-4515.

Robust monitoring of CAPM portfolio betas

M. Hušková
Charles University in Prague, Czech Republic,
huskova@karlin.mff.cuni.cz

O. Chochola, Z. Prášková
Charles University in Prague, Czech Republic,
chochola@karlin.mff.cuni.cz, praskova@karlin.mff.cuni.cz

J. Steinebach
University of Cologne, Germany,
jost@math.uni-koeln.de

The talk will concern robust sequential procedures for the detection of structural breaks in capital asset pricing models (CAPM). Most of the existing procedures for these models are based on ordinary least squares (OLS) estimates. Here we present a class of procedures based on M-estimates and related partial weighted sums of M-residuals.
The theoretical results (limit behavior) will be accompanied with a simulation study that compares the proposed procedures with those based on OLS estimates. An application to real data set will be also presented.
The results will be presented by Hušková.

*Keywords:*capital asset pricing models, on-line test procedures, robust procedures

Testing overdispersion in a mixture model

Maria Iannario
Department of Political Sciences, University of Naples Federico II, Italy.
maria.iannario@unina.it

A practical problem with large scale survey data is the possible presence of overdispersion. It occurs when data display more variability than predicted by the variance-mean relationship for the assumed sampling model. It can be due to design effects, hidden clusters, interviewer effects, number of modalities for each response in rating context or, more generally, to the absence of relevant predictors in the model. Whatever be the underlying cause, overdispersion can be represented either by a positive correlation between the responses or by variation in the response probabilities.

This paper describes some alternative tests for detecting overdispersion in the context of a mixture of discrete random variables denoted as CUBE (Iannario, 2012):

$$Pr(R = r) = \pi\, \beta_r(\xi, \phi) + (1 - \pi)\, U_r\,, \quad r = 1, 2, \ldots, m\,, \qquad (1)$$

where $\beta_r(\xi, \phi)$ specifies the Beta Binomial and U_r the Uniform distributions, respectively.

This mixture has been successfully implemented for ordinal data characterized by some interpersonal overdispersion and encompasses other models, as CUB model (Piccolo, 2003; Iannario and Piccolo, 2012), for instance. Moreover, the mixture has been also extended with the inclusion of predictors, implemented in cases of design effects and scale usage heterogeneity.

In this framework, given sample data, we are able to check if a CUB model, for which is true $H_0 : \phi = 0$, is more adequate than a CUBE model, for which is true $H_1 : \phi > 0$.

Thanks to the asymptotic theory for maximum likelihood estimators (MLE) we may build, under specific conditions, the log-likelihood ratio

(LRT), the Wald, and the Score tests. These statistics require different information: a LRT requires the computation of both full and restricted MLE; instead Wald and Score test require only computation of full and restricted MLE, respectively. In addition, all of them should be carefully applied since we are testing a borderline hypothesis, and thus the asymptotic distribution of these statistics does not converge to the standard χ^2 random variable.

However, previous tests do not consider the possible conditioning of the other parameters in the estimation of ϕ. Thus, we introduce a *generalized likelihood ratio* statistic based on the profile log-likelihood function which allows to avoid the conditioning of *nuisance parameters* $\boldsymbol{\eta} = (\pi, \xi)'$.

In our problem, a natural statistic is the log ratio of maximized likelihoods:

$$W_p(\phi_0) = 2 \left[\ell(\hat{\phi}, \hat{\eta}) - \ell(\phi_0, \hat{\eta}_{\phi_0}) \right],$$

where $\ell(\hat{\phi}, \hat{\eta})$ is computed on the basis of the full ML estimates where $\ell(\phi_0, \hat{\eta}_{\phi_0})$ is computed on the basis of the restrictive model. Under common conditions, this statistic converges to a χ^2_1 distribution.

In this paper several simulation studies are conducted to investigate the empirical performance of all these statistics for varying values of the parameters of the maintained model.

Keywords: Overdispersion, CUBE model, Asymptotic tests, Profile log-likelihood.

References

Iannario M. (2012) CUBE models for interpreting ordered categorical data with overdispersion, Quad. Stat., **14**, 137–140 (2012)

Iannario M., Piccolo D. (2012) CUB models: Statistical methods and empirical evidence, in: Kenett R. S. and Salini S. (Eds.), *Modern Analysis of Customer Surveys: with applications using R*. Chichester: J. Wiley & Sons, pp. 231–258.

Piccolo D. (2003) On the moments of a mixture of uniform and shifted binomial random variables, Quad. Stat., **5**, 85–104.

Numerical studies of space filling designs: optimization algorithms and subprojection properties

Bertrand Iooss
EDF R&D
bertrand.iooss@edf.fr

Guillaume Damblin, Mathieu Couplet
EDF R&D
guillaume.damblin@edf.fr, mathieu.couplet@edf.fr

Quantitative assessment of the uncertainties tainting the results of computer simulations is nowadays a major topic of interest in both industrial and scientific communities. One of the key issues in such studies is to get information about the output when the numerical simulations are expensive to run (Fang et al., 2006). In this communication, we consider the problem of exploring the whole space of variations of the computer model input variables in the context of a large dimensional spaces, that is building initial exploratory designs of numerical experiments, called "Space Filling Designs". In our targeted applications, while the number of points in the design is of the order one hundred, the design dimensions range from two to fifty.

Various properties of Space Filling Designs are studied and justified: interpoint-distances, L^2-discrepancies, Minimal Spanning Tree (MST) criteria. A specific class of design, the optimized Latin Hypercube Sample, is then considered. Several optimization algorithms, coming from the literature, are studied via intensive numerical tests in terms of convergence. For comparison studies, we argue that MST criteria are preferable to the well-known interpoint minimum distance criterion. We show that the Enhanced Stochastic Evolutionary (ESE) algorithm (Jin et al., 2005) converges more rapidly towards good solutions than the classical simulated annealing algorithm (Morris & Mitchell, 1995).

Another contribution of this communication is the deep analysis of the space filling properties of the design 2D-subprojections: among the tested designs (LHS, maximin LHS, several L^2-discrepancy optimized LHS, Sobol' sequence), only the centered and wrap-around discrepancy optimized LHS prove to be robust in high dimension. For example, the other tested designs with one hundred points are no longer robust when their dimension is larger than 10. Moreover, centered-discrepancy optimized LHS are generally more regular than the wrap-around discrepancy optimized LHS.

Keywords: discrepancy, space fillling, Latin Hypercube Sampling, computer experiment.

Acknowledgements: Part of this work has been backed by French National Research Agency (ANR) through COSINUS program (project COSTA BRAVA noANR-09-COSI-015).

References

Fang K-T., Li R., Sudjianto A. (2006): *Design and modeling for computer experiments*, Chapman & Hall/CRC.

Jin R., Chen W., Sudjianto A. (2005): An efficient algorithm for constructing optimal design of computer experiments, *Journal of Statistical Planning and Inference*, Vol. 134, pp 268-287.

Morris M.D., Mitchell T.J. (1995): Exploratory designs for computationnal experiments, *Journal of Statistical Planning and Inference*, Vol. 43, pp 381-402.

Et tu "Brute Force"? No! A Statistically-Based Approach to Catastrophe Modeling

Mark E. Johnson
Dept. of Statistics, University of Central Florida
mejohnso@mail.ucf.edu

Charles C. Watson, Jr.
Watson Technical Consulting
cwatson@methaz.org

Large scale simulations of natural catastrophes such as hurricanes have been widely used since the 1980's to assess the expected losses from extreme events. A more direct statistical approach is described here that involves simulating only the events in the historical record and then fitting probability distributions to the impacts on a site specific basis. This approach eliminates the possibility of generating unrealistic events while employing the historical data in a coherent and logically-consistent fashion. Special attention is paid to the upper percentiles of the distributions of wind, damage and loss as this region is of paramount interest in the re-insurance context.

Keywords: hurricanes, Weibull distribution, tropical cyclones, cross-validation.

References

Johnson, M. E., Watson, C. C., Jr. (2007): Fitting Statistical Distributions to Data in Hurricane Modeling, *American Journal of Mathematical and Management Sciences*, Vol. 27.

Johnson, M.E., Watson, C.C., Jr. (1999): Hurricane Return Period Estimation, 10th Symposium on Global Change Studies, Dallas, TX, American Meteorological Society, pp. 478-479.

Watson, C. C., Jr., Johnson, M. E. (2004): Hurricane Loss Estimation Models: Opportunities for Improving the State of the Art, *Bulletin of the American Meteorological Society*, Vol. 85, pp. 1713-1726.

Monte Carlo modeling in non-stationary problems of laser sensing of scattering media

Kargin B.A.
Novosibirsk State University
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch RAS
bkargin@osmf.sscc.ru

Kablukova E.G.
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch RAS
Jane_k@ngs.ru

Non-stationary problems of laser sensing are of great interest in connection with the wide practice of using laser sensors (LIDARs) placed on the ground, on aircraft and in space to quickly diagnose aerosol contaminants in the atmosphere, the space-time transformation of aerosols' microphysical properties and various types of cloud particles, to determine remotely the hydrophysical properties of the ocean, to discover the upper and lower boundaries of cloudiness and to solve a variety of other problems of optical remote sensing of the natural medium. It's possible to get acquainted in more detail with a list of certain modern physical problems' statements for problems of laser sensing of the atmosphere and ocean and corresponding methods of statistical modeling in the quite recent review [1]. The justification of using the Monte-Carlo methods as well as elaboration of corresponding algorithms for solving non-stationary problems of transfer theory of optical radiation in scattering and absorbing media have been performed quite a long time ago in a series of early works, which were summarized in [2]. The laser sensing problems under consideration are different from many other problems of atmosphere optics because of the complex boundary conditions related to the finite size of the source radiation beam and small phase volume of the detector as well as the principally non-stationary character of the

transfer process being modeled. This condition is responsible for typical requirements to the technique of statistical modeling and determines the necessity of using local estimates which, though labor-intensive, are the only possible method of calculating sought for radiation properties registered by a detector with a small phase volume. Such problems belong to the class of so-called "big" problems of mathematical physics and for their solution serious computing resources, including multiprocessor computer systems, are required. That's why one topical issue in solving practical problems of remote laser sensing of natural media by the Monte-Carlo method is the problem of reducing an algorithm's "the cost of computation". For this issue a new optimization of local estimates by means of integrating based upon a discrete-stochastic version of "importance sampling" and using a variation of the "splitting" method is presented in the paper. The computation formulae are presented in detail, including those used to estimate optimal splitting parameters and the node grid for discrete-stochastic integrating of local estimates. Some results of a great series of numerical calculation of a laser signal reflected off a cloud layer in dependence on algorithm optimization parameters are presented.

Keywords: Monte Carlo, transport theory, local estimates, laser sensing, atmospheric optics.

Acknowledgements: This work has been done under the financial support of the RFFR (grants 12-01-00034 and 13-01-00032), Integrational Project SB RAS No. 52, RAS Presidium Project No. 15.9-1-1, DMS RAS Project No. 1-3-3-2.

References

1 Krekov G.M.(2007): Monte Carlo method in problems of atmospheric optics, */Atmospheric and Oceanic*, Vol. 20, No. 09, pp. 757-766.

2 Marchuk G.I., Mikhailov G.A., Nazaraliev M.A., Darbinian R.A., Kargin B.A., Elepov B.S.(1980): Monte Carlo Methods in Atmospheric Optics, */SpringerVerlag*, -208p.

Adjusting for selection bias in single-blinded randomized controlled clinical trials

Lieven N. Kennes
RWTH Aachen University
lkennes@ukaachen.de

Selection bias affects the evaluation of clinical trials (Berger, 2005). In absence of double-blinding, the investigator is able to predict future allocation based on past assignments yielding the possibility to cause inhomogeneous treatment groups. To quantify the impact of selection bias in a clinical trial, type one error rate elevation is used (Proschan, 1994). Even in randomized, but single-blinded, controlled clinical trials, serious type one error rate elevation can occur, especially unter permuted block randomization (Kennes, 2011). Based on the principle of maximum likelihood (ML), a test adjusting for selection bias as well as confidence intervals for the true treatment difference and the selection effect are derived. The established ML-estimators are consistent and asymptotic normal (Kennes, 2013). Formulas for confidence intervals for all parameters are given. The confidence interval for the selection effect can be used to detect selection bias. Results of simulation studies illustrate that while the t-test exceeds the nominal significance level in presence of selection bias, the ML-test corrects for selection bias and asymptotically holds the nominal significance level.

Keywords: Selection bias, randomized controlled clinical trial, blinding, random walk, maximum likelihood.

References

Berger, V. W. (2005): *Selection Bias and Covariate Imbalances in Randomized Clinical Trials*, Wiley, Chichester.

Kennes, L. N., Cramer, E., Hilgers, R. D., Heussen, N. (2011): *The Impact of Selection Bias on Test Decisions in Randomized Clinical Trials*, Statistics in Medicine, Vol. 30, pp. 2573-2581.

Kennes, L.N. (2013): *The Effect of and Adjustment for Selection Bias in Randomized Controlled Clinical Trials*, Dissertation, RWTH Aachen University.

Proschan, M. (1994): *Infuence of Selection Bias on Type 1 Error Rate under Random Permuted Block Designs*, Statistica Sinica, Vol. 4, pp. 219-231.

Asymptotic Permutation Tests and Confidence Intervals for Paired Samples

Frank Konietschke
University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany
fkoniet@gwdg.de

In many psychological, biological, or medical trials, data are collected in terms of a matched pairs design, e.g. when each subject is observed repeatedly under two different treatments or time points. When the normality assumption of the data (or of the differences of the paired observations) is violated, e.g. in case of skewed distributions or even ordered categorical data, nonparametric (ranking) procedures, which use ranks over all dependent and independent observations, are preferred for making statistical inferences (Munzel, 1999b).

Most of these approaches, however, are restricted to testing problems and cannot be used for the computation of confidence intervals for the treatment effects. Particularly, different variances of the paired observations occur in a natural way, e.g. when data are collected over time. The derivation of nonparametric procedures, which allow the data to have different variances or shapes even under the null hypothesis, is a challenge.

We study various bootstrap and permutation methods for matched pairs, whose distributions can have different shapes even under the null hypothesis of no treatment effect. Although the data may not be exchangeable under the null hypothesis, we investigate different permutation approaches as valid procedures for finite sample sizes. In particular, we derive the limit of the studentized permutation distribution under alternatives, which can be used for the construction of $(1-\alpha)$-confidence intervals. Simulation studies show that the new approach is more accu-

rate than its competitors.

The procedures are illustrated using a real data set.

Keywords: confidence intervals, permutation tests, heteroscedastic designs, resampling.

References

Konietschke, F. and Pauly, M. (2012): A studentized permutation test for the non-parametric Behrens-Fsher problem in paired data. *Electron. J. Stat.* 6, pp. 1358-1372.

Konietschke, F. and Pauly, M. (2012): Bootstrapping and Permuting paired t-test type statistics. *Statistics and Computing*, DOI 10.1007/s11222-012-9370-4.

Munzel, U. (1999): Nonparametric methods for paired samples. *Statistica Neerlandica* 53, pp. 277-286.

Monte Carlo methods for reconstructing a scattering phase function from polarized radiation observations

Korda A.S., Ukhinov S.A.
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk State University, 630090, Novosibirsk, Russia
asc@osmf.sscc.ru, sau@sscc.ru

Atmospheric optics considers the problem of reconstructing the scattering phase function of the atmosphere, $g(\mu)$, by using ground-based observations of sky brightness in the solar almucantar, i.e., in various directions that make the same angle θ_s with the zenith as the line of sight to the Sun. In the single-scattering approximation, the observed brightness values are proportional to the corresponding phase function values. Hence, to estimate the phase function, one can use an iterative algorithm in which the contribution of single scattering to the observed brightness is successively refined by mathematical modeling. A number of iterative algorithms were constructed previously. In these algorithms, mathematical modeling is used to successively refine the phase function values by using available information about the angular distribution of brightness on the underlying surface under the assumption that the contribution to brightness of the single-scattered radiation is rather large. Two such algorithms were extended to the case of polarized radiation, and a new algorithm of this type was constructed.

The objective of this study is to numerically substantiate the convergence of these methods. For this purpose, an algorithm of Jacobi matrices calculation for the iteration operators of the methods was developed, and calculations were carried out for various parameters of the atmosphere.

The results of calculations of spectral radii of Jacobi matrices explain the numerical results of the phase function reconstruction. For

instance, a method converges if the spectral radius of the corresponding Jacobi matrix for the same parameters of the atmosphere is less than unity, and a method diverges if the spectral radius is greater than unity. The comparison of the results of different methods with different parameters of atmosphere shows, that method converges faster when corresponding spectral radius is smaller.

Also a study of the influence of measurement errors on the reconstruction of the scattering matrix was carried out. Test calculations showed the stability of algorithms to errors in the initial data.

Keywords: Monte Carlo methods, polarized radiation, inverse problems.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (13-01-00441, 12-01-00034, 12-01-00727, 12-01-31328), and by MIP SB RAS (A-47, A-52)

References

Mikhailov G.A., Ukhiniov S.A., Chimaeva A.S. (2009): Monte Carlo Algorithms for Reconstruction of the Scattering Indicatrix Adjusted for Polarization, *Russian Journal of Numerical Analysis and Mathematical Modelling*, Vol. 24, N. 5, pp. 455-465.

Ukhinov S.A., Chimaeva A.S. (2011): Convergence of Monte Carlo Algorithms for Reconstructing the Scattering Phase Function with Polarization, *Numerical Analysis and Applications*, Vol. 4, N. 1, pp. 81-92.

Monte Carlo Algorithm for Simulation of the Vehicular Traffic Flow within the Kinetic Model with Velocity Dependent Thresholds

M.A. Korotchenko
Institute of Computational Mathematics and Mathematical Geophysics
SB RAS, prospect Akademika Lavrentjeva, 6, Novosibirsk, Russia, 630090
kmaria@osmf.sscc.ru

We consider an acceleration oriented vehicular traffic flow (VTF) model, which was proposed in (Waldeer, 2003a). A special feature of this model is introduction of the acceleration variable into the set of phase coordinates, which describe the state of a vehicle, together with its velocity and spatial coordinate, traditionally used in kinetic models. In contrast to the gas dynamics, the interactions in the system result not in velocity, but in acceleration jumps. Such modification of the phase space allowed in (Waldeer, 2003a) to extend this kinetic model onto a wider class of the VTFs, as it also adequately describes the case of partly constrained flows, and denser VTFs.

For spatial homogeneous case we use the interaction profiles based on velocity dependent thresholds, which were introduced in (Waldeer, 2003b, 2004). In such profiles an interaction occurs only if the distance between interacting vehicles is equal to one of the threshold distances, whicg depend on the velocity of the follower. On each of these thresholds an individual acceleration change occurs for the follower.

For the original probabilistic VTF model we construct an integral equation of the second kind, which is related to a linear N-particle model describing the vehicle system evolution. We also propose Monte Carlo algorithms for estimating the functionals of the solution to the obtained equation. Analogous algorithms were constructed in (Burmistrov & Korotchenko, 2011) for simpler interaction profiles.

The practical suitability of this approach to the solution of traffic

problems is demonstrated by numerical experiments in which we estimate various functionals, such as velocity and acceleration distribution, the vehicle density dependence of the traffic density (so called fundamental diagram), mean velocity, velocity scattering, and acceleration noise.

Keywords: Evolution of Many-Particle System, Acceleration Jump Process, Interaction Profile, Fundamental Diagram.

Acknowledgements: This work was partly supported by the Russian Foundation for Basic Research (grants 11-01-00252, 12-01-31134) and Siberian Branch of the Russian Academy of Sciences (Interdisciplinary Integration Grant No. 47).

References

Waldeer K.T. (2003a): The Direct Simulation Monte Carlo Method Applied to a Boltzmann-Like Vehicular Traffic Flow Model, *Computer Physics Communications*, Vol. 156, N. 1, pp. 1-12.

Waldeer K.T. (2003b): Vergleich der Ergebnisse eines beschleunigungsorientierten, Boltzmannartigen Verkehrsflußmodells mit Messungen, *Proceedings of 19th Dresden Conference on Traffic and Transportation Science*, pp. 84.1-84.16. (in German)

Waldeer K.T. (2004): Numerical Investigation of a Mesoscopic Vehicular Traffic Flow Model Based on a Stochastic Acceleration Process, *Transport Theory and Statistical Physics*, Vol. 33, N. 1, pp. 31-46.

Burmistrov A.V., Korotchenko M.A. (2011): Application of Statistical Methods for the Study of Kinetic Model of Traffic Flow with Separated Accelerations, *Russian Journal of Numerical Analysis and Mathematical Modelling*, Vol. 26, N. 3, pp. 275-293.

Uniform generation of acyclic digraphs and new MCMC schemes via recursive enumeration

Jack Kuipers, Giusi Moffa
University of Regensburg
jack.kuipers@ur.de, giusi.moffa@ukr.de

Directed acyclic graphs (DAGs) are the basic representation of the structure underlying Bayesian networks, widely applied for example in biology and social sciences. Our contribution here is twofold: to uniform generation from the space of DAGs, and to sampling Monte Carlo Markov chain (MCMC) methodology for inference in the context of Bayesian graphical models.

A uniform sample is a crucial requirement in simulation studies for removing any structure related bias. The current methods of choice rely on Markov chain methods (Melançon et al., 2001, Ide and Cozman, 2002). Due to their convergence and computational issues practitioners often need to content themselves instead by sampling on the much simpler space of triangular matrices. Here we consider a method based on the recursive enumeration of DAGs (Robinson, 1970). An analysis of its complexity suggests that it is indeed advantageous with respect to the Markov chain approach. More importantly we propose a fast and highly accurate approximation based on the limiting behaviour of their distribution which allows us to sample arbitrarily large acyclic digraphs.

The slow mixing and convergence issues of the standard MCMC on graphical structure (Madigan and York, 1995) constitute a serious practical limitation to application in high-throughput biological studies, leading to the development of MCMC over orders (Friedman and Koller, 2003) and more recently to a new edge reversal proposal move (Grzegorczyk and Husmeier, 2008). We propose instead an adaptation of the ideas based on recursive enumeration to build a novel MCMC scheme

on the space of DAGs to sample from the posterior distribution of graphs conditional on the data.

Keywords: Acyclic digraphs, Sampling, Graphical models, Bayesian networks, Markov Chain Monte Carlo.

References

Friedman N. and Koller D. (2003): Being Bayesian about Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, Vol. 50, pp. 95–125

Grzegorczyk M. and Husmeier D. (2008): Improving the Structure MCMC Sampler for Bayesian Networks by Introducing a New Edge Reversal Move. *Machine Learning*, Vol. 71, pp. 265–305

Ide J. S. and Cozman F. G. (2002): Random Generation of Bayesian Networks *Brazilian Symposium on Artificial Intelligence*, pp. 366–375

Madigan D. and York J. (1995): Bayesian Graphical Models for Discrete Data. *International Statistical Review*, Vol. 63, pp 215–232

Melançon G., Dutour I. and Bousquet-Mélou M. (2001): Random Generation of Directed Acyclic Graphs. *Electronic Notes in Discrete Mathematics* Vol. 10, pp. 202-207

Robinson R. W. (1970): Enumeration of Acyclic Digraphs. *Proceedings of the second Chapel Hill conference on combinatorial mathematics and its applications*, pp. 391–399, Chapel Hill, 1970 University of North Carolina

Two-Stage Adaptive Optimal Design with Fixed First Stage Sample Size

Adam Lane and Nancy Flournoy
University of Missouri
aclpp9@mail.missouri.edu, nflournoy@missouri.edu

An adaptive optimal design uses the data from all previous stages to estimate the locally optimal design of the current stage. Many, including Box and Hunter (1965), Fedorov (1972), White (1975), Silvey (1980) and others have suggested using such designs. Recently, Dragalin et al. (2008), Lane et al. (2012), Yao and Flournoy (2010). have investigated the properties and performance of these procedures.

Asymptotics for regular models with a fixed number of stages are straightforward if one assumes the sample size at each stage goes to infinity with the overall sample size. However, it is not uncommon for a small pilot study of fixed size to be followed by a much larger experiment. We study the large sample behavior of two-stage studies under this scenario. For simplicity, we assume

$$y_{ij} = \eta(x_i, \theta) + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right),\ i = 1, 2,\ j = 1, \ldots, n_i, \quad (1)$$

where n_i and x_i are the sample size and treatment level for the i^{th} stage, respectively and $\eta(x, \theta)$ is some nonlinear mean function. The first stage treatment level, x_1, is considered fixed. However, the second stage treatment level will be selected using an adaptive optimal procedure. The total sample size $n = n_1 + n_2$.

For model 1 we show that the distribution of the maximum likelihood estimates (MLE) converges to a scale mixture family of normal random variables, i.e.,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathscr{D}} UQ$$

as $n_2 \to \infty$, where $\hat{\theta}_n$ is the MLE, $Q \sim \mathcal{N}(0, \sigma^2)$ and $U = \left(\frac{d\eta(x_2,\theta)}{d\theta}\right)^{-1}$ is a random function of the mean of the first stage data. We compare the behavior of these estimates with those obtained from the normal distribution that results when samples from both stage are large.

Keywords: adaptive design, optimal design, pilot study, scale mixtures.

References

Box, G. and Hunter, W. (1965): Sequential Design of Experiments for Nonlinear Models. Proceedings of the Scientific Computing Symposium: Statistics. White Plains: IBM.

Fedorov, V. (1972): Theory of Optimal Experiments. Academic Press. Adademic Press. New York.

Dragalin, V., Fedorov, V. and Wu, Y. (2008): Adaptive Designs for Selecting Drug Combinations Based on Efficacy-Toxicity Response. *Journal of Statistical Planning and Inference*. Vol. 2, pp. 352-373.

Silvey, S. (1980): Optimal Design: An Introduction to the Theory for Parameter Estimation. Chapman and Hall. London.

Lane, A., Yao, P. and Flournoy, N. (2012): Information in a Two-Stage Adaptive Optimal Design. Isaac Newton Institute for Mathematical Sciences. Preprint Series,
www.newton.ac.uk/preprints.html.

Yao, P. and Flournoy, N. (2010): Information in a Two-Stage Adaptive Optimal Design for Normal Random Variables Having a One Parameter Exponential Mean Function. In Giovagnoli, A., Atkinson, A.C., Torsney, B. and May, C. *MoDa 9 – Advances in Model-Oriented Design and Analysis*. Springer. pp. 229-236.

Optimal Bayesian designs for prediction in deterministic simulator experiments

Erin R. Leatherman, Thomas J. Santner
The Ohio State University
leatherman.33@osu.edu, tjs@stat.osu.edu

Angela Dean
University of Southampton and The Ohio State University
amd@stat.osu.edu

The use of deterministic simulators as experimental vehicles has become widespread in applications such as engineering, biology, and physics. Such simulators are based on complex mathematical models which describe the relationship between the input and output variables in a physical system. One use of a computer simulator is for prediction; given sets of system inputs, the simulator is run to find the corresponding predicted outputs of the system. However, when the mathematical model is complex, a simulator can be computationally expensive, taking many hours or days to produce a single output. In this case, a cheaper statistical metamodel (emulator) is often used to make predictions of the system outputs.

This talk introduces a design criterion for the construction of designs which lead to metamodels that give accurate predictions over the input space. The criterion minimizes the Bayesian Integrated Mean Squared Prediction Error (BIMSPE) which is calculated assuming a hierarchical Bayesian model for the outputs of the simulator. Comparisons of the prediction accuracies of BIMSPE-optimal designs, IMSPE-optimal designs and space-filling designs will be presented via examples. It will be shown that, as measured by root mean squared prediction error, BIMSPE-optimal designs tend to outperform the space-filling Latin hypercube and minimum Average Reciprocal distance designs over a wide class of response surfaces, while not requiring unknown model

parameters to be specified.

Keywords: Bayesian design criterion, computer experiment, deterministic simulator, prediction error, process-based estimation

Acknowledgements: This research was sponsored, in part, by the National Science Foundation under Agreement DMS-0806134 (The Ohio State University).

Modeling the anesthesia unit and surgical wards in a Chilean hospital using Specification and Description Language (SDL)

Jorge Leiva Olmos
Head of the Department of Data Administration, Information and Communication. Hospital Dr. Gustavo Fricke. Chile
jleivahgf@gmail.com

Pau Fonseca i Casas, Jordi Ocaña
Department of Statistics and research operative Polytechnic University of Catalonia and Department of Statistics, University of Barcelona
pau@fib.upc.edu,jocana@ub.edu

This work addresses the problem of performing a formal modeling of the processes related to the anesthesia unit and surgical wards (UAPQ) of a Chilean hospital. Modeling was performed using the Specification and Description Language (SDL), specifically SDL/GR , using Microsoft Visio as a tool to represent the diagrams. The channels of information, signals, directionality and hierarchy between the different processes have been defined. The full definition of the characteristics of the system allowed a clear identification of relationships between system elements. The design includes protocols for administrative activities such as those associated with support units, including those tasks and knowledge (often tacit) transmitted through informal channels. This complete model representation and the modular model definition allows continuing different research lines, such as simulating the UAPQ processes or studying the optimization of some of their processes. In summary, the study allowed the creation of clear documentation and an understanding of the processes of Anesthesia and surgical wards by clinical and administrative staff, through a graphic model.

Keywords: Modeling language, SDL, Surgical ward

References

Brade D. (2000): Enhancing modeling and simulation accreditation by structuring verification and validation results, *Winter Simulation Conference*.

Chile Ministry of Health (2009): Compromisos de Gestin, *Via del Mar, Chile*.

Fonseca i Casas, Pau (2008): SDL distributed simulator, *Winter Simulation Conference. Miami: INFORMS*.

Reed Rick (2000): SDL-2000 form New Millenium, *Systems Telektronikk 4.2000*, pp. 20-35.

Sargent Robert (2007): Verification and validation of simulation models, *Winter Simulation Conference. Washinton: IEEE*.

[6] Telecommunication standardization sector of ITU (1999): Specification and Description Language (SDL), Series Z, *http://www.itu.int/ITU-T/studygroups/com17/languages/index.html*.

Simultaneous t-Model-Based Clustering Applied to Company Bankrupt Prediction

Alexandre Lourme
Université Bordeaux 4 & Institut de Mathématiques de Bordeaux (France)
alexandre.lourme@u-bordeaux1.fr

Christophe Biernacki
Université Lille 1 & CNRS (France)
biernack@math.univ-lille1.fr

Mixture models are common cluster analysis tools. When a sample is suspected to harbor homogeneous subgroups a standard mixture model-based clustering method consists in (McLachlan and Peel, 2000): (a) modeling this sample with a finite mixture of K parametric components (b) estimating the mixture parameter by Maximum Likelihood thanks to an Expectation-Maximization algorithm (Dempster et al. 1977) and (c) allocating each data point by maximum a posteriori to the component corresponding to the highest conditional probability.

Here, we are interested in numerous cases where not only one, but several samples $\mathbf{x}^1, \ldots, \mathbf{x}^H \subset \mathbb{R}^d$ have to be clustered (i) the H samples being described by a common set of variables (ii) the statistical units being of same nature in $\mathbf{x}^1, \ldots, \mathbf{x}^H$ and (iii) K-class partitions with identical interpretations being expected in all samples. For instance heterogeneous populations evolving over time can provide such samples.

In this context Lourme and Biernacki (2013) displays a simultaneous clustering method based on a Gaussian assumption: the data from each group k ($k \in \{1, \ldots, K\}$) in $\mathbf{x}^h$ ($h \in \{1, \ldots, H\}$) are supposed to be realizations of a conditional random vector $\mathbf{X}^h_{|\mathfrak{G}_k} \in \mathbb{R}^d$ normally distributed. But as Gaussian parameters are sensitive to extreme values, when the data are suspected to include outliers we propose to assume alternatively that each $\mathbf{X}^h_{|\mathfrak{G}_k}$ is distributed according to a d-dimensional Student's t-distribution. Then, The following relationship is assumed

for any h, h' and k:

$$\mathbf{X}^{h'}_{|\mathfrak{G}_k} \overset{\mathscr{D}}{=} \mathbf{D}^{h,h'}_k \mathbf{X}^{h}_{|\mathfrak{G}_k} + \mathbf{b}^{h,h'}_k, \tag{1}$$

$\mathbf{D}^{h,h'}_k \in \mathbb{R}^{d\times d}$ being a diagonal positive definite matrix and $\mathbf{b}^{h,h'}_k$ a vector in $\mathbb{R}^d$ ($\bullet \overset{\mathscr{D}}{=} \bullet$ indicates two random vectors having the same distribution). (1) is a stochastic link between the corresponding conditional populations; it aims to formalize (i), (ii) and (iii) which characterize our clustering context and it enables to learn some information about each sample from the other samples at the inference step.

The simultaneous clustering procedure based on Student's t-mixtures is involved in Lourme and Biernacki (2011) in detecting bankrupt or healthy companies ($K = 2$) among two samples of firms ($H = 2$) differing over the year and described by a common set of six financial ratios ($d = 6$). The longitudinal study highlights a more complex hidden structure than the expected one, distinguishing three groups: two clear groups of healthy and bankrupt companies, but also a third group of firms with unpredictable health.

Keywords: Stochastic linear link, t-mixture, model-based clustering, EM algorithm, model selection.

References

Dempster A.P., Laird N.M., Rubin D.B. (1977): Maximum likelihood from incomplete data (with discussion), *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1-38.

Lourme A., Biernacki C. (2011): Simultaneous t-Model-Based Clustering for Time Dependent Data: Application to a Study of the Financial Health of Corporations, *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, Vol. 4, N. 2, pp. 73-82.

Lourme A., Biernacki C. (2013): Simultaneous Gaussian model-based clustering for samples of multiple origins, *Computational Statistics*, Vol. 28, N. 1, pp. 371-391.

McLachlan G.J., Peel D. (2000): Finite mixture models. Wiley, New York, p. 29.

Random Walks Methods for Solving BVP of some Meta Elliptic Equations

Vitaliy Lukinov
Russian Academy of Sciences ICM&MG SB RAS,
Novosibirsk State University
Vitaliy.Lukinov@ngs.ru

This talk present new Monte Carlo methods for solving BVP of some meta harmonic equations with Dirichlet, Neumann, or mixed boundary conditions. Despite the slow rate of convergence of statistical methods for low-dimensional spaces, in comparison with classical numerical methods, their use is advantageous in finding a solution to a small area or for calculating the statistical characteristics of the solutions with random right-hand sides. In this talk, we consider the meta harmonic equation with random inputs functional parameters. Here, a new scalar walk by spheres estimates of covariance for the solution were constructed by an parametric differentiation of special bounce estimates of solution to special constructed problems. First, this approach was proposed by G. Mikhailov (1999). Besides vector estimates developed by Mikhailov and Tolstolytkin (1994), proposed scalar estimates had been fully investigated: the finiteness of variances has been proved, absolute errors have been evaluated, laboriousness have been estimated, the problem of optimal choice of method parameters to achieve a given error level have been solved. Compared to SVD approach proposed by Sabelfeld (2009) the offered method can to solve a problems with a random spectral parameter. It seems that the SVD approach is less time consuming, but a special comparison of methods was not carried out.

We obtained estimates for the Dirichlet BVP and some special Neumann and mixed BVP. Further investigation is aimed at building a cost-effective methods for mixed and Neumann BVP. This work is mostly theoretical. However, the proposed estimates is easy to extend to the real problems(see Bolotin (1979)) with the own geometry of the bound-

ary.

Keywords: Monte Carlo methods; meta harmonic equations; BVP; 'random walks' algorithms

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (grant N11-01-00252-a) and Novosibirsk State University

References

Bolotin V.V. Random fluctuations of elastic systems. (Moscow.: Nauka, 1979).

Lukinov V.L., Mikhailov G.A.*The probabilistic representation and Monte Carlo methods for the first boundary value problem for a polyharmonic equation. Rus. J. Num. Anal. and Math. Modell. 2004, v. 19, N.5, p. 434-449* .

Mikhailov G.A., Tolstolytkin D.V. *A new Monte Carlo method for calculating the covariance function of the solutions of the general biharmonic equation.* // Dokl. Akad. Nauk – 1994. – V. 338, No 5. – P. 601–603.

Ermakov S.M., Nekrutkin V.V., Sipin A.S. *A random procees for solving a classic equations of mathematical physics.* – Moskow.: Nauka, 1984.

Mikhailow G.A. *Parametric Estimates by the Monte Carlo method* -Utrecht, The Nederlands,1999.

Sabelfeld K.K., Mozartova N. *Sparsified Randomization Algorithms for low rank approximations and applications to integral equations and inhomogeneous random field simulation.* – Monte Carlo Methods Appl. Vol. 15 No. 3 (2009), pp. 16.

Optimization via Information Geometry

Luigi Malagò
Dipartimento di Informatica, Università degli Studi di Milano, Italy
malago@di.unimi.it

Giovanni Pistone[1]
Collegio Carlo Alberto, Moncalieri, Italy
giovanni.pistone@carloalberto.org

A non-parametric version of the Information Geometry (IG) that Amari started in 1982 has been developed in a number of papers beginning with a 1995 paper by Pistone and Sempi. In the finite state space case, the set of all positive densities $\mathcal{P}$ is considered a manifold with tangent space $T_p = \{u\colon \mathrm{E}_p(u) = 0\}$ at $p \in \mathcal{P}$. On each tangent space there is the Rao metric $g_p(u, v) = \mathrm{E}_p(uv)$. If we fix a reference density p and a vector basis $u_1, \ldots, u_d$ of T_p, then all the positive densities are of the exponential form $q = \exp(\sum_j \beta_j u_j - \psi(\beta)) \cdot p$. Both the parameters β and $\eta = \nabla\psi(\beta)$ provide affine charts. In general sample space, a similar construction applies to an exponential family $\mathcal{E}$, when the u_j's are linearly independent and span a subspace of T_p.

The geometrical setting, together with the theory of exponential families, provides convenient tools to compute the derivatives of a real function $F\colon \mathcal{E} \to \mathbb{R}$. For example, the derivative at p in the direction $u \in T_p(\mathcal{E})$ has the form of the natural gradient $\widetilde{\nabla} F(p) \in T_p(\mathcal{E})$ with respect to the Rao metric g_p, $\mathrm{D}_u F(p) = \mathrm{E}_p(\widetilde{\nabla} F(p) u)$, see Amari (1998). The flow of the gradient of F is controlled by the differential equation $\dot{p}(t)/p(t) = \widetilde{\nabla} F(p(t))$.

The IG framework has given good results in optimization, where the problem $\max_x f(x)$ is relaxed to $\max_{p \in \mathcal{E}} \mathrm{E}_p(f)$, $\mathcal{E}$ a parametric exponential family. The relaxed problem is solved using a discretized and sampled version of the flow equation, see Malagò et al. (2011) for the

[1] Presenting author

discrete sample space, Wierstra et al. (2008) and Arnold et al. (2011) for the continuous sample space.

In this talk we discuss the first findings of further research in progress: 1) the natural gradient in the natural parameters β of the exponential family is given by a regression formula $I^{-1}(\beta)\mathrm{Cov}_p(f, U)$, where U is the vector of sufficient statistics and $I(\beta) = \mathrm{E}_\beta(UU^t)$ is the Fisher Information Matrix; 2) the setting in the gaussian case is similar; 3) the discrete approximation of the flow equation can be improved e.g., by considering second derivatives.

Keywords: Exponential families, Information Geometry, Optimization, Gradient Flow, Natural Gradient

References

Amari, S.-I. (1998): Natural gradient works efficiently in learning, *Neural Computation*, Vol. 10, N. 2, pp. 251–276.

Arnold L., Auger A., Hansen N., Ollivier Y. (2011): Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles, arXiv:1106.3708.

Malagò L., Matteucci M., Pistone G. (2011): Towards the geometry of estimation of distribution algorithms based on the exponential family. In: *Proceedings of FOGA '11, Schwarzenberg, Austria*, pp. 230–242.

Wierstra D., Schaul T., Peters J., Schmidhuber J. (2008): Natural evolution strategies. In *Proc. of IEEE CEC 2008*, pp. 3381–3387.

Using coarse-grained and fine-grained parallelization of the Monte Carlo method to solve kinetic equations

Mikhail Marchenko
Novosibirsk State University,
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch RAS
mam@osmf.sscc.ru

Kinetic equations which describe the evolution of the electron showers may be solved numerically by the Monte Carlo method. In the course of the stochastic modeling of the single realization of the electron shower we simulate the individual tracks of the electrons and thereby modelling different stochastic events: electron's emission, scattering, nonelastic interaction or ionization; simulate how other electrons and ions affect the behavior of the given one.

While carrying out simulations of the electron showers in gases, we faced the following problem. By necessity, the number of the time steps in the simulation scheme is quite large and the number of the electrons in a single realization is increasing according to the exponential law. Therefore, after a while the number of the electrons in the shower reaches the value of order 10^8 or even 10^9.

To simulate the behavior of the showers with large number of electrons we implemented several parallelization techniques. For parallel implementation, we used either the MPP based supercomputers (without the accelerators or coprocessors) or the hybrid supercomputers with the accelerators such as NVIDIA CUDA GPUs and Intel Xeon Phi.

The coarse-grained parallelization technique is usually used for the Monte Carlo computations: the simulation of realizations is distributed among processor cores. For the simulations on the MPP based supercomputer (without the accelerators) we used the PARMONC software library (Marchenko, 2011). But using the PARMONC the simulation

was quite time consuming. A way to additionally reduce the time cost was to implement the fine-grained parallelization technique and use the supercomputers with the accelerators. Namely, conditionally independent branches of the shower were simulated on the different cores of the accelerator. In order to balance the workload, the parallel threads on the cores were exchanging test particles in the course of simulation.

In order to simulate the independent branches on each core of the accelerator we used the parallel random numbers generator presented in (Marchenko, Mikhailov, 2007). The generator also enabled us to correlate the results of different computations in order to study the parametric dependence of the stochastic model.

Keywords: Electron showers, stochastic simulation, parallelization, supercomputers.

Acknowledgements: This work was supported by the Interdisciplinary Integration Projects of the Siberian Branch RAS, grants No. 39, 126 and 130; Russian Foundation for Basic Research, grants No. 12-01-00034 and 12-01-00727.

References

Marchenko, M.A., Mikhailov, G.A. (2007): Distributed computing by the Monte Carlo method, *Automation and Remote Control*, Vol. 68, issue 5, pp. 888-900.

Marchenko M. (2011): PARMONC - A Software Library for Massively Parallel Stochastic Simulation, *LNCS*, Vol. 6873, pp. 302-316.

Heuristic Optimization for Time Series Analysis

Dietmar Maringer
Economics and Business Faculty, University of Basel, Peter Merian-Weg 6,
CH-4002 Basel, Switzerland.
dietmar.maringer@unibas.ch

Recent advances in econometrics and statistics have not only brought more sophisticated and demanding models, but have also increased the awareness for numerical issues in empirical applications. For example, model estimation can require the maximization of of some goodness-of-fit measures, but no analytical solution exists, and one has to rely on numerical procedures. These measures are rarely as well-behaved as out-of-the-box optimization algorithms require: Likelihood functions can have local optima, information criteria are not necessarily continuous and smooth functions, and adding constraints on the feasible values for parameters adds another layer of complicatedness. In most cases, toolboxes use traditional optimization methods which are based on the assumption of a convex, continuous and frictionless search space. Consequently, it is not guaranteed that the reported solution really is correct: unbeknownst to the user, border solutions, local optima or arbitrary solutions can be reported.

It therefore seems desirable to have methods that can deal reliably with the particularities of econometric and statistical estimation and calibration problems such as heuristic methods. These can deal with problems such as local optima and discontinuities. Moreover, for many of these methods, convergence proofs exist and, when calibrated adequately, they will find the correct solution. This presentation outlines the working of some popular and powerful heuristics and demonstrates how they can be used in time series analysis. Examples include GARCH-, STAR- and VEC-models, where only continuous parameters have to be estimated, but might also come with the highly demanding combina-

torial problem of lag selection, which is even more challenging if one allows for "holes" in the lag structure. Furthermore, it is shown how these methods can facilitate concepts from robust statistics that can only be approach numerically.

Keywords: Heuristics, optimization, maximum likelihood, lag selection, model selection.

Nonparametric Zhang tests for comparing distributions

Marco Marozzi
University of Calabria
marco.marozzi@unical.it

We consider nonparametric tests because parametric tests require many assumptions to be applied but in practice assumptions like normality are seldom satisfied or sample sizes are not sufficiently large to rely on the central limit theorem. In particular we consider permutation tests because they are particularly suitable for combined testing. Moreover, they do not even require random sampling, only exchangeability of observations between samples under the null hypothesis that the parent distributions are the same. Permutation testing is valid even when a non random sample of units is randomized into two groups to be compared. This circumstance is very common in biomedical studies.

Many nonparametric tests have been developed for comparing the distribution functions of two populations. They may be classified as tests for detecting (i) location differences, (ii) scale differences, (iii) joint location and scale differences; (iv) any differences between the distributions. Nonparametric combined tests have been very useful to address problems (i), (ii) and (iii). Combined testing is an effective strategy because generally non combined (single) tests show good performance only for particular distributions (Marozzi, 2011). Since in many actual situations there is no clear knowledge about the parent distribution, the problem of which test should be chosen in practice arises. The aim of the paper is to see whether nonparametric combined tests are useful also to address the general two sample problem (iv). We aiming at proposing a test that even though was not the most powerful one for every distribution, it has good overall performance under every type of distribution, a combined test that inherits the good behavior shown by a certain number of single tests in particular situations.

We consider two combining functions: the direct one and the one based on Mahalanobis distance. A generic combined test statistic for the general two sample problem is defined as $T_\psi = \psi(\mathbf{T})$ where ψ is a proper combining function, $\mathbf{T} = (T_1, ..., T_K)'$, $T_k = \frac{|S_k - E(S_k)|}{\sqrt{VAR(S_k)}}$, S_k is a two sided test statistic for the general two sample problem whose large values speak against H_0, $E(S_k)$ and $VAR(S_k)$ are respectively the mean and the variance of S_k, $k = 1, ..., K$, K is a natural number with $2 \leq K < \infty$. Traditional tests for the general two sample problem are the Kolmogorov Smirnov, Cramer Von Mises and Anderson Darling tests. Zhang (2006) proposed an unified approach that not only generates the traditional tests but also new nonparametric tests based on the likelihood ratio. We consider the Zhang tests that are analog to the traditional tests. We would like to study the type-one error rate of Zhang tests, that has not be studied before, and to see whether the combined test framework is useful when is applied to these tests. It is very difficult to derive theoretically optimality properties for nonparametric tests with completely unknown distributions of the populations behind the samples. Therefore to study and compare type-one error rate and power of the tests we rely upon Monte Carlo simulation. If the simulation size is sufficiently large then simulated type-one error rate and power of the tests will be reasonably close to the true values.

Keywords: nonparametric testing, general two sample problem, permutation testing, combined testing.

References

Marozzi, M. (2011): Levene type tests for the ratio of two scales. *Journal of Stat. Comput. and Simul.*, Vol. 81, pp. 815-826.

Zhang, J. (2006): Powerful two-sample tests based on the likelihood ratio. *Technometrics*, Vol. 48, pp. 568-587.

Turning simulation into estimation

Maarten Marsman
Cito, Institute for Educational Measurement, the Netherlands
Maarten.Marsman@cito.nl

Gunter Maris, Timo Bechger
Cito, Institute for Educational Measurement, the Netherlands
gunter.maris@cito.nl, timo.bechger@cito.nl

Cees Glas
Department of Research Methodology, Measurement, and Data Analysis
University of Twente, the Netherlands
c.a.w.glas@utwente.nl

Many models exist for which it is easy to generate data, but it is difficult or impossible to generate from its parameter(s) posterior distribution(s). We introduce a new class of composition algorithms to sample from these posterior distributions, based on the idea that once we can generate data from the model, we can simulate from its parameters posterior.

Posterior distributions are proportional to a product of the distribution of the data $\mathbf{X}$ conditional on the model parameter(s) θ and the prior distribution for θ. To sample from the posterior we set up a Markov chain using a variant of the Metropolis algorithm as follows. We generate data using composition, i.e., we generate a candidate parameter value θ^* from the prior and with this candidate value generate a proposal data set $\mathbf{X}^*$. The candidate value θ^* is a draw from a posterior distribution of a parameter with realization $\mathbf{X}^*$, i.e., $\theta|\mathbf{X}^*$, and of the same form as the target posterior. We then feed θ^* and $\mathbf{X}^*$ to the Metropolis algorithm as a proposal for θ (with realization $\mathbf{X}$). This is known as the Single-Variable-Exchange algorithm (Murray, Ghahramani, & MacKay, 2006).

We consider its use in latent variable models, i.e., models written as a product of a conditional distribution $\mathbf{X}|\theta$, where θ is a latent variable, and a distribution of the θ's, which usually involves a large number of la-

tent variables θ. Generating values from the two distributions is usually straightforward, but it is difficult to simulate from the posterior $\theta|\mathbf{X}$. For these models we consider modifications of the basic algorithm to obtain high efficiency, i.e., few rejected samples and large steps in the parameter space.

We will provide an illustration of composition in an application of the Signed Residual Time model (Maris & van der Maas, 2012) for accuracy and response time in a massive computer adaptive test (CAT) called the Maths Garden (Mathsgarden.com). The Maths Garden is a web-based CAT used by a large number of pupils which respond on a frequent basis to items from a large item bank.

Keywords: Bayesian Statistical Methods, Item Response theory, MCMC, Metropolis algorithm, Single-Variable-Exchange algorithm.

References

Maris, G., van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika, 77* , 615-633.

Murray, I., Ghahramani, Z., MacKay, D. (2006). Mcmc for doubly-intractable distributions. In R. Dechter & T. Richardson (Eds.), *Uncertainty in artificial intelligence* (p. 359-366). AUAI Press.

Bayesian Estimation of Multidimensional IRT Models for Polytomous Data

Irene Martelli, Mariagiulia Matteucci, and Stefania Mignani
Department of Statistical Sciences, University of Bologna, Italy
irene.martelli2@unibo.it,m.matteucci@unibo.it,
stefania.mignani@unibo.it

The aim of the work is to conduct a simulation study to verify item parameter recovery of confirmatory item response theory (IRT) models for ordinal data, assuming correlated abilities, by using the Gibbs sampler for model estimation.

Recently, estimation of IRT via simulation, *i.e.* using Markov chain Monte Carlo (MCMC) techniques, has become very popular due to the high capability of fitting to different models and to the increasing availability of cheap computing power that limited its use in the past. In this context, a fully Bayesian approach is adopted with the advantages of estimating the item parameters and individual abilities jointly, including uncertainties about item parameters and abilities in the prior distributions, and using Bayesian model comparison techniques. MCMC estimation of IRT models can be viewed as an alternative to marginal maximum likelihood (MML) estimation, where the approximation of multiple integrals involved in the likelihood function, especially for increasingly complex models, may represent a serious problem.

In the set of complex IRT models, we can include multidimensional models (Béguin and Glas, 2001; Edwards, 2010; Sheng and Wikle, 2009) and multilevel models (Fox, 2005; 2010). In particular, we focus on the first class of models which are used when the assumption of unidimensionality does not hold. From an empirical point of view, this assumption is explicitely violated when a test consisting of different subtests is submitted to a sample of candidates and different abilities are involved in the response process. Among different multidimensional approaches, we refer to confirmatory models where the relationship be-

tween the latent variables and the response variables is specified in advance. Following the approach of Sheng and Wikle (2009), we consider the multi-unidimesional model, where each latent trait is related to a single set of items, and the additive model, involving the existence of an overall ability. Both models may assume correlated abilities and, for this reason, the additive model can be distinghuished in the literature from the well-known bi-factor model.

In this work, we propose the Gibbs sampler algorithm for the estimation of the multi-unidimensional and the additive model in their extension from binary to ordinal data. The assessment of item parameter recovery is conducted through a simulation study on a bidimensional case by varying the number of response categories, the sample size, the test and subtest length and the ability correlations.

Keywords: Multidimensional IRT models, MCMC estimation, ordinal data.

References

Béguin, A., Glas, C.A.W. (2001): MCMC Estimation and Some Model-Fit Analysis of Multidimensional IRT Models, *Psychometrika*, Vol. 66, pp. 541-562.

Edwards M.C. (2010): A Markov Chain Monte Carlo Approach to Confirmatory Item Factor Analysis, *Psychometrika*, Vol. 75, N. 3, pp. 474-497.

Fox J.P. (2005): Multilevel IRT Using Dichotomous and Polytomous Response Data, *British Journal of Mathematical and Statistical Psychology*, Vol. 58, pp. 145-172.

Fox J.P. (2010): *Bayesian Item Response Modeling: Theory and Application*, Springer, New York.

Sheng, Y., Wikle, C. (2009): Bayesian IRT Models Incorporating General and Specific Abilities, *Behaviormetrika*, Vol. 36, pp. 27-48.

The use of the scalar Monte Carlo estimators for the optimization of the corresponding vector weight algorithms

Ilya N. Medvedev
Novosibirsk State University
Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
min@osmf.sscc.ru

In this talk the problem of constructing the weight Monte Carlo estimators with finite variance for estimating the solution of the system of integral equations of the second kind

$$\phi_i(x) = \sum_{j=1}^{m} \int\limits_X k_{ij}(x,y)\phi_j(y)dy + h_i(x)$$

or in the vector form $\Phi = \mathbf{K}\Phi + H$, where $H^{\mathrm{T}} = (h_1, \ldots, h_m)$,

$$\mathbf{K} \in [L_\infty \to L_\infty], \;\; ||H||_{L_\infty} = \mathrm{vrai} \sup_{i,x} |h_i(x)|.$$

is studied. The weight skalar estimator is defined as in I.N.Medvedev, G.A.Mikhailov (2011):

$$\xi_{(i,x)} = h(i,x) + \delta_{(i,x)} q((i,x),(j,y))\xi_{(j,y)}.$$

We present some new modifications of our criterion (G.A. Mikhailov, I.N. Medvedev 2006) for finiteness of weight skalar estimator variance that is based on the use of adjoint system of integral equations with majorant kernels. It is shown that under some given conditions the weight skalar estimator variance is always greater then the corresponding weight vector estimator variance.

Also we present the scalar weight collision estimator with the use of branching the trajectory into ν random independent branches

$$\zeta_{(i,x)} = h(i,x) + \delta_{(i,x)} \sum_{n=1}^{\nu} \zeta_{(j,y)}^{(n)}, \quad E\nu \equiv \frac{k_{ij}(x,y)}{p_{ij}(x,y)}$$

where $p_{ij}(x,y)$ is transition distribution density from x to y. It is proved that the $E\zeta_{(i,x)}^2$ and average simulation time for one trajectory are always bounded if the initial functional $\phi_i(x)$ is bounded. Finally we present some remarks about vector weight collision estimator with the use of branching the trajectory into ν random independent branches

$$\boldsymbol{\zeta}_x = H(x) + \delta_x \frac{Q(x,y)}{E\nu(x,y)} \sum_{n=1}^{\nu} \boldsymbol{\zeta}_y{}^{(n)}.$$

Keywords: solution of the system of integral equations, scalar (vector) weight Monte Carlo estimator, finite variance, branching ..., keyword5.

Acknowledgements: This work was supported by Russian Foundation of Basic Research (grants 12-01-00034 and 12-01-31328)

References

G.A. Mikhailov, I.N. Medvedev (2006) A new criterion of weight estimate variance finiteness in statistical modeling // *Proceedings 7-th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* pp.561-576, Ulm, Germany, 2006

I.N.Medvedev, G.A.Mikhailov (2011): Probabilistic-algebraic algorithms of Monte Carlo methods // *Russian Journal of Numerical Analysis and Mathematical Modelling*, Vol. 26, 3. P. 323-336

On comparison of regression curves based on empirical Fourier coefficients

Viatcheslav Melas, Andrey Pepelyshev
St.Petersburg State University
vbmelas@post.ru, andrey@ap7236.spb.edu

Luigi Salmaso, Livio Corain
University of Padova
luigi.salmaso@unipd.it,livio.corain@unipd.it

Comparison of regression curves is one of the important problems in applied regression analysis and in many fields of applications only two regression functions are observed. Let observation results be described by the relation

$$Y_{l,j} = f_l(t_{l,j}) + \varepsilon_{l,j}, \quad j = 1, \ldots, n_l,\ l = 1, 2, \tag{1}$$

where design points are equidistant and after an appropriate re-scaling belong to the unit interval, $t_{l,j} = (j-1)/(n_l - 1)$, $j = 1, 2, \ldots, n_l$; $f_l : [0, 1] \to R$ are unknown continuous real valued functions; $\varepsilon_{l,j}$, $j = 1, \ldots, n_l$, $l = 1, 2$, are i.i.d. random variables with zero mean and finite variance σ_l^2. Suppose that the random variables $\varepsilon_{l,j}$, $j = 1, \ldots, n_l$, $l = 1, 2$, are mutually independent.

We consider the problem of testing the null hypothesis

$$H_0 : f_1 = f_2 \tag{2}$$

under the alternative hypothesis

$$H_1 : f_1 \neq f_2.$$

To deal with this problem, we introduce a new method for testing the equality of regression curves of unknown functional form. In general, there are many methods for solving this problem but, roughly speaking,

they can be divided in groups of two types: fully prior methods and methods with auxiliary parameters that should be chosen using some prior information about regression functions. Two examples of methods of the first type are given in Delgado (1993) and Munk and Dette (1998). Examples of methods of the second type are mainly based on smoothing regression curves and need to fix a smoothing parameter [see King et al. (1991) among many others]. The approach suggested in Mohdeb et al. (2010) can be referred to the second type since it assumes a fixed number of Fourier coefficients to be indicated prior to calculations. Typically, the second type tests can be more powerful than fully prior ones but this holds under the condition that a good prior information is available.

In contrast, our approach does not belong these two types since the proposed method uses the adaptive choice of the number of Fourier coefficients. We prove theoretically and numerically that in typical cases the power of our adaptive approach is very close to the power of the approach based on the best prior choice of the number of the coefficients with account of explicit form of the regression functions. The work by V. Melas was partly supported by RFBR (project 12-01-00747a).

References

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. Statist. Probab. Lett. 17: 199–204.

King, E.C., J.D. Hart and T.E. Wehrly (1991), Testing the equality of two regression curves using linear smoothers, Statist. Probab. Lett. 12, 239-247.

Mohdeb, Z., Mezhoud, K.A., and Boudaa, D. (2010). Testing the equality of nonparametric regression curves based on Fourier coefficients. Journal Afrika Statistika, 5(4): 219–227.

Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. Ann. Statist., 26: 2339–2368.

Bayesian estimation with INLA for logistic multilevel models

Silvia Metelli
Leonardo Grilli, Carla Rampichini
Department of Statistics, Computer Science, Applications - University of Florence, Italy
silvia.metelli1@gmail.com ,
grilli@disia.unifi.it, rampichini@disia.unifi.it

In multilevel models for binary responses, estimation is computationally challenging due to the need to evaluate intractable integrals. In this paper, we investigate the performance of a recently proposed Bayesian method for deterministic fast approximate inference, Integrated Nested Laplace Approximation INLA (Rue et al., 2009) through an extensive simulation study, making comparisons with Bayesian MCMC Gibbs sampling and maximum likelihood estimation with Adaptive Gaussian Quadrature (AGQ). Particular attention is devoted to the case of small sample size and to the specification of the prior distribution for the variance component. We use three different specifications for the precision (inverse of level 2 variance): (*i*) $\Gamma(1, 0.0005)$, default choice of the `inla` function; (*ii*) $\Gamma(0.001, 0.001)$, default choice of the widespread BUGS software (Lunn, Thomas et al., 2000); (*iii*) $\Gamma(0.5, 0.0164)$, a prior specification recently suggested by Fong et al. (2010). Our findings show accurate estimates for the fixed effects, whereas the estimates of the variance component are quite sensitive to the number of clusters and highly dependent on the choice of the prior distribution. When the number of clusters is large, all the considered priors yield satisfactory results. However, in the case of few clusters, only the prior $\Gamma(0.5, 0.0164)$ gives an acceptable bias on the variance component. From the comparison of the three methods we can gather that the patterns are similar. Overall, INLA is more accurate than MCMC, which is in turn more accurate than AGQ, even if the differences vanish as the number of clusters in-

creases. Both MCMC and INLA results strongly depend on the choice of the prior distribution and the bias for the level 2 variance can be either positive or negative. On the contrary, with AGQ we know the direction of the bias, which is always negative.
Finally, INLA is considerably faster than MCMC and it has computational times similar to AGQ.

Keywords: Integrated Nested Laplace Approximations, Logistic multilevel models, MCMC estimation, Prior specification.

Acknowledgements: This work was supported by the Italian government FIRB 2012 project n. RBFR12SHVV_003: *Mixture and latent variable models for causal inference and analysis of socio-economic data*.

References

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71, pp. 319-392.

Fong Y., Rue H. and Wakefield, J. (2010). Bayesian inference for Generalized Linear Mixed Models. *Biostatistics*, 11, pp. 397-412.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 7:34.

Probability model of the interacting particles ensemble evolution and the parametric estimate of the nonlinear kinetic equation solution

Mikhailov G.A.
Novosibirsk State University
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch RAS
gam@osmf.sscc.ru

Rogasinsky S.V.
Novosibirsk State University
Institute of Computational Mathematics and Mathematical Geophysics,
Siberian Branch RAS
svr@osmf.sscc.ru

Model process of stochastic kinetics of N particles system is a uniform Markov chain in which transitions are carried out as a result of elementary pair interactions. Time distribution between their is defined by the system status and is the generalized exponential with a variable parameter. We consider this Markov chain described above in an extended phase space introducing the number of the pair realizing a collision in the system into the set of phase variables. This important initial point leads to the "fibering" of the collision distribution according to the number of the pair of particles in the system. Such transformation of the phase space is necessary for derivation of a special integral equation used as the base for the construction of weight modifications of statistical modelling of the many-particle system considered here. It is known that under the assumption of "molecular chaos" the appropriate one-particle density asymptotically satisfies the Boltzmann equation for $N \to \infty$ and, as a rule, the corresponding error has the order of $O(1/N)$. The "molecular chaos" means factorization of the "two-particle" density, which can be assumed as the independence of random frequencies for nonintersectings subdomains of the phase space, i.e., the Poisson property of the ensemble of the particles. It is noted that under this assumption the Boltzmann equation describes the balance of the

mean number of particles in phase space. Thus, the Boltzmann equation is, in particular, a balance equation for a Poisson ensemble of pairwise interacting particles.

For the purpose of optimization of the direct statistical simulation method (DSM) in this work "the modified method of the majorant frequency" as a statistical modeling algorithm of the generalized exponential distribution with parameter $\sigma(\bullet) = \sum\limits_{i=1}^{m} \sigma_i(\bullet)$ in the assumption that $\sigma_i(\bullet) \leq \sigma_i^*(\bullet)$, moreover the values $\sigma_i^*(\bullet)$ and, hence, $\sigma(\bullet) = \sum\limits_{i=1}^{*} m\sigma_i^*(\bullet)$, are being calculated sufficiently simply is built and proofed. Note that in [1] case by $\sigma_i^*(\bullet) = \sigma^*(\bullet)/m$ was actually considered.
Under of the assumption that the particles ensemble is the Poisson particles ensemble in this work mean squared optimization of a global estimate of the one-particle density by the histogram is realized. For the given error δ gives suitable values of a step h of a averaging grid and number n of independent implementations of a basic Markov chain are obtained.
When using a method of the majorant frequency, elaboration of the functional estimates of the DSM method grows linearly on N.
It is shown that weight modifications of DSM allow to build unbiased estimates of parametric derivative of functionals and on this basis to solve appropriate inverse problems.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (projects 12-01-00034-a, 13-01-00746-a, 13-01-00441-a, 12-01-00727-a), , the Integration grant 2012 -No.47, 126 of the Siberian Branch RAS.

References

Ivanov M.S., Rogasinsky S.V.(1990): Statistical simulation of rarefied gas flows by the majorant frequency principal, *Doklady USSR Academy of Sciences*, Vol. 312, N. 2, pp. 315-320.

Mikhailov G.A., Rogazinskii S.V. (2012): Probabilistic model of many-particle evolution and estimation of solutions to a nonlinear kinetic equation, *Rus. J. Numer. Analys. and Math. Modelling*, Vol.27, N.3, pp. 229-242.

Mathematical problems of statistical simulation of the polarized radiation transfer

Mikhailov G.A., Korda A.S., Ukhinov S.A.
Institute of Computational Mathematics and Mathematical Geophysics
SB RAS, Novosibirsk State University, Novosibirsk, 630090, Russia
gam@sscc.ru, asc@osmf.sscc.ru, sau@sscc.ru

Various aspects of usage and substantiation of standard vectorial algorithm of statistical modeling of polarized radiation transfer are considered. Additional researches of the variant of the matrix-weight algorithm based on direct simulation of "scalar" transfer process are carried out. Due to the fact that the appropriate statistical estimates can have the infinite variance, the method of "ℓ-fold polarization", in which recalculation of a Stokes vector on a "scalar" trajectory is carried out no more, than ℓ times, is offered deprived of this deficiency. Thus polarization is not exactly taken into account, but errors of required estimates can be quite small. First of all, it is important for studying the changes in estimates of radiation intensity (first element of a Stocks vector) when you select vectorial or scalar model in inverse problems of atmospheric optics.

We also have considered the double local estimate, used for calculation of vectorial intensity in the given point of a phase space. The practically effective evaluation of the bias which here arises when "cutting" an auxiliary small sphere for limitation of a mean squared error is given. It is specified, how to apply a method of "ℓ-fold polarization" to implementation of local estimates.

The problem on a variance finiteness of appropriate standard vectorial estimates of a Monte-Carlo method is studied also. For this purpose the system of integral equations defining a matrix of the second moments of a weight vectorial estimate has been considered. By means of numerical evaluations on the basis of resolvent iteration procedure it is shown that the spectral radius of an appropriate matrix-integral operator

is rather close to product of similar spectral radius for an infinite medium which is calculated analytically, on simply estimated spectral radius of the scalar integral operator. For the purpose of the extension of possibilities of analytical studies of this practically important factorization it is presented dual (in relation to considered earlier) representation of mean squares of the Monte Carlo estimates of studied functionals.

Keywords: Monte Carlo method, vector integral equations, radiation transfer.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (13-01-00441, 12-01-00034, 12-01-00727, 12-01-31328), and by MIP SB RAS (A-47, A-52)

References

Mikhailov G.A., Ukhinov S.A., Chimaeva A.S. (2006): Variance of a Standard Vector Monte Carlo Estimate in the Theory of Polarized Radiative Transfer, *Computational Mathematics and Mathematical Physics*, Vol. 46, N. 11, pp. 2006-2019.

Mikhailov G.A., Ukhinov S.A. (2011): Dual Representation of the Mean Square of the Monte Carlo Vector Estimator, *Doklady Mathematics*, Vol. 83, N. 3, pp. 386-388.

Mikhailov G.A., Lotova G.Z. (2012): A Numerical-Statistical Estimate for a Particle Flux with Finite Variance, *Doklady Mathematics*, Vol. 86, N. 3, pp. 743-746.

Cluster Weighted Modeling with B-splines for longitudinal data

Simona C. Minotti
Dipartimento di Statistica e Metodi Quantitativi, Università di Milano-Bicocca (Italy)
simona.minotti@unimib.it

Giorgio A. Spedicato
Dipartimento di Scienze Bancarie, Università Cattolica di Milano (Italy)
spedicato_giorgio@yahoo.it

Cluster-Weighted Modeling (CWM) is a flexible family of mixture models for fitting the joint density of a response variable and a set of explanatory variables. Let $(\mathbf{X}, Y)$ be the pair of random vector $\mathbf{X}$ and random variable Y defined on Ω with joint probability distribution $p(\mathbf{x}, y)$, where $\mathbf{X}$ is the d-dimensional input vector with values in some space $\mathcal{X} \subseteq \mathbb{R}^d$ and Y is a response variable having values in $\mathcal{Y} \subseteq \mathbb{R}$. Thus, $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d+1}$. Suppose that Ω can be partitioned into G disjoint groups, say $\Omega_1, \ldots, \Omega_G$, that is $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$. CWM decomposes the joint probability $p(\mathbf{x}, y)$ as follows:

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(y|\mathbf{x}, \Omega_g)\, p(\mathbf{x}|\Omega_g)\, \pi_g, \tag{1}$$

where $p(y|\mathbf{x}, \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\mathbf{x}$ and Ω_g, $p(\mathbf{x}|\Omega_g)$ is the probability density of $\mathbf{x}$ given Ω_g, $\pi_g = p(\Omega_g)$ is the mixing weight of Ω_g, ($\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$), $g = 1, \ldots, G$, and $\boldsymbol{\theta}$ denotes the set of all parameters of the model.

Hence, the joint density of $(\mathbf{X}, Y)$ can be viewed as a mixture of local models $p(y|\mathbf{x}, \Omega_g)$ weighted (in a broader sense) on both local densities $p(\mathbf{x}|\Omega_g)$ and mixing weights π_g.

The original formulation, proposed by Gershenfeld, Schoener and Metois (1999) under Gaussian and linear assumptions, was developed in the context of media technology to build a digital violin with traditional inputs and realistic sound. Quite recently, Ingrassia, Minotti and Vittadini (2012) reformulated CWM in a statistical setting and proposed an extension to Student-t distributions.

The idea from Gershenfeld, Schoener and Metois (1999) is that signals that are nonlinear, non-stationary, non-gaussian, and discontinous can be described by expanding the probabilistic dependence of the future on the past around local models of their relationships. Thus, predictor vector $\mathbf{x}$ might be a time-lag vector, while y could be the future value of a series. Since B-splines enable a flexible modeling of the functional development over time (see e.g. Luan and Lin, 2004), in this paper we propose to model $p(y|\mathbf{x}, \Omega_g)$ in (1) by means of B-splines. The proposal will be supported by means of simulations in order to provide a comparison with linear Gaussian CWM and traditional mixture models for longitudinal data.

Keywords: Cluster Weighted Modeling, B-spline, longitudinal data.

References

Gershenfeld N., Schoener B., Metois E. (1999): Cluster-weighted modelling for time-series analysis, *Nature*, Vol. 397, pp. 329-332.

Ingrassia S., Minotti S.C., Vittadini G. (2012): Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, Vol. 29, N.3, pp.363-401.

Luan Y., Lin H. (2004): Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioninformatics*, Vol. 19, N.4, pp.474-482.

Statistical Challenges in Estimating Frailty Models on Mortality Surfaces

Trifon I. Missov
Max Planck Institute for Demographic Research
missov@demogr.mpg.de

Lifetable adult mortality data (death counts and exposures in the absence of explanatory variables) are usually fit parametrically by a gamma-frailty model with a Gompertz-Makeham baseline (Vaupel et al. 1979). Its straight application to cohort mortality data produces, though, dubious parameter estimates as it does not incorporate improvements in age-specific mortality rates that occur yearly. One possibility of dealing with this is to design an estimation procedure on mortality surfaces for a fixed age range over a fixed period of time:

$$\bar{\mu}(x,y) = \bar{z}(x_0, y-x) \cdot \bar{\boldsymbol{S}}^{\gamma} \cdot a(x_0,y) \cdot e^{b(x-x_0)} \,,$$

where $\bar{\mu}(x,y)$ is the marginal hazard at age x and year y; $\bar{z}(x_0, y-x)$ is the average frailty among survivors to age x_0 from the cohort born in year $y-x$; $\boldsymbol{S}$ is the $(y-x)$-cohort survivorship between ages x_0 and x; γ is the squared coefficient of variation of the frailty distribution; $a(x_0,y)$ and b are the Gompertz parameters. This model can be estimated in a number of special cases.

Formally, let $D(x,y)$ and $E(x,y)$ denote death counts and exposure at age x in year y. We can assume that death counts are Poisson-distributed $D(x,y) \sim \mathcal{P}\left[E(x,y) \cdot \bar{\mu}(x,y)\right]$. The log-hazard can be represented as a linear combination of a model matrix and a set of parameters: $\ln\left[\mathtt{vec}(\bar{\mu}(x,y))\right] = \boldsymbol{X}\boldsymbol{\beta}$ and estimation is performed by using (penalized) iteratively re-weighted least-squares:

$$(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{1}$$

where $\boldsymbol{W}$ and $\tilde{\boldsymbol{z}}$ are derived from the Poisson assumption (McCullagh

and Nelder 1989). Instead of enforcing parametric structure, we assume that both average frailty among survivors and mortality progress change smoothly over cohorts and years, respectively. This is captured in strategies 2. and 3. by the penalty term $\boldsymbol{P}$, which measures the roughness of $\bar{z}(x_0, y - x)$ and $a(x_0, y)$ with differences of order d, weighted by a positive regularization parameter (Camarda 2012). The performance of different modelling strategies is illustrated on Swedish female mortality data from HMD for years 1955-2000 and ages 80-104.

Keywords: frailty models, mortality surfaces, Poisson regression, iteratively re-weighted least squares

Acknowledgements: This work was supported by the *Research Project on the Rate of Aging* funded by the Max-Planck-Gesellschaft.

References

McCullagh, P., Nelder J.A. (1989): *Generalized Linear Model*, Monographs on Statistics and Applied Probability 37, Chapman and Hall, London.

Camarda C.G. (2012): MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines, *Journal of Statistical Software*, Vol. 50, pp. 1-24.

Vaupel J.W., Manton K.G., Stallard E. (1979): The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality, *Demography*, Vol. 16, pp. 439-454.

On the Generalized $\triangle^2$-Distribution for Constructing Exact D-Optimal Designs

Trifon I. Missov
Max Planck Institute for Demographic Research & Univeristy of Rostock
missov@demogr.mpg.de

Sergey M. Ermakov
Department of Statistical Modelling, Faculty of Mathematics and Mechanics,
State Petersburg State University
sergej.ermakov@gmail.com

We suggest a procedure for constructing exact D-optimal designs with a predefined number of points. The method is implemented in three steps: first, sampling from the generalized Δ^2-distribution (Missov and Ermakov, 2009), second, estimating the resulting sample modes, and, third, running a differential evolution algorithm for precisely allocating the global maximum of the information matrix (). The method depends neither on the choice of design region, nor on the selection of linearly independent functions in it.

We present results for polynomial regression in uni-, two-, and three-dimensional regions. In $[0,1]^2$ the exact D-optimal design is located 1) on the vertices of $[0,1]^2$ with different weights (linear regression), 2) on the vertices and sides of $[0,1]^2$ with not more than one internal point (quadratic regression), 3) on the vertices and sides of $[0,1]^2$ with not more than three internal points (cubic regression). In $[0,1]^3$ is located on the vertices of the design region with different weights. In all cases the exact D-optimal design is not unique.

Keywords: Exact D-optimal designs, generalized Δ^2-distribution, differential evolution.

Acknowledgements: This work was supported by RFBR grant 11-01-00769-a.

References

Ermakov S.M., Zolotukhin V.G. (1960): Polynomial Approximations and the Monte Carlo Method, *Theor. Probability Appl.*, Vol. 5, pp. 428-431.

Podkorytov A.N. (1975): On the Properties of D-Optimal Designs for Quadratic Regression, *Vestnik LGU*, Vol. 2, N. 7, pp. 163-166.

Price K.V., Storn R., Lampinen J.. (2005): *Differential Evolution: A Practical Approach to Global Optimization*, Springer, Berlin-Heidelberg-New York.

Missov T.I., Ermakov S.M. (2009): On Importance Sampling in the Problem of Global Optimization, *Monte Carlo Methods and Applications*, Vol. 15, N. 2, pp. 135-144.

A new and easy to use method to test for interaction in block designs

Karl Moder
Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Vienna, Austria

Block designs are often used designs to evaluate influences of a factor in the presence of some disturbance variables. Although this kind of design is widely used, it suffers from one drawback. As there is only one observation for each combination of a block and factor level it is not possible to test interaction effects because the mean square value for interaction has to serve for the error term. Although there are some attempts to overcome this problem these methods however, have not been adopted in practice and have not been broadly disseminated. Many of these tests are based on nonlinear interaction effects (e.g. Tukey 1949, Mandel 1961, Johnson and Graybill 1972). Others are based on the sample variance for each row in the block design (Milken an Ramuson 1977) with some modification by Piepho (1994). A somehow similar method was proposed by Kharrati-Kopaei and Saddooghi-Alvandi (2007). A review on such tests is given by Karabatos (2005) and Alin and Kurt (2006). Rasch et al. (2009) proposed the use of nonlinear regression which is fitted to Tukey's model and is tested by the likelihood ratio test. Here a new model is introduced to test interaction effects in block designs. It is based on an additional assumption regarding the columns of the block design which is intuitive and common in Latin Squares. The application of this model is very simple and a test on interaction effect is very easy to calculate based on the results of an appropriate analysis of variance. The method as such is applicable for fixed effect models as well as for mixed and random effect models.

Keywords: block designs, test on interaction, power .

References

Alin, A. and S. Kurt (2006). Testing non-additivity (interaction) in two-way anova tables with no replication. *Stat. in Medicine* 15, 63-85.

Johnson, D. E. and F. A. Graybill (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association* 67, 862-869.

Karabatos, G. (2005). Additivity Test. In B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*, pp. 25-29. Wiley.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association* 56, 878-888.

Millken, G. A. and D. Rasmuson (1977). A heuristic technique for testing for the presence of interaction in nonreplicated factorial experiments. *Australian Journal of Statistics* 19 (1), 3238.

Piepho, H.-P. (1994). On tests for interaction in a nonreplicated two-way layout. *Australian Journal of Statistics* 36 (3), 363369.

Rasch, D., T. Rusch, M. Simeckova, K. Kubinger, K. Moder, and P. Simecek (2009). Tests of additivity in mixed and fixed effect two-way anova models with single sub-class numbers. *Statistical Papers* 50 (4), 905-916.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics* 5, 232-242.

Sample size in approximate sequential designs under several violations of prerequisites

Karl Moder
University of Life Sciences, Vienna, Austria
karl.moder@boku.ac.at

In sequential designs for means (one-and two-sample problems) the test statistic is compared with the α-percentile of the standard normal distribution, Therefore the test is approximate and the formulae for the expected sample size too. By simulation for several situations we calculated empirical sample sizes for the OBrien Fleming test and for triangular sequential designs as described in Rasch et al. (2011). The tests and situations considered are: The group sequential test OBrien-Fleming Method (1979), the triangular method of Kittelson and Emerson (1999) and the method of Whiteheads and Stratton for continuous monitoring. Simulations were carried out for several situations of assumed population variances and deviations from these assumptions for the sample distributions. Various situations of non normal distributions in regard to skewness and kurtosis (based on the Fleishman (1978)) and their influence on alpha and power were examined. All simulations very carried out using SAS 9.2 (2008) by means of proc seqdesign and proc seqtest. In general we found that deviations from the normal distribution in shape are of less importance than those of heterogeneous variances.

Keywords: group sequential test, continuous monitoring, sample size.

References

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4):521-532.

Kittelson, J. M. and Emerson, S. S. (1999), A Unifying Family of Group Sequential Test Designs, *Biometrics*, 55, 874882.

OBrien, P. C. and Fleming, T. R. (1979), A Multiple Testing Procedure for Clinical Trials, *Biometrics*, 35, 549556.

Rasch, D.; Pilz, J., Gebhardt, A. and Verdooren, R.L. (2011), *Optimal Experimental Design with R*, Boca Raton, Chapman and Hall,

SAS Institute Inc. (2008). *SAS/STAT® 9.2 Users Guide*. Cary, NC: SAS Institute Inc.

Whitehead, J. and Stratton, I. (1983), Group Sequential Clinical Trials with Triangular Continuation Regions, *Biometrics*, 39, 227236.

A wild-bootstrap scheme for multilevel models

Lucia Modugno
Department of Statistical Sciences, University of Bologna
lucia.modugno@unibo.it

Simone Giannerini
Department of Statistical Sciences, University of Bologna
simone.giannerini@unibo.it

In this work we propose a modified version of the wild bootstrap procedure for multilevel data. Inference in multilevel models usually relies upon maximum likelihood methods (e.g. Skrondal and Rabe-Hesketh (2004)) that mostly use asymptotic approximations for the construction of test statistics and estimation of variances. If the sample size is not large enough, the asymptotic approximation is not reliable and can lead to incorrect inferences. By using bootstrap methods, under some regularity conditions, it is possible to obtain a more accurate approximation of the distribution of the statistics. Three general resampling approaches are well established in the case of hierarchical data (discussed for example in Van der Leeden et al. (2008); Goldstein (2010)): the parametric, the residual and the cases bootstrap.

The wild bootstrap, developed by Liu (1988), is a technique aimed to obtain consistent estimators for the covariance matrix of the coefficients of a regression model when the errors are heteroscedastic. Further evidences and refinements are provided in Flachaire (2004) and Davidson and Flachaire (2008). Here, we introduce a modified version of the wild bootstrap procedure which is particularly suitable to hierarchical data. We assess the finite size performances of the proposed bootstrap scheme and compare it with the three resampling schemes used for multilevel models by means of a Monte Carlo study where we vary sample size, error distribution and error variance. Both the cases bootstrap and

the wild bootstrap do not require homoscedasticity and do not make distributional assumptions on the error processes. However, the performance of the two schemes is very different in terms of coverage and length of confidence intervals. Also, for big sample sizes the wild bootstrap outperforms the three competitors in all the scenarios considered including the Gaussian homoscedastic case.

Keywords: Multilevel model; Wild bootstrap, Heteroscedasticity; Cases bootstrap.

References

Davidson R., Flachaire E. (2008): The wild bootstrap, tamed at last. *Journal of Econometrics*, Vol. 146, pp. 162–169.

Flachaire E. (2004): Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap, *Computational Statistics & Data Analysis*, Vol. 49, pp. 361–376.

Goldstein H. (2010): *Multilevel Statistical Models* (4th ed.), J. Wiley & Sons, Chichester.

Liu R. Y., (1988): Bootstrap procedures under some non-i.i.d. models, *Annals of Statistics*, Vol. 16, pp. 1696–1708.

Skrondal A., and Rabe-Hesketh S. (2004): *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*, Chapman & Hall, New York.

Van der Leeden R., Meijer E., Busing F. (2008): Resampling multilevel models. In de Leeuw J., Meijer E. (Eds.), *Handbook of Multilevel Analysis*, Springer, New York, Chapter 11, pp. 401–433.

The A-criterion: Interpretation and Implementation

John P. Morgan
Virginia Tech
jpmorgan@vt.edu

Jonathan W. Stallings
Virginia Tech
jstallin@vt.edu

In a suitable model framework, let C_d be the information matrix for estimation of parameters $\boldsymbol{\theta}$ when employing design d. The A-value for design d is the trace of the M-P inverse of C_d, trace(C_d^+), and a design is A-optimal if it minimizes the A-value over all competing designs. Necessary and sufficient conditions are established on $M = (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_p)'$ for the A-value to be proportional to

$$\text{trace}(MC_d^+M') \propto \textstyle\sum_{i=1}^p Var_d(\widehat{\boldsymbol{m}_i'\boldsymbol{\theta}}),$$

establishing the full range of interpretation for the A-optimality criterion. Examples are given for both full-rank and rank-deficient models. This reveals, for instance, a plethora of useful interpretations for the A-criterion when estimating contrasts in a simple one-way classification model.

The A-value under linear transformations of the targeted parameter vector is also examined, extending the above result. It is shown that A-value for a transformed parameter is equivalent to *weighted A-value* for the original parameter, and necessary and sufficient conditions are provided, parallel to those described above, for the weighted problem. It is shown how several seemingly disparate optimality problems are encompassed under the umbrella of weighting, and implications for algorithmic design are discussed. An interesting application is to fractional design for 2^m experiments where, instead of the traditional orthogonal

parametrization (OP), a baseline parametrization (BP) into main effects and interactions is employed. OP and BP models lead to markedly different A-optimal designs, though the *balanced array* concept is central to both.

Keywords: A-criterion, algorithmic design, balanced array, optimal design, weighted optimality.

A Bayesian Clinical Trial Design for Targeted Agents in Metastatic Cancer

Peter Müller
University of Texas, Austin, USA
pmueller@math.utexas.edu

We describe a Bayesian clinical trial design for cancer patients who are selected based on molecular aberrations of the tumor. The primary objective is to determine if patients who are treated with a targeted therapy that is selected based on mutational analysis of the tumor have longer progression-free survival than those treated with conventional chemotherapy.

The design includes a probability model for a random partition of patients into subgroups with similar molecular aberrations and baseline covariates and a cluster-specific sampling model for progression free survival. Inference includes an estimation of an overall treatment effect for matched targeted therapy that allows to address the primary objective of the trial.

Patients are randomized to targeted therapy or standard chemotherapy.

Keywords: Bayesian design, clinical trial.

Experimental Design for Engineering Dimensional Analysis

Christopher J. Nachtsheim
Carlson School of Management, University of Minnesota

Dimensional Analysis (DA) is a fundamental method in the engineering and physical sciences for analytically reducing the number of experimental variables affecting a given phenomenon prior to experimentation. Two powerful advantages associated with the method, relative to standard design of experiment (DOE) approaches are: (1) a priori dimension reduction, (2) scalability of results. The latter advantage permits the experimenter to effectively extrapolate results to similar experimental systems of differing scale. Unfortunately, DA experiments are underutilized because very few statisticians are familiar with them. In this paper, we first provide an overview of DA and give basic recommendations for designing DA experiments. Next we consider various risks associated with the DA approach, the foremost among them is the possibility that the analyst might omit a key explanatory variable, leading to an incorrect DA model. When this happens, the DA model will fail and experimentation will be largely wasted. To protect against this possibility, we develop a robust-DA design approach that combines the best of the standard empirical DOE approach with our suggested design strategy. Results are illustrated with some straightforward applications of DA. This effort represents joint work with Mark Albrecht, Thomas Albrecht, and Dennis Cook.

Keywords: dimension reduction, robust design, dimension analysis.

Goodness-of-fit tests for the power function distribution based on Puri–Rubin characterization

Ya. Yu. Nikitin
Saint-Petersburg State University, Russia
yanikit47@gmail.com

K. Yu. Volkova
Saint-Petersburg State University, Russia
efrksenia@gmail.com

Consider the family $\mathcal{P}_\lambda$ of power function distributions having the d.f. $G(x) = x^\lambda,\ x \in [0,1],\ \lambda > 0$. Power function distributions often appears in applications, e.g. in economics, in queueing and reliability theory.

We are interested in goodness-of-fit tests for this family independent of unknown nuisance parameter λ. The only attempt to build such tests has been traced by Martynov (2009) who outlined the traditional approach based on empirical processes with estimated parameters.

We develop completely different way by introducing two tests based on the characterization of the power function distribution by Puri and Rubin(1970): *Let X and Y be i.i.d. non-negative random variables with the continuous d.f. Then the equality in law of X and* $\min(\frac{X}{Y}, \frac{Y}{X})$ *takes place iff X has some d.f. from the family $\mathcal{P}_\lambda$.*

Let $X_1, X_2, \dots$ be i.i.d. observations with the continuous d.f. F and let F_n be the empirical d.f. based on the sample $X_1, \dots, X_n$. We are testing the hypothesis $H_0 : F \in \mathcal{P}_\lambda$ against the alternative $H_1 : F \notin \mathcal{P}_\lambda$. Let $H_n(t) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \mathbf{1}\{\min(\frac{X_i}{X_j}, \frac{X_j}{X_i}) < t\}$, $t \in (0,1)$ be the U-empirical d.f. related to the characterization.

Consider two statistics which can be used for testing H_0 against H_1 :

$$I_n = \int_0^1 (H_n(t) - F_n(t))\, dF_n(t),\ D_n = \sup_{t \in [0,1]} |H_n(t) - F_n(t)|.$$

We describe their large deviation asymptotics under H_0 : as $a \to 0$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(I_n > a) \sim -10.8a^2, \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(D_n > a) \sim -2.84a^2.$$

This allows us to calculate their local Bahadur efficiency under numerous local alternatives concentrated on $(0, 1)$. It is seen that the Kolmogorov statistic D_n is less efficient than the integral statistic I_n as usually happens in goodness-of-fit testing, see Nikitin (1995).

However, the local efficiencies for both sequences of statistics are reasonably high, and they can be recommended for testing. Finally we describe the conditions of local optimality of considered statistics.

Keywords: Power function distribution; U-statistics; Bahadur efficiency; goodness-of-fit test.

Acknowledgements: This work was partially supported by the grant of RFBR 13-01-00172a and within Meropriyatie 2 of Saint-Petersburg University.

References

Martynov G.V. (2009): Cramér-von Mises test for the Weibull and Pareto Distributions, *Proceedings of Dobrushin Internat. Conf.*, Moscow, pp. 117 - 122.

Nikitin Y. (1995): *Asymptotic efficiency of nonparametric tests*, Cambridge University Press, New York.

Puri P.S., Rubin H. (1970): A Characterization Based on the Absolute Difference of Two I. I. D. Random Variables. *Ann. Math. Stat.*, Vol. 41, pp. 2113 - 2122.

Flexible regression models in survival analysis

Mikhail Nikulin
IMB UMR 5251, Universite Victor Segalen, Bordeaux, France
mikhail.nikouline@u-bordeaux2.fr

Mariya Vedernikova
Novosibirsk State Technical University, Novosibirsk, Russia
vedernikova.m.a@gmail.com

We shall analyze survival data from clinical trials in oncology, when the crossing effects of survival functions could be observed. Accelerated life models, based on counting processes, are used more and more often in carcinogenesis studies for such kind of problems to relate lifetime distribution to the time-depending explanatory variables. Classical examples are the well-known data concerning effects of chemotherapy and chemotherapy plus radiotherapy on the survival times of gastric and lung cancers patients; Stablein and Koutrouvelis (1985), Piantadosi (1997), Wu, Hsieh and Chen (2002), and Bagdonavicius, Hafdi and Nikulin (2004). Following Bagdonavicius, Levuliene, Nikulin (2009), Bagdonavicius, Kruopis, Nikulin (2011), Nikulin and Wu (2006) we give examples to illustrate and compare possible applications of the Hsieh model (2001) and Bagdonavicius and Nikulin's (2002, 2005, 2006) simple cross effect (SCE) model, both of them are particularly useful for the analysis of survival data with one crossing point.

Keywords: goodness-of-fit tests, accelerated life models, cross-effect, carcinogenesis studies.

References

Bagdonavicius, V. and Nikulin, M. (2002). *Accelerated Life Models*. Boca Raton: Chapman and Hall/CRC.

Bagdonavicius, V. and Nikulin, M. (2005). Analysis of survival data with non-proportional hazards and crossings of survival functions. In: *Quantitative Methods in Cancer and Humain Risk Assesment, (eds. L.Edler, Ch.Kitsos)*, J.Wiley, N.Y., 193-209.

Bagdonavicus, V., Levuliene, R., Nikulin, M. (2009). Testing absence of hazard rates crossings, *Comptes Rendus de l'Academie des Sciences de Paris,Ser. I*, 346, 7-8,445-450.

Bagdonavicus, V., Kruopis, V., Nikulin, M. (2011). *Non-parametric tests for censored data*, ISTE-WILEY, 233p.

Hsieh, F. (2001). On heteroscedastic hazards regression models: theory and application. *Journal of the Royal Statistical Society*, B 63, 63-79.

Martinussen, T., Scheike, T. (2006). *Dynamic regression models for survival analysis*, Springer: New York.

Nikulin, M., Wu, H.-D. (2006). Flexible regression models for carcinogenesis data. *Probability and Statistics*, 10. Steklov Mathematical Institute in St.Petersburg, RAS , 78-101.

Piantadosi, S. (1997). Clinical Trials, J.Wiley: New York.

Stablein, D.M., Koutrouvelis, I.A. (1985). A two sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, 41, 643-652.

Wu, H-D.I. (2006). Statistical Inference for two-sample and regression models with heterogeneity effects: a collected-sample perspective. In: *Probability, Statistics and Modeling in Public Health*, (Eds. M. Nikulin, D. Commenges, C. Huber), Springer: New York, 452-465.

Wu, H-D.I. (2007). A Partial score test for difference among heterogeneous populations. *Journal of Statistical Planning and Inference*, 137, 527-537.

Conditions for minimax designs

Hans Nyquist
Department of Statistics, Stockholm University
Hans.Nyquist@stat.su.se

The construction of an optimum design of an experiment generally requires knowledge of unknown parameters. Two proposed approaches out of this dilemma include using optimum on-the-average designs, which uses a weighted average of criterion functions, and minimax designs, that finds the best guaranteed designs as the unknown parameters vary in a specified subset of the parameter space. The minimax design has shown to be mathematically intractable and numerically difficult to construct, which has restricted its practical use. The aim of this paper is to consider some relations between optimum on-the-average designs and minimax designs. With these relations clarified, the two approaches to optimum design will be better understood and the application of minimax designs will be easier to use in practice. In particular, an algorithm for construction of minimax designs is suggested.

Keywords: Least favorable distribution, Optimum design, Optimum on-the-average design.

Acknowledgements: This work was supported by the Swedish Research Council

Numerical stochastic models of meteorological processes and fields and some their applications

V.A. Ogorodnikov
Novosibirsk State University,
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
ova@osmf.sscc.ru

N.A. Kargapolova , O.V. Sereseva
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
nkargapolova@gmail.com, seresseva@mail.ru

Some approaches to the stochastic modeling of meteorological processes and fields are considered. One of them in based on modeling of an inhomogeneous vector Markov chain with a periodic in time matrix of transition probabilities (Kargapolova, 2012). Every vector component is an indicator function of the values of a meteorological process greater than a given level. the periodic properties of such a Markov chain allow one to take into account the daily variations of real processes. With this model some joint indicator series of various meteorological elements (for example, air temperature, components of wind velocity and daily precipitation) for given levels are constructed.

For modeling homogeneous with respect to the spatial variables and stationary in time fields of daily liquid precipitation on regular grid the following approach is used: Two types of interdependent fields are simulated. One field is an indicator field with a given correlation function and probabilities of precipitation. The other field represents the amount of precipitation with an appropriate correlation function and a one-dimensional distribution. The resulting field is the product of the above fields. The spatial - temporal correlation functions of both these fields are described by the product of the exponential spatial and temporary correlation functions with parameters describing the speed of

their decrease and spatial orientation. Selecting parameters determining the degree of dependence of the fields, it is possible to obtain a better approximation of the correlation structure of real field in model. On the basis of this model, an approximate numerical stochastic model of conditional non-Gaussian precipitation fields with given values at the weather stations is constructed. Long-term observations data on precipitation in the Novosibirsk region are used for estimating the model parameters. The above-considered models are used for stochastic interpolation of daily precipitation fields from weather stations to grid points, for estimation of the statistical properties dangerous rainfall modes, etc. Similar approaches were considered in e.g. (Evstafieva, 2005), (Kleiber, 2012).

Keywords: vector Markov chain, stochastic spatial-temporal fields, daily precipitation, extreme weather events, stochastic interpolation.

Acknowledgements: This work was supported by the Russian Foundation for Basis Research (grants 11-01-00641, 12-01-00727 and 12-05-00169).

References

N.A.Kargapolova, V.A. Ogorodnikov. Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorologocal processes // Russ. J. of Numer. Anal. and Math. Modelling, (2012), V. 27, No. 3. pp. 213-228

A.I. Evstafieva, E.I.Khlebnikova, V.A. Ogorodnikov. Numerical stochastic models for complexes of time series of weather elements. Russ. J. of Numer Anal. and Math. Modelling,(2005), V.20, No. 6, pp. 535-548.

Kleiber, W., R. W. Katz and B. Rajagopalan, 2012: Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. Water Resources Research, 48, doi:10.1029/2011WR011105.

A stochastic numerical model of daily precipitation fields based on an analysis of synchronous meteorological and hydrological data

V.A.Ogorodnikov
Novosibirsk State University,
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
ova@osmf.sscc.ru

V.A.Shlychkov
Institute for Water and Environmental Problems SB RAS
slavhome@ngs.ru

For stochastic modeling of atmospheric precipitation fields an approach based on a combined use of meteorological and hydrological observation data is proposed. The approach is developed for stochastic modeling of meteorological fields with scanty of meteorological information. The observation stations of the Siberian Arctic are few and far between on the huge territory. Therefore, it is difficult to construct and interpret appropriate mathematical models for the regions of Northern Siberia where serious consequences of global climate change are occurred.

A study has been carried out on the development of parametric numerical stochastic models of meteorological processes and fields with input information based on scanty available observation. For instance, the meteorological time series in stochastic modeling of homogeneous spatial - temporal fields of daily precipitate can be used in combination with hydrological observations. When both kinds of the data are synchronous in time and space, the approach allows one to receive information on river catchment produced by precipitation in each realization of the precipitation field. This additional information in combination

with the general climatic information, for example about wind direction, precipitation and other atmospheric parameters at the available meteorological stations can be used to correct the parameters of correlation function of fields of precipitation.

The model has been tested for the Berd river catchment in the vicinity of the of Novosibirsk with 2 meteorological and 2 hydrological stations. The calculations of the of precipitation where carried out with hydrological data and a combined deterministic model of precipitation and river runoff (Krysanova, 1989). The results of calculations of ensemble realizations of the precipitation field have shown that the statistical characteristics of modeling fields are close to the characteristics of real precipitation fields.

Keywords: stochastic spatial-temporal fields, daily precipitation, meteorological and hydrological observations, determined model, river catchment.

Acknowledgements: This work was supported by Russian Foundation for Basis Research (grants 11-01-00641, 12-01-00727 and 12-05-00169).

References

Simulation of system ”catchment basin - river - sea bay”. Edited by V. Krysanova and Kh. Luik. Tallinn. 1989. 428 P.

Structure of Life Distributions

Ingram Olkin
Stanford University
olkin@stat.stanford.edu

Nonnegative random variables arise in many contexts, as magnitudes of physical objects, as wind speeds, material strengths, life lengths, as well as in economic data. Whereas the normal distribution plays a central role for random variables that take on both positive and negative values, it is not generally a good model for nonnegative data. However, for nonnegative data, there is no distribution as basic as the normal distribution, with its foundation in the central limit theorem. Indeed, there are a number of competitors that serve as models for nonnegative data.

We classify methods as parametric, nonparametric, and semiparametric. Distribution-free methods do not depend on any assumptions about the underlying distribution. A more extensive category is that of qualitatively conditioned methods. For example, the density is unimodal or has a heavy right-hand tail.

Parametric families for survival analysis include the exponential, gamma, Weibull, lognormal, Gompertz, and others. Semiparametric families include distributions that have both a real parameter and a parameter that is itself a distribution. For lifetime data in survival or reliability families that contain, for example, a scale or power parameter, proportional hazards parameters are introduced. Interesting results on stochastic orderings are obtained for such families.

In this talk we provide an overview of these families with an emphasis on their characteristics.

Keywords: reliability, survival analysis, nonparametric families, semiparametric families.

TOPICS

1. Distributions
   (a) Lorenz curve
   (b) total time on test transform
2. Ordering distributions
3. Mixtures
4. Nonparametric families in reliability theory
5. Semiparametric families
   (a) location, scale, power, hazard, convolution
   (b) residual life families
6. Parametric families; origins
7. Competing risks
8. Coincidences of semiparametric families

BOOK REFERENCE

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

This book contains an extensive bibliography of over 500 references.

Change detection in a Heston type model

Gyula Pap, Tamás T. Szabó
Bolyai Institute, University of Szeged
papgy@math.u-szeged.hu, tszabo@math.u-szeged.hu

In the talk our main objective will be to define a process with the help of which we can introduce a change detection procedure for a special type of the Heston model. The process in question will be

$$\begin{cases} dY_t = (a - bY_t)\,dt + \sigma_1\sqrt{Y_t}\,dW_t, \\ dX_t = (m - \beta Y_t)\,dt + \sigma_2\sqrt{Y_t}\big(\varrho\,dW_t + \sqrt{1-\varrho^2}\,dB_t\big), \end{cases} \quad t \in \mathbf{R}_+, \tag{1}$$

where $a \in \mathbf{R}_+$, $b, m, \beta \in \mathbf{R}$, $\sigma_1, \sigma_2 \in \mathbf{R}_{++}$, $\varrho \in [-1, 1]$ and $(W_t, B_t)_{t\in\mathbf{R}_+}$ is a 2-dimensional standard Wiener process. We will restrict our attention to the ergodic case, i.e., when $a > \frac{1}{2}, b > 0$.

The solution to the first equation is the CIR process, which deserves independent interest. It will be instructive to restrict our attention to this equation first. Y_t is the scaling limit of a sequence of branching processes, it is therefore natural to try to apply our results from Pap and T. Szabó (2013). To this end we introduce the martingale $M_s = Y_s + \int_0^s (bY_r - a)dr$, and show that we have

$$\left(\frac{1}{\sqrt{T}}\int_0^{tT}\begin{bmatrix}\frac{1}{Y_s}\\ 1\end{bmatrix} dM_s\right)_{t\in[0,1]} \xrightarrow{\mathcal{D}} \left(I^{1/2}W_t\right)_{t\in[0,1]},\ T\to\infty, \tag{2}$$

where W is a standard two-dimensional Brownian motion and the matrix I depends on the parameters. In the next step we replace the parameters by their ML estimates given by Overbeck (1998). The resulting process converges in distribution—scaled appropriately—to a Brownian bridge. To prove this, we require an analogue of the strong LLN and the martingale CLT for continuous time-continuous state processes.

The joint parameter estimation for (1) follows Barczy *et al.* (2013), the setup of which does not cover our model. The methods are, however,

applicable and one can deduce the loglikelihood function of (Y_t, X_t). This is a result by G. Pap and M. Barczy; a sketch proof will be provided in the talk. We introduce a two-dimensional martingale by deducting from our process the conditional expectation on (Y_0, X_0), similarly to the martingale introduced before (2), and a result similar to (2) can be proved. The ML estimates can be substituted into this martingale to obtain a change detection method. Under the null hypothesis (i.e., no change), the procedure can also be thought of as a model-fitting test. Investigations under alternatives (as in, e.g., Pap and T. Szabó (2013) is under way, and possible approaches will be presented in the talk.

Keywords: Change-point detection; Heston model

Acknowledgements: This work was supported by by the Hungarian Scientific Research Fund under Grant No. OTKA T-079128, the Hungarian–Chinese Intergovernmental S & T Cooperation Programme for 2011-2013 under Grant No. 10-1-2011-0079 and the TÁMOP-4.2.2/B-10/1-2010-0012 project.

References

Barczy, M., Doering, L., Li, Z., and Pap, G. (2013). Parameter estimation for an affine two factor model. (arXiv:1302.3451)

Overbeck, L. (1998). Estimation for continuous branching processes. *Scandinavian Journal of Statistics*, Vol. 25, pp. 111–126.

Pap, G., T. Szabó, T. (2013). Change detection in INAR(p) processes against various alternative hypotheses. To appear in *Communications in Statistics: Theory and Methods* (arXiv:1111.2532)

Comparison of randomization techniques for non-causality hypothesis

Angeliki Papana
University of Macedonia, Thessaloniki, Greece
angeliki.papana@gmail.com

Catherine Kyrtsou
University of Macedonia, University of Strasbourg, BETA,
University of Paris 10, ISC-Paris, Ile-de-France
ckyrtsou@uom.gr

Dimitris Kugiumtzis
Aristotle University of Thessaloniki, Greece
dkugiu@gen.auth.gr

Cees Diks
University of Amsterdam, The Netherlands
C.G.H.Diks@uva.nl

When estimating Granger causality in multivariate time series it is important to assess the statistical significance of the estimate. While for linear estimators of Granger causality parametric significance tests have been established, for nonlinear estimators resampling techniques have to be employed and the null distribution for the causality estimator is formed from an ensemble of generated resampled (randomized) time series. We propose here a randomization technique that aims to improve the effectiveness of a test statistic and is particularly useful for distinguishing indirect causal effects. We use as a test statistic, the partial transfer entropy (Papana et al., 2002). Our randomization technique is similar to the time-shifting surrogates (Quian Quiroga et al., 2002). We form the surrogates by time-shifting the reconstructed vectors of both the driving and driven time series. By keeping the components of the reconstructed vectors unchanged, we aim to utilize all the information of the original dynamical properties of the systems that generate the time

series, while the null hypothesis of no causal effects is fulfilled. The proposed significance test is evaluated on a simulation study where we consider (1) time-shifted surrogates, (2) the suggested surrogates and (3) the stationary bootstrap (Politis & Romano, 1994). For the three resampling techniques, we examine two cases: (a) randomizing the driving time series, (b) randomizing both the driving and driven time series. The size of the test is improved using the suggested randomization method (2b).

Keywords: randomization, causality, significance, partial transfer entropy.

Acknowledgements: The research project is implemented within the framework of the Action "Supporting Postdoctoral Researchers" of the Operational Program "Education and Lifelong Learning" (Action's Beneficiary: General Secretariat for Research and Technology), and is co-financed by the European Social Fund (ESF) and the Greek State.

References

Papana A., Kugiumtzis D., Larsson P.G. (2012): Detection of Direct Causal Effects and Application to Electroencephalogram Analysis, *International Journal of Bifurcation and Chaos*, Vol. 22, N. 9, 1250222.

Politis D., Romano J.P. (1994): The Stationary Bootstrap, *Journal of the American Statistical Association*, Vol. 89, N. 428, pp. 1303-1313.

Quian Quiroga R., Kraskov A., Kreuz T., Grassberger P. (2002): Performance of different Synchronization Measures in Real Data: A Case Study on Electroencephalographic Signals, *Physical Review E*, Vol. 65, 041903.

Daniels' sequence and reliability of fiber composite material

Yuri Paramonov
Aeronautical Institute, Riga Technical University
yuri.paramonov@gmail.com

V. Cimanis, S.Varickis
Aeronautical Institute, Riga Technical University
chiamian@inbox.lv, serhiomail@inbox.lv

Connection between the cumulative distribution function (cdf), $F_X(s)$, of random variable (r.v.) X, which is the tensile strength of longitudinal items (LI) (fibers or strands) of unidirectional fiber reinforced composite material (UFRCM), and fatigue durability of this material can be defined by the random Daniels' sequence $\{s_0, s_1, s_2, ...\} : s_{i+1} = s/(1-\hat{F}_X(s_i))$, $i = 0, 1, 2, ...$, where $s_0 = s$, s is maximum value of cyclic stress, $\hat{F}_X(s)$ is estimate of $F_X(s)$ using results of tensile strength test of n LIs . RDS is non-decreasing sequence but there are two types of RDS. For RDS_1, when $s > max\ x(1-\hat{F}_X(x))$, items of RDS increase up to infinity and there is such i^* that $s_{i^*} < s_c$ but $s_{(i^*+1)} \geq s_c$, where s_c is critical stress corresponding to failure of UFRCM. The value of $N_D = i^*$ we call as *random Daniels fatigue life*. For RDS_2, when $s < max\ x(1-\hat{F}_X(x))$, there is such smallest i^{**} that $s_{i^{**}+1} = s_{i^{**}}$. The s_c never will be reached and N_D is equal to infinity. The maximum value of s for which this phenomenon takes place we call *random Daniels fatigue limit*: $S_D = max\ x(1-\hat{F}_X(x))$.
The character of RDS is very similar to character of some physical parameters; existence of RDS_2 explains existence of fatigue limit (infinite life). But the value of N_D at the middle value of stress is too small, the value of S_D is too large in comparison with real values of UFRCM. Required decreasing of S_D can be reached by decreasing of local tension strength, X_L (for example, $X_L = X/k_f, k_f \geq 1$), in comparison with

nominal strength X. The values of N_D for RDS_1 can be "stretched out" by the use of Markov chains (MCh) [1]. Let $R_{(s,x_{L,1:n})^*}$ is the population of pairs $(s, x_{L,1:n})^*$, corresponding to RDS_1, where $(x_{L,1:n})$ is sample from population with cdf $F_{X_L}(x)$. Let first $(n_D + 1)$ states of MCh correspond to first $(n_D + 1)$ items of RDS ; n_D is realization of N_D; $(n_D + 2)$-th state, s_{n_D+1}, is absorbing state (local stress become more than s_c). The estimate $\hat{F}_{X_L}(x)$ defines the realization of random matrix of transition probabilities, $\hat{P}$, the cdf of time to absorption in MCh with this matrix, $\hat{T} = \hat{T}_1 + \hat{T}_2 + ... + \hat{T}_{n_D+1}$, where $\hat{T}_i$ is the time of transition from state s_{i-1} to state s_i . Mean value and variance of $\hat{T}$ can be easily calculated, taking into account that r.v. $\hat{T}_i$ is geometric r.v. The mean of random cdf $\hat{F}_T(t)$, mean value and quantiles of r.v. T as function of initial stress s should be calculated under condition that pair $(s, x_{L,1:n}) \in R_{(s,x_{L,1:n})^*}$. It is supposed, that the number of cycles up to failure of UFRCM in fatigue test $N_{k_m D} = k_m T$, where k_m is scale factor. By the use of the considered model we can make fitting of fatigue test data and predict fatigue life changes as consequence of tensile strength parameter changes. Numerical example is given.

Keywords: Markov chain, tensile strength, fatigue life.

References

Paramonov Yu, Kuznetsov A, Kleinhofs M. (2011): *Reliabilty of fatigue-prone airframes and composite materials*, Riga: RTU. $(http://gnedenko-forum.org/library/Paramonov/Reliability\_Paramonov.pdf)$

Minimax decision for reliability of aircraft fleet and airline

Yuri Paramonov
Aeronautical Institute, Riga Technical University
yuri.paramonov@gmail.com

Maris Hauka, Sergey Tretyakov
Aeronautical Institute, Riga Technical University
maris.hauka@gmail.com, sergejs.tretjakovs@gmail.com

The planning of inspection interval is considered. The purpose is to limit probability of any fatigue failure (FFP) in the fleet of N aircraft (AC) and to provide the economical effectiveness of airline (AL) under limitation of fatigue failure rate (FFR). The decision is based on result of acceptance fatigue tests during which the estimates of parameters of fatigue crack growth trajectory are obtained. If result of acceptance test is too bad then this new type of aircraft will not be used in service. The redesign of this project should be done. If result of acceptance tests is too good then the reliability of aircraft fleet and an airline will be provided without inspection. So there is maximum of mean FFP $p_{fNw} = E_R((1-w)^R)$ as function of parameter θ. This maximum is defined by the distribution of random times when fatigue crack becomes detectable, T_d, and critical, T_c. Here R is the total random number of inspections in whole fleet of new type of aircraft before first failure in this fleet, taking into account that operation of every AC is limited by retirement at specified service life (SL),t_{SL}, or before first failure in this fleet;$E_R(.)$ is symbol of expectation in accordance with cdf of r.v. R ; w is probability of crack detection when there is crack (human factor). If parameter θ of cdf of T_d, T_c and cdf of r.v. R, $F_R(r;\theta,d)$ are known then the interval between inspection d should be chosen under condition of limitation of p_{fNw} for all θ allowable for $F_R(r;\theta,d)$. In order to take into account the economic effectiveness of AL the pro-

cess of operation of one AC of specific AL is considered as absorbing semi-Markov process with reword (SMPW) with $(n+4)$ states, where $n = ([t_{SL}/d] - 1)$ is the inspection number. The states $E_1, E_2, ..., E_{n+1}$ correspond to operation of AC in time intervals between inspection: $[t_0, t_1), [t_1, t_2), ..., [t_n, t_{SL})$. States E_{n+2} (retirement), E_{n+3} (fatigue failure), and E_{n+4} (fatigue crack detection) correspond to return of corresponding Markov chain to state E_1 (new aircraft is purchased and AL operation returns to first interval). Using matrix of transition probabilities P for SMPW, which is defined by parameter θ, we can get the vector of stationary probabilities and the airline gain, which is defined by the reward related to successful transition from one operation interval to the following one ; by the cost of one inspection; by the cost of transition to states E_{n+3}, E_{n+4} and E_1. If θ is known we calculate the gain of AL as function of n, $g(n, \theta)$, and choose the number n_g corresponding to the maximum of gain. Then we calculate FFR as function of n, $\lambda_F(n, \theta)$, and choose n_λ in such a way that for any $n \geq n_\lambda$ the function $\lambda_F(n, \theta)$ will be equal or less than some value λ_{FD} (the "designed" FFR). And finally we choose $n = n_{g\lambda}(\lambda_{FD}, \theta) = max(n_g, n_\lambda)$.
In fact, we do not know θ. Here it is shown, how using the estimate of this parameter, $\hat{\theta}$, (after acceptance test) one of two decisions should be made: 1) to do the redesign of new type of AC if result of test is "too bad" ; 2) make the choice of the number of inspection as function of $\hat{\theta}$ in such a way that this inspection number provides maximum of expectation of AL gain under limitation of FFR or FFP.

Keywords: inspection, reliability, airline, economic effectivnes.

References

Paramonov Yu, Kuznetsov A, Kleinhofs M. (2011): *Reliabilty of fatigue-prone airframes and composite materials*, Riga: RTU. $(http://gnedenko-forum.org/library/Paramonov/Reliability\_Paramonov.pdf)$

Asymptotic Permutation Tests in Factorial Designs - Part II

Markus Pauly
University of Düsseldorf, Institute of Mathematics, Universitätsstrasse 1,
40225 Düsseldorf, Germany
pauly@math.uni-duesseldorf.de

Permutation and randomization procedures are classical tools for statistical inference. However, permutation tests are often developed under the assumption of exchangeability of the data or are based on simulation results of heuristic procedures. For non-exchangeable observations, as in the classical Behrens-Fisher settings, the permutation procedure has to be dealt with care. In particular, Huang et al. (2006) have exemplified that this approach may lead to invalid results. Nevertheless, in the previous talk by Edgar Brunner it has been demonstrated that the permutation idea can be applied in general unbalanced, heteroscedastic factorial designs where the error distributions may come from a quite general class. In the present talk it is illustrated why the permutation idea works in this generality. In a first step the theory is explained for comparing the means in unpaired two-sample designs with unequal variances. Moreover, it is pointed out that a similar approach is also valid in other two-sample problems such as paired observations, variances, treatment effects or correlations. This idea is then extended to general factorial designs where the errors in the different samples may even be heterogeneous. This is enabled by a detailed investigation of the asymptotic distribution of the corresponding permutation statistic. The main theorem states that its permutation distribution asymptotically mimics the null distribution of the related test statistics of the respective hypotheses. Note, that this property is not shared by the ANOVA-type statistic (Brunner et al., 1997) and the Welch-James test (Johansen, 1980), which are in general only approximations, even for large samples. The proof of the main result is based on conditional central limit

theorems derived by Janssen (2005) and Pauly (2011). The simulation results presented in the previous talk by Edgar Brunner are in agreement with these theoretical results. For the ease of convenience, however, this asymptotic property is graphically demonstrated instead of presenting the sophisticated mathematical proof in detail (Pauly et al., 2013). Furthermore, for exchangeable data, this permutation test possesses the favorable characteristic of being exact for finite sample sizes.

Keywords: Analysis of variance, permutation tests, heteroscedastic designs.

References

Brunner, E., Dette, H., Munk, A. (1997): Box-Type Approximations in Nonparametric Factorial Designs, *Journal of the American Statistical Association* 92, pp. 1494-1502.

Huang, Y., Xu, H., Calian, V., and Hsu, J.C. (2006): To permute or not to permute, *Bioinformatics* 22, pp. 2244-2248.

Janssen, A. (2005): Resampling Students t-Type Statistics, *Annals of the Institute of Statistical Mathematics* 57, pp. 507-529.

Johansen, S. (1980): The Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression, *Biometrika* 67, pp. 85-92.

Pauly, M., Brunner, E., and Konietschke, F. (2013): Asymptotic Permutation Tests in General Factorial Designs, Preprint.

Pauly, M. (2011): Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics* 5, pp. 41-52.

Nonparametric Change Detection Under Dependent Noise

Mirosław Pawlak
Dept. of Electrical & Computer Eng., University of Manitoba
Miroslaw.Pawlak@ad.umanitoba.ca

In this paper a nonparametric version of the sequential signal detection problem is studied. Our signal model includes a class of time-limited signals for which we collect data in the sequential fashion at discrete points in the presence of correlated noise. Hence, suppose we are given noisy measurements

$$y_k = f(k\tau) + \epsilon_k,$$

where τ is the sampling period, $\{\epsilon_k\}$ is a zero mean noise process, and $f(t)$ is an unspecified signal which belongs to some signal space. We are interested in the following on-line detection problem: given a reference (target) parametric class of signals $\mathcal{S} = \{f(t;\theta) : \theta \in \Theta\}$, where Θ is a subset of a finite dimensional space, we wish to test the null hypothesis $H_0 : f \in \mathcal{S}$ against an arbitrary alternative $H_1 : f \notin \mathcal{S}$. Throughout the paper, we assume that the signal $f(t)$ of interest is observed over a finite time frame, i.e., $t \in [0, T]$, for some $0 < T < \infty$. For such a setup we introduce a novel signal detection algorithm relying on the post-filtering smooth correction of linear sampling schemes. Our detector is represented as a normalized partial-sum continuous time stochastic process, for which we obtain a functional central limit theorem under weak assumptions on the correlation structure of the noise. Particularly, our results allow for noise processes such as ARMA and general linear processes as well as α-mixing processes. The established limit theorems allow us to design monitoring algorithms with the desirable level of the probability of false alarm and able to detect a change with probability approaching one.

Keywords: change-point problems, sequential detection, nonparametric regression, correlated noise, Donsker's Theorem.

Acknowledgements: This work was supported by NSERC

References

Poor, H.V. and Hadjiliadis, O. (2009): *Quickest Detection*. Cambridge: Cambridge University Press.

Pawlak, M., Rafajłowicz E., and Krzyżak, A (2003): Post-filtering versus pre-filtering for signal recovery from noisy samples. *IEEE Trans. Information Theory*, Vol. 49, pp. 569-587.

Rafajłowicz E., Pawlak, M., and Steland, A (2010): Nonparametric sequential change-point detection by a vertically trimmed box method. *IEEE Trans. Information Theory*, Vol.56, pp. 3621-3634.

SSA change-point detection and applications

Andrey Pepelyshev
RWTH Aachen University
pepelyshev@stochastik.rwth-aachen.de

Change-point detection is an important problem in applied statistics. The present work is devoted to investigation of a promising approach to sequential change-point detection based on Singular Spectrum Analysis (SSA) proposed in (Moskvina, Zhigljavsky, 2003). We show that this approach is capable to detect changes in mean, amplitude, frequency and phase of damped periodic series and changes in coefficients of linear recurrent formulas governing series observed in the presence of noise. We also illustrate the performance of SSA for discovering changes in environmental data and quality control of a photovoltaic modules production line.

Keywords: Singular Spectrum Analysis, change-point detection, quality control.

References

Moskvina, V., Zhigljavsky, A. (2003). An algorithm based on singular spectrum analysis for change-point detection. Communications in Statistics - Simulation and Computation, 32(2), 319–352.

Implementation of Bayesian methods for sequential design using forward sampling

Juergen Pilz
Alpen-Adria-University Klagenfurt
Department of Statistics
juergen.pilz@aau.at

Keywords: Bayes sequential design, backward induction, forward sampling, expected utility maximization.

We deal with decision making and utility maximization in one- and two-sample problems in a Bayesian sequential framework. When data samples arrive over time then Bayes' rule may be used sequentially just by treating the current posterior as the prior for the next update. In the Bayesian framework the problem of deciding whether or not to continue sampling is usually tackled by the method of backward induction, see e.g. DeGroot (1970). This is, however, difficult to implement when the number of decision stages is medium to large.
As an alternative we present a forward sampling algorithm, based on ideas developed in Carlin and Louis (2009), which can produce the optimal stopping boundaries for a broad class of decision problems at far less computational expense. The computational complexity grows only linearly in the number of decision stages, instead of exponentially. In the case of a k-parameter exponential family likelihood, the class of possibly optimal sequential rules includes arbitrary functions of the k-vector of sufficient statistics. We illustrate the approach for the one-sample problem of testing a normal mean with unknown variance and for the two-sample problem of comparing Poisson distributions with dif-

ferent rates. Finally, we indicate extensions of our forward sampling approach based on ideas propagated by Brockwell and Kadane (2003) and Mueller et al. (2007).

References

Brockwell, A.E., Kadane, J.B. (2003): A gridding method for Bayesian sequential decision problems, *Journal of Computational and Graphical Statistics*, Vol. 12, pp. 566-584.

Carlin, B.P., Louis, T.A. (2009): *Bayesian Methods for Data Analysis*, Chapman and Hall, Boca Raton.

DeGroot M.H. (1970): *Optimal Statistical Decisions*, McGraw-Hill, New York.

Mueller, P., Berry, D., Grieve, A., Smith, M., Krams, M. (2007): Simulation-based sequential Bayesian design, *Journal of Statistical Planning and Inference*, Vol. 137, pp. 3140-3150.

Model-free prediction intervals for regression and autoregression

Dimitris N. Politis
Department of Mathematics
University of California at San Diego
La Jolla, CA 92093-0112, USA
dpolitis@ucsd.edu

With the advent of widely accessible powerful computing in the late 1970s, computer-intensive methods such as the *bootstrap* created a revolution in modern statistics; see Efron and Tibshirani (1993) and the references therein. However, even bootstrapping has not been able to give a definitive solution to prediction. Consider the popular model-based, i.e., residual-based, bootstrap method for the construction of prediction intervals for regression (and autoregression). Even when the model assumptions are correct, bootstrap prediction intervals have been plagued by *undercoverage* even in the simplest case of linear regression; see Olive (2007) for a recent review of the state-of-the-art.

Furthermore, in many instances the model assumptions can be violated in which case any model-based inference will be invalid. In this talk, the problem of statistical prediction is revisited with a view that goes beyond the typical parametric/nonparametric dilemmas in order to reach a fully model-free environment for predictive inference, i.e., point predictors and predictive intervals. The *'Model-Free (MF) Prediction Principle'* of Politis (2007, 2010) is based on the notion of transforming a given set-up into one that is easier to work with, namely i.i.d. or Gaussian. The two important applications are regression and autoregression whether an additive parametric/nonparametric model is applicable or not. To elaborate, consider a vector of observed data $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ where the objective is to predict Y_{n+1}. If the $Y_1, \ldots, Y_n$ were i.i.d., the prediction problem would be trivial. The essence of the MF Prediction Principle is to *transform* the Y–data to-

wards "i.i.d.–ness", and thus attempt to trivialize the problem. In summary, the MF Prediction Principle amounts to using the structure of the problem in order to *find an invertible transformation H_m that can map the non-i.i.d. vector $\underline{Y}_m$ to a vector $\underline{\epsilon}_m = (\epsilon_1, \ldots, \epsilon_m)'$ that has i.i.d. components;* here m could be taken equal to either n or $n+1$ as needed. Letting H_m^{-1} denote the inverse transformation, we have

$$\underline{Y}_m \stackrel{H_m}{\longmapsto} \underline{\epsilon}_m \text{ and } \underline{\epsilon}_m \stackrel{H_m^{-1}}{\longmapsto} \underline{Y}_m. \tag{1}$$

If the practitioner is successful in implementing the MF procedure, i.e., in identifying the transformation H_m to be used, then the prediction problem is reduced to the trivial one of predicting i.i.d. variables. To see why, note that eq. (1) with $m = n+1$ yields $\underline{Y}_{n+1} = H_{n+1}^{-1}(\underline{\epsilon}_{n+1}) = H_{n+1}^{-1}(\underline{\epsilon}_n, \epsilon_{n+1})$. But $\underline{\epsilon}_n$ can be treated as known given the data $\underline{Y}_n$; just use eq. (1) with $m = n$. Since the unobserved Y_{n+1} is just the $(n+1)^{th}$ coordinate of vector $\underline{Y}_{n+1}$, the former can also be expressed as a function of the unobserved ϵ_{n+1}.

Keywords: Bootstrap, predictive distributions, prediction intervals.

References

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

Olive, D.J. (2007). Prediction intervals for regression models. *Comput. Statist. and Data Anal.*, 51, pp. 3115–3122.

Politis, D.N. (2007). Model-free prediction, in *Bulletin of the International Statistical Institute–Volume LXII*, 22–29 August 2007, Lisbon, pp. 1391-1397.

Politis, D.N. (2010). Model-free model-fitting and predictive distributions, Discussion Paper, Department of Economics, UCSD. Retrieved from: `http://escholarship.org/uc/item/67j6s174`. To appear as an Invited Discussion Paper in journal *Test* in 2013.

Estimation of Change-in-Regression-Models based on the Hellinger Distance for Dependent Data

Annabel Prause
Institute for Statistics, RWTH Aachen University, Germany
prause@stochastik.rwth-aachen.de

Ansgar Steland, Mohammed Abujarad
Institute for Statistics, RWTH Aachen University, Germany
steland@stochastik.rwth-aachen.de,
abujarad@stochastik.rwth-aachen.de

We study minimum Hellinger distance estimation (MHDE) based on kernel density estimators for a large class of parametric regression models including models with a change in the regression function. To be more precise we consider for an observed random sample

$$(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n) \tag{1}$$

of bivariate random vectors, whose density belongs to the parametric family $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^l$, the (non-linear) regression model $Y_t = g_{\theta_0}(X_t) + \sigma_{\theta_0}(X_t)\varepsilon_t, \;\; t \geq 0$. If we take the function g as

$$g_\eta(x) = \widetilde{g}_{\eta_1}(x) + [\widetilde{g}_{\eta_2}(x) - \widetilde{g}_{\eta_1}(x)]\mathbf{1}_{\{x \geq c\}}, \;\; \eta = (\eta_1, \eta_2), \tag{2}$$

with $c \in \mathbb{R}$ known we get an example for a model with a change in the regression function. Moreover, also the (non-linear) autoregressive model of order one can be considered as a special case of the general regression model.

The idea of the Minimum Hellinger distance approach now is to minimize the L_2 distance between the square roots of $p_\theta(x, y)$ and some nonparametric density estimator $\widehat{q}_n(x, y)$, which in this work is chosen to be a two dimensional kernel density estimator.

It is shown that consistency and asymptotic normality of the MHDE basically follow from the uniform consistency of the density estimate

and the validity of the central limit theorem for its integrated version. Those conditions hold true for i.i.d. as well as for strong mixing observations under fairly general conditions. In the mixing case these conditions include smoothness conditions on the kernel functions and the joint density of (X_0, Y_0) as well as a certain decay of the α-mixing coefficients. As an important difference to the case of univariate observations, the asymptotic normality of the MHDE can only be shown when correcting with a certain bias term. This bias term can either be random or deterministic and is a consequence of conflicting orders of convergence of certain terms that appear in the decomposition of $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$. To the best of our knowledge no other MHD results have been developed for multivariate dependent data for our class of parametric regression models yet, i.e. for time series.

Keywords: central limit theorem, consistency, density estimation, mixing, time series.

References

Beran R. (1977): Minimum Hellinger Distance Estimates for Parametric Models, *The Annals of Statistics*, Vol. 5, N. 3, pp. 445-463.

Hansen B.E. (2008): Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory*, Vol. 24, pp. 726-748.

Liebscher E. (1996): Central limit theorems for sums of alpha-mixing random variables, *Stochastics and Stochastic Reports*, Vol. 59, N. 3-4, pp. 241-258.

Tamura R.N., Boos D.D. (1986): Minimum Hellinger Distance Estimation for Multivariate Location and Covariance, *Journal of the American Statistical Association*, Vol. 81, N. 393, pp. 223-229.

The Clouds and the Sea Surface Stochastic Models in the Atmosphere Optics

Sergei M. Prigarin
Novosibirsk State University, and Institute of Computational Mathematics and Mathematical Geophysics, Russian Academy of Sciences (Siberian Branch), Novosibirsk, Russia
sergeim.prigarin@gmail.com

This paper deals with numerical stochastic structure models of the clouds and of the sea surface. The models are based on nonlinear transformations of Gaussian fields and are applicable to solving problems of the atmosphere-ocean optics.

The light interaction with clouds and underlying surface is a major factor affecting the radiation balance in the Earth's atmosphere. The stochastic structure of geophysical fields brings about a considerable uncertainty in the climate and radiation transfer models. That is why it is a challenging problem to construct numerical models of clouds and underlying surface, including the sea surface roughness with allowance for their random optical properties and geometry. A simplified description of the models that were used to simulate the stochastic structure of the clouds and the sea surface can be presented by the formula

$$w(x) = F(G(x))$$

where $G(x)$ is a numerical model of a Gaussian field and F is a pointwise nonlinear function. Spectral models from (Mikhailov, 1978) and (Prigarin, 2001) were used to simulate Gaussian random fields:

$$G(x) = \sum_j \left[\xi_j \cos < \lambda_j, x > + \eta_j \sin < \lambda_j, x >\right].$$

Here $< ., . >$ denotes the scalar product, while distributions of random variables ξ_j, η_j, λ_j and the number of random harmonics should be chosen according to desired properties of the field.

The main objective of this paper is to present some new simulation examples to demonstrate how the method works, in addition to the previous investigations described, for example, in (Kargin, Oppel, Prigarin,1999), (Prigarin, 2005), (Prigarin, Marshak, 2009).

Keywords: numerical simulation, random fields, clouds, sea surface, threshold and spectral models.

Acknowledgements: The work was supported by the Russian Foundation for Basic Research (grant nos. 11-01-00641, 12-05-00169).

References

Kargin B.A., Oppel U.G., Prigarin S.M. (1999): Simulation of the undulated sea surface and study of its optical properties by Monte Carlo method, *Proc. SPIE*. Vol. 3583, pp.316-324.

Mikhailov G.A. (1978): Numerical construction of a random field with given spectral density, *Dokl. USSR Ac. Sci.*, Vol. 238, N. 4, pp. 793-795.

Prigarin S.M. (2001): *Spectral Models of Random Fields in Monte Carlo Methods*, VSP, Utrecht.

Prigarin S.M. (2005): *Numerical Modeling of Random Processes and Fields*, ICMMG Publisher, Novosibirsk [in Russian].

Prigarin S., Marshak A. (2009): A simple stochastic model for generating broken cloud optical depth and cloud top height fields, *Journal of the Atmospheric Sciences*, Vol. 66, N. 1, pp. 92-104.

Simulation of Extreme Ocean Waves by Peaks of Random Functions

Sergei M. Prigarin
Novosibirsk State University, and Institute of Computational Mathematics and Mathematical Geophysics, Russian Academy of Sciences (Siberian Branch), Novosibirsk, Russia
sergeim.prigarin@gmail.com

Kristina V. Litvenko
Institute of Computational Mathematics and Mathematical Geophysics, Russian Academy of Sciences (Siberian Branch), Novosibirsk, Russia
litchristina@gmail.com

This paper deals with numerical simulation of formation and development of the extreme ocean waves by using specific models of random fields. The extreme waves known as "rogue" or "freak" waves represent a poorly understood natural phenomenon whose existence was distrusted because of the absence of reliable evidence. In contrast to tsunami waves, the solitary extreme waves are of 20, 30, or more meters of height, essentially exceeding the heights of other waves, appearing suddenly, and vanishing far from the shore without visible causes; sometimes this occurs in a slight sea under relatively light wind. For the first time a rogue wave was instrumentally detected only in 1995, and nowadays strenuous efforts are mounted to observe the extreme waves and to study the rogue wave phenomenon both theoretically and experimentally (see, for example, Pelinovsky and Kharif (2008)).

For numerical simulation of the extreme waves, we used conditional spectral models of random fields proposed by Prigarin (1998). It is assumed that the sea surface roughness is sufficiently well described by a spatio-temporal random field. Along with the spectrum of a random field, the simulation of the extreme wave requires additional informa-

tion concerning the wave profile, i.e., the field of the sea surface elevations should be specified at certain points at given time moments. The conditional spectral models allow us to numerically simulate a set of independent spatio-temporal implementations of the sea level passing through given points and hence to study typical features of the development and spread of the extreme waves. In particular, an unexpected result of numerical experiments was the appearance of groups consisting of three extreme waves, whereas the extreme level of roughness was fixed at one point only (see Prigarin and Litvenko (2012)). This type of extreme waves is well known and is called the "three sisters".

Keywords: numerical simulation, extreme ocean waves, random fields, conditional spectral models, freak-waves, rogue waves.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research (grant 11-01-00641).

References

Pelinovsky E., Kharif Ch., Eds. (2008): *Extreme Ocean Waves*, Springer, Berlin.

Prigarin S.M. (1998): Conditional spectral models of Gaussian homogeneous fields. *Russian Journal of Numerical Analysis and Mathematical Modelling*, Vol. 13, N. 1, pp. 57-68.

Prigarin S.M., Litvenko K.V. (2012): Conditional spectral models of extreme ocean waves, *Russian Journal of Numerical Analysis and Mathematical Modelling*, Vol. 27, N. 3, pp. 289-302.

A conjecture about BIBDs

Dieter Rasch
University of Life Sciences, Vienna, Austria
dieter.rasch@boku.ac.at

Friedrich Teuscher, L. Rob Verdooren
FBN Dummerstorf, FB Genetik und Biometrie, Germany, Danone Research, Wageningen, The Netherlands
teuscher@fbn-dummerstorf.de, Rob.Verdooren@danone.com

The main purpose of this paper is to collect arguments supporting a conjecture. We consider v elements $(1, 2, 3, , v)$ called treatments which have to be allocated to b sets called blocks. Such an allocation is called a block design.
A (completely) *balanced incomplete block design* (BIBD) is a proper and equireplicated incomplete block design with the additional property that all treatment pairs occur in the same number λ of blocks. A BIBD is called symmetrical if $b = v$ and $r = k$. Necessary conditions for the existence of a BIBD are:

$$b \geq v \quad (1)$$
$$vr = bk \quad (2)$$
$$\lambda(v-1) = r(k-1) \quad (3)$$

Parameters fulfilling the necessary conditions are called admissible. A BIBD is called trivial if it is identical to the set of all $\binom{n}{k}$ possible k-tupels, then we have $b = \binom{v}{k}, r = \binom{v-1}{k-1}$ and $\lambda = \binom{v-2}{k-2}$. A BIBD for a pair (v, k) is called elementary if it can not be split into at least two BIBDs for this (v, k). A BIBD for a pair (v, k) is called the smallest BIBD if for this pair r (and by this also k and λ) is a minimum. A complementary design to a given BIBD for a pair (v, k) is the design for $(v, v-k)$ with the same number of blocks so that each block of the complementary design contains the treatments not in the corresponding block of the original BIBD.

Of course a smallest (v, k) BIBD is elementary but not all elementary BIBDs are smallest.
Conjecture: If $3 < k < \frac{v}{2}$ the case $(v, k) = (8, 3)$ is the only one where the trivial BIBD is elementary.
This conjecture is supported by the fact that there is no $b < 56$ in the case $(8, 3)$ for which the necessary conditions are fulfilled. The following theorem also supports the conjecture:
Theorem: The conjecture is correct if at least one of the following conditions is fulfilled:

a) $v < 26$
b) $k < 6$
c) for $v > 8$ if v is prime or a prime power.

Keywords: Balanced incomplete block designs (BIBDs), trivial BIBDs, necessary conditions for the existence of BIBDs.

References

Abel, R.J.R. Bluskov, I. and Greig, M. (2001): Balanced Incomplete Block Designs with block size 8, J. *Combin. Des.* 9, 233268.

Abel, R. J. R.; Bluskov, I.and Craig, M. (2004): Balanced incomplete block designs with block size 9: Part III *AUSTRALASIAN JOURNAL OF COMBINATORICS*,Volume 30, Pages 5773.

Colbourn, C.J. and Dinitz, J.H. (eds.) 2nd ed. (2006): *Handbook of Combinatorial Designs*, Chapman and Hall.

Hanani, H. (1975): Balanced incomplete block designs and related designs, *Discrete Mathem.* 11, 275-289.

Rasch, D.; Pilz, J., Verdooren, R., Gebhardt, A. (2011): *Optimal Experimental Design with R*, Chapman and Hall, Boca Raton.

Rasch, D. and Verdooren, L. R. (2013): *A conjecture on BIBs*. Paper to be presented at MODA 10 International Workshop on Model-Oriented Design and Analysis agw Lubuski, Poland, 10-14 June 2013.

Stochastic Modification of a Knapsack Problem

Marina Rebezova
Transport Clearing House
rebezova@tch.ru

Nikolay Sulima, Roman Surinov
Exigen Services Latvia, Rosaviaconsorcium
sulima@gmail.com, surinov@rosaviaconsorcium.ru

A considered problem can be formulated as a problem of optimal planning of various project development. Let us introduce the following notations: k - project's number, $i = 1, 2, \ldots, k$; n_i - number of development variants for the i-th project; $z_{i,j}$ - expenditure on a realization of the j-th variant for the i-th project; $c_{i,j}$ - reward on a realization of the j-th variant for the i-th project; d_{i,j,i^*,j^*} - additional reward, conditioning by simultaneously realization of the j-th variant for the i-th project and the j^*-th variant for the i^*-th project.

Additionally a total size of money Z is known, that must be distributed between different projects. For that each project has to be developed accordingly to unique variant only.

To give mathematical setting of the problem, let us introduce Boolean variables $x_{i,j}$: $x_{i,j} = 1$, if the i–th project is developed by the j–th variant, and 0, otherwise.

Now it is possible forming the considered problem as a modification of a knapsack problem (Gilmore, Gomory, 1965), (Hu, 1970).

Maximize a total reward:

$$f(x) = \sum_{i=1}^{k} \sum_{j}^{n_i} (c_{i,j} - z_{i,j})\, x_{i,j} + \sum_{i=1}^{k} \sum_{j}^{n_i} \sum_{i^*=1}^{k} \sum_{j^*}^{n_i} d_{i,j,i^*,j^*} x_{i,j} x_{i^*,j^*}, \tag{1}$$

subject to restrictions:

$$\sum_{j=1}^{n_i} x_{i,j} = 1,\ i = 1, \ldots, k,\ \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j} z_{i,j} \leq Z. \qquad (2)$$

Now we suppose that reward is a random variable $C_{i,j}$, given known distribution function $F_{i,j}(x) = \mathsf{P}\{C_{i,j} \leq x\}$. Above considered value $c_{i,j}$ is its mean: $c_{i,j} = \mathsf{E}\,(C_{i,j}) = \int x dF_{i,j}(x)$. We suppose that additional rewards $\{d_{i,j,i^*,j^*}\}$ are constants.

In this case constants $c_{i,j}$ in (1) are replaced by random variables $C_{i,j}$. Therefore total reward (1) becomes a random variable. In the above considered case, average value of the total reward (1) was maximized. Further we consider *ruin probability* as optimization criterion: it is a probability that total reward is less than prescribed value R^*:

$$F(R^*) = \mathsf{P}\{R \leq R^*\}, \qquad (3)$$

where R is calculated by formula (1) by changing $c_{i,j}$ by $C_{i,j}$.

Now we must find a solution that satisfies (2) and minimizes ruin probability (3). Paper contains corresponding numerical algorithm. Some statistical problems and practical applications are considered too.

Keywords: knapsack problem, ruin probability.

References

Hu T.C. (1970): *Integer Programming and Network Flows*, Addison-Wesley, Reading, Massachusetts.

Gilmor P.C., Gomori R.E. (1965): The Theory of Computation of Knapsack Functions, *Journal of ORSA*, Vol. 14, N. 6, pp. 1045-1974.

Advances in Multilevel Modeling: a review of methodological issues and applications

Giulia Roli, Paola Monari
Dipartimento di Scienze Statistiche, Universit di Bologna
g.roli@unibo.it, paola.monari@unibo.it

Multilevel modeling is a recently new class of statistical methods, firstly introduced in 1987 by Goldstein and later by Raudenbush and Bryk (1992) and Hox (1995). This approach for data analysis is a generalization of linear and generalized linear regressions, when the structure of data is nested, i.e. base level units are grouped into higher level units involving their own variability and dependencies among the related observations. The nested or *multilevel* structure of data is a common phenomenon, especially in behavioral and social sciences, where the study of the relationship between individuals and society is of crucial importance and, thus, the dependence of data becomes a focal interest of the research. Moreover, the hierarchy of data can be generated by the sampling design, such as in the multi-stage sampling, which is frequently employed in the traditional surveys to reduce the costs of data collection. In such cases, data dependence is treated as a nuisance which requires further adjustments during the analysis.

Mainly thanks to the wide range of applicability and the great increase of statistical softwares (de Leeuw and Kreft, 1999), in the last decades multilevel modeling has enjoyed an explosion of published papers and books in both methodological and application field. Currently, there is a need to not only develop the research on multilevel approach for the analysis of complex data, but also to have instructions to properly address the usage.

This work aims at summarizing methodological aspects related to multilevel models, illustrating good-practices, advantages and limits and reviewing applications in various fields, such as socio-economic, educational, health and medical sciences. To date, only few reviews are

available to report practices of multilevel applications, mainly restricted to education field (see, e.g., Dedrick *et al.*, 2009). We further focus our attention on the latest advances of multilevel modeling towards the inclusion of latent variables, such as multilevel structural equation and latent class models (Skrondal and Rabe-Hesketh, 2004), and the increasing use of Bayesian inference approach (Gelman *et al.*, 2003).

Keywords: multilevel model, hierarchical regression, nested data.

References

Bryk A.S., Raudenbush S.W. (1992): *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods*, First Edition, Newbury Park, CA: Sage Publications.

Dedrick R.F., Ferron J.M., Hess M.R., Hogarty K.Y., Kromrey J.D., Lang T.R., Niles J., Lee R.(2009), Multilevel Modeling: A Review of Methodological Issues and Applications, *Journal: Review of Educational Research*, vol. 79, no. 1, pp. 69-102.

de Leeuw J., Kreft I.G.G. (1999): Software for Multilevel Analysis, *Department of Statistics Papers, Department of Statistics, UCLA, UC Los Angeles*, http://escholarship.org/uc/item/5dk9x4p0.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2003): *Bayesian data analysis*, Second Edition, Chapman & Hall, New York.

Goldstein H (1987): *Multilevel models in education and social research*, London, England: Charles Griffin & Co; New York, NY, US: Oxford University Press.

Hox J.J. (1995): *Applied Multilevel Analysis*, Amsterdam: TT-Publikaties.

Skrondal, A. and Rabe-Hesketh, S. (2004): *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL: Chapman & Hall/CRC.

Kriging based adaptive sampling in metrology

Daniele Romano
University of Cagliari, Italy
daniele.romano@dimcm.unica.it

Statistical sampling is a fundamental tool in science, and metrology is no exception. The merit of a sample is its efficiency, i.e. a good trade-off between the information collected and the sample size. Although the sample points - i.e. the sites - are ordinarily decided upon prior to the measurements, a different option would be to select them one at a time. This strategy is potentially more informative as the next site can be chosen on the basis of the measurements made up to then. The core of the method is to drive the next-site selection by a non-parametric model known as kriging, namely a stationary Gaussian stochastic process with a given autocorrelation structure. The main feature of this model is the ability to promptly reconfigure itself, changing the pattern of the predictions and their uncertainty each time a new measurement comes in. Since the model is re-estimated after each added point, the sampling procedure is an adaptive one. The next sampling site can be selected via a number of model-based criteria, inspired by the principles of reducing prediction uncertainty or optimizing an objective function, or a combination of the two.

The methodology has been already demonstrated by the author [see Ascione *et al.*(2013) and Pedone *et al.*(2009)] in a metrological application, i.e. the design of adaptive inspection plans for measuring geometric errors using touch-probe Coordinate Measuring Machines (CMM). Results have shown that both the non-adaptive statistical plans (Random, Latin Hypercube sampling, uniform sampling) and two adaptive deterministic plans from the literature were largely outperformed by the proposed plans both in terms of accuracy and cost. Here we further investigate a number of important questions related to adaptive kriging:

what the best trade-off is between the number of adaptive and non-adaptive points (e.g. chosen according to uniform coverage), which next-site selection criteria are more suitable for particular objectives (e.g. capturing extreme values vs minimizing the root mean square prediction error (RMSPE)), which kriging correlation function is preferable, whether the Limit kriging predictor (a modification of the standard predictor) should be used or not and when the sampling should be stopped.

Keywords: Adaptive sampling, Kriging, Metrology, Jackknife variance.

References

Ascione R., Moroni G., Petro' S., Romano D. (2013): Adaptive inspection in coordinate metrology based on Kriging models, *Precision Engineering*, Vol. 37, pp. 44-60.

Pedone P., Vicario G., Romano D., (2009), Kriging-based sequential inspection plans for coordinate measuring machines, *Applied Stochastic Models in Business and Industry*, Vol. 25, N. 2, pp. 133-149.

Monte Carlo Techniques for Computing Conditional Randomization Tests

William F. Rosenberger
George Mason University
wrosenbe@gmu.edu

Victoria Plamadeala
Precision Therapeutics
vplamadeala@ptilabs.com

Randomization tests are a distribution-free inferential procedure that arise naturally from a randomized clinical trial. They are useful because randomized clinical trials do not follow a population model with random sampling (Rosenberger and Lachin, 2002). It is easy to create a Monte Carlo procedure to compute unconditional randomization tests, where there are 2^n possible randomization sequences, some of which may not be equiprobable (Zhang and Rosenberger, 2012). However, it is a prohibitively large computational problem to compute conditional randomization tests, which condition on the number of observed treatment assignments. In this case, generating a huge number of sequences and then "picking out" those that satisfy the condition results in a test, but the huge number of sequences may be computationally infeasible. We present a method that generates sequences directly from the conditional distribution, by computing certain conditional probabilities relying on combinatorics. We apply this technique to Efron's (1971) biased coin design.

Sequential monitoring in clinical trials is often employed to allow for early stopping and other interim decisions, while maintaining the type I error rate. However, sequential monitoring is typically described only in the context of a population model. We describe a computational method to implement sequential monitoring in a randomization-based context. We also describe the computation of a randomization-based

analog of the information fraction. These techniques require derivation of certain conditional probabilities and conditional covariances of the randomization procedure. We employ combinatoric techniques to derive these for the biased coin design.

The content of this paper appeared recently in *The Annals of Statistics* (Plamadeala and Rosenberger, 2012).

Keywords: biased coin design, conditional reference set, random walk, restricted randomization, sequential analysis.

Acknowledgements: This work was supported by a grant from the National Science Foundation under the 2009 American Recovery and Reinvestment Act.

References

Efron, B. (1971): Forcing a Sequential Experiment to Be Balanced, *Journal of the American Statistical Association*, Vol. 58, pp. 403-417.

Plamadeala V., Rosenberger, W.F. (2012): Sequential Monitoring of Conditional Randomization Tests, *Annals of Statistics*, Vol. 40, pp. 30-44.

Rosenberger, W.F., Lachin, J.M. (2002): *Randomization in Clinical Trials: Theory and Practice*, Wiley, New York.

Zhang, L., Rosenberger, W.F. (2012): Adaptive Randomization in Clinical Trials. In: Hinkelmann, K. (Ed.) *Design and Analysis of Experiments. Volume 3: Special Designs and Applications*, Wiley, Hoboken, pp. 251-282.

Non-symmetrical Passenger Flows Estimation Using the Modified Gravity Model

Diana Santalova
University of Tartu
diana.santalova@ut.ee

In the present research the modified gravity model proposed and estimated in (Andronov, Santalova, 2012) is improved in order to increase the precision of estimation. The detailed analysis of the model estimates properties was performed using principles of simulation modeling in (Santalova, 2013).

The modified gravity model is a nonlinear regression model for passenger correspondences estimation for pairs of spatial points depending on distance between them, population at every such point, and various predictors. The model for a correspondence between points i and l can be written as

$$Y_{i,l} = \frac{(h_i h_l)^\nu}{(d_{i,l})^\tau} exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}), \qquad (1)$$

where a, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)^T$ and $\beta = (\beta_1, \beta_2, ..., \beta_m)^T$ are unknown regression parameters, ν and τ are unknown shape parameters, $c_{(i)} = (c_{i,1}, c_{i,2}, ..., c_{i,m})$ and $g_{(i,l)} = (c_{i,1}c_{l,1}, ..., c_{i,m}c_{l,m})$ are known m-vector-rows, $\{V_{i,l}\}$ are i.i.d. random variables with zero mean and unknown variance σ^2.

Unknown parameters of the model and correspondences are estimated using aggregated data, i.e. total passengers departures from each point for considered time interval. The total number of departures Y_i from point i is presented as sum of correspondences over other points l:

$$Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} Y_{i,l} = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\nu}{(d_{i,l})^\tau} exp(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}). \quad (2)$$

One of assumptions of the model is that estimated correspondences $Y^*_{i,l}$ are symmetric. It means that estimated total number of departed passengers Y^*_i at every point i must be equal to the total number of arrived passengers at the same point. In fact, in the real life this requirement quite often is violated. In this paper we estimate two vectors of parameters on the basis of total numbers of departed (*embarked*) passengers Y^E_i, and of total number of arrived (*disembarked*) passengers Y^D_i. It diminishes the mean squared error of estimation.

Keywords: gravity model, correspondences, non-symmetric flows.

Acknowledgements: The research was supported by the European Union through the European Social Fund (Mobilitas grant No GMTMS280MJ) and Estonian Science Foundation (grant No ETF9127).

References

Andronov A.M., Santalova D. (2012): Statistical estimates for modified gravity model by aggregated data, *Communications in Statistics - Simulation and Computation*, Vol. 41, N. 6, pp. 730-745.

Santalova D. (2013): Experimental Analysis of Statistical Properties of Parameter Estimation of Modified Gravity Model, *Automatic Control and Computer Sciences*, Vol. 47, N. 2, in print.

Bivariate Lorenz Curves based on the Sarmanov-Lee Distribution

José María Sarabia, Vanesa Jordá
Department of Economics, University of Cantabria, Spain
sarabiaj@unican.es, vanesa.jorda@unican.es

The extension of the univariate Lorenz curve to higher dimensions is not an obvious work. Koshevoy and Mosler (1996) introduced the concepts of Lorenz zonoid and Gini zonoid index. However, these concepts are difficult to implement in parametric families of multivariate income distributions. In this paper, we use the alternative definition of bivariate Lorenz curve proposed by Arnold (1987) to generate closed expressions of the Lorenz curve in some relevant bivariate distributions. Let (X_1, X_2) be a bivariate random variable with bivariate probability distribution function F_{12} on R_+^2 having finite second and positive first moments. The Lorenz surface of F_{12} is the graph of the function,

$$L(u, v; F_{12}) = \frac{\int_0^s \int_0^t xy dF_{12}(x, y)}{\int_0^\infty \int_0^\infty xy dF_{12}(x, y)}, \tag{1}$$

where $u = \int_0^s dF_1(x)$, $v = \int_0^t dF_2(y)$, $0 \leq u, v \leq 1$ and F_1 and F_2 are the marginal distribution functions of F_{12}. Note that if F_{12} is a product distribution function, then $L(u, v; F_{12})$ is just the product of the marginal Lorenz curves. Now, let (X_1, X_2) be the class of bivariate distributions with given marginals described by Sarmanov and Lee (Lee, 1996; Sarmanov, 1966). In this paper, using definition (1), a closed expression for the bivariate Lorenz curve in the case of the Sarmanov-Lee distribution is given. The expression obtained can be easily interpreted as a convex linear combination of products of classical and generalized Lorenz curves. Several special cases and other alternative families of bivariate distributions with given marginals are studied, including the classical Farlie-Gumbel-Morgenstern (FGM) distribution, iterations of the FGM, and the variations proposed by Huang and Kotz and Bairamov and Kotz. Models of bivariate distributions based on conditional specification (Arnold, Castillo, Sarabia, 1999) are also considered. A closed

expression for the bivariate Gini index (Arnold, 1987) in terms of the classical and generalized Gini indices of the marginal distributions is given. Specific models with Beta, GB1, Pareto and lognormal marginal distributions are studied in detail. Some concepts of stochastic dominance are explored. Extensions to higher dimensions are included. Finally, some numerical illustrations with real income data are given.
Keywords: Sarmanov-Lee distribution, Lorenz curve, bivariate Gini index, stochastic dominance.

Acknowledgements: The authors thank to Ministerio de Economía y Competitividad (project ECO2010-15455) and Ministerio de Educación (FPU AP-2010-4907) for partial support.

References

Arnold, B.C. (1987): *Majorization and the Lorenz Curve*, Lecture Notes in Statistics 43, Springer Verlag, New York.

Arnold, B.C., Castillo, E., Sarabia, J.M. (1999): *Conditional Specification of Statistical Models*, Springer, New York.

Koshevoy, G., Mosler, K. (1996): The Lorenz Zonoid of a Multivariate Distribution, *Journal of the American Statistical Association*, Vol. 91, pp. 873-882.

Lee, M-L.T. (1977): Properties of the Sarmanov Family of Bivariate Distributions, *Communications in Statistics, Theory and Methods*, Vol. 25, N. 6, pp. 1207-1222.

Sarmanov, O.V. (1966): Generalized Normal Correlation and Two-Dimensional Frechet Classes, *Doklady (Soviet Mathematics)*, Vol. 168, pp. 596-599.

Additive Model for Cost Modelling in Clinical Trial

Nicolas Savy and Guillaume Mijoule
Institut Mathématiques de Toulouse
nicolas.savy@math.univ-toulouse.fr
guillaume.mijoule@math.univ-toulouse.fr

Vladimir Anisimov
Predictive Analytics, Innovation, Quintiles, UK
Vladimir.Anisimov@quintiles.com

In the framework of a clinical trial or more generally in medical research, an important question is how long it takes to recruit a given number of patients N_R. Indeed, this is of a paramount interest for planning trials because of scientific concern, economic and ethical reasons. Ethical, because it is not satisfactory to continue a study in vain and economical, because an improvement of the planning and monitoring of a trial allows reducing costs and save money. Scientific concern, because new drugs are increasingly developed and approved by regulatory agencies and when accrual rates are too low, there may be new information available during the enrollment period such as the results of other trials or a change in the understanding of the underlying biology. Meanwhile, a huge variability of the recruitment process makes the question quite hard to investigate, thus, stochastic modelling has to be developed. There were many investigations on this way and now we are able to claim that the easier to handle and more relevant model is a so-called Poisson-Gamma model (Mijoule *et al* (2012)) introduced in Anisimov *et al* (2007). This model assumes that patients arrive at different centres according to Poisson processes where the rates are $\Gamma(\alpha, \beta)$-distributed.

The talk aims to give the very first steps of a model for multicentric clinical trial cost. The main ingredients are the recruitment process N_t^C which is a sum of C Poisson-Gamma processes modelling recruitment in each centre, two constants, F_c, that is a fixed cost of a centre c and K, that is a fixed cost per patient, and a function $k(\cdot)$ representing a cost per patient as a function of t. The total cost at time t can be written (T_i

is the arrival instant of i-th patient):

$$C_t = KN_t^C + \sum_{i=1}^{N_t^C} k(t, T_i) + \sum_{c=1}^{C} F_c.$$

The first term of the right-hand term is the cost generated by the N_t recruited patients, the second one, the cumulative cost of the different patients and F is the fixed cost of the centre. This can be rewritten

$$C_t = KN_t^C + \int_0^t k(t, s) dN_s^C + \sum_{c=1}^{C} F_c. \tag{1}$$

It is easy to calculate the expectation $\mathbb{E}[C_t]$ for a given t but the instant of interest is the first time denoted by τ when the process N_t^C attains N_R ($\tau = \inf\{t : N_t \geq N_R\}$). Obviously τ is a stopping time for the natural filtration but unfortunately, the integral involved in (1) is not of Itô's type and the calculation of $\mathbb{E}[C_\tau]$ cannot be made by the use of standard stochastic calculus arguments.

The main objective is to present the main ingredients of the cost process modelling and especially the recruitment process. We also explain how to calculate the total cost ($\mathbb{E}[C_\tau]$) at the end of clinical trial.

Keywords: Clinical trials, recruitment, Bayesian statistics, Poisson process.

Acknowledgements: This research has benefited from the help of IRESP during the call for proposals launched in 2012 in the setting of Cancer Plan 2009-2013.

References

Mijoule G., Savy S., Savy N. (2012): Models for patients' recruitment in clinical trials and sensitivity analysis, *Statistics in Medicine*, Vol. 31, N. 16, pp. 1655-1674.

Anisimov V., Fedorov V. (2007): Modelling, prediction and adaptive adjustment of recruitment in multicentre trials, *Statistics in Medicine*, Vol. 26, N. 27, pp. 4958-4975.

Simulating from the copula that generates the maximal probability for a joint default under given (inhomogeneous) marginals

Matthias Scherer
Technische Universität München
scherer@tum.de

Jan-Frederik Mai
XAIA Investment GmbH
jan-frederik.mai@xaia.com

Given univariate survival functions, we compute the dependence structure that maximizes the probability of a joint default. For inhomogeneous marginals, this is not the comonotonicity copula, opposed to a common modeling (mal-)practice in the financial industry. We present a stochastic model that respects the marginal laws and attains the upper bound for joint defaults. We explain how one can simulate from this copula / joint distribution. We illustrate the theoretical findings by bootstraping default probabilities from credit default swap contracts referencing on EU peripherals and Germany and we compute the upper bound for the probability of Germany defaulting jointly with one of the peripherals.

Keywords: Joint default, singular component, copula.

References

Mai, J.-F.; Scherer, M., The maximal probability for a joint default under given marginals, working paper, 2013.

The analysis of time course ranking data by nonparametric inference

Michael G. Schimek
Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, 8036 Graz, Austria
michael.schimek@medunigraz.at

Marcus D. Bloice, Vendula Švendová
Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, 8036 Graz, Austria
marcus.bloice@medunigraz.at, vendula.svendova@medunigraz.at

Recently various approaches for the statistical analysis of several ranked lists that comprise the same set of objects have been proposed. The general assumption is that assessors or assessing devices rank these objects independently of each other. This assumption holds for many applications (e.g. Schimek and Bloice, 2012) such as the integration of rank order data from different biotech laboratories or from Web search engines. In this paper we are taking an interest in ranked lists which originate from one and the same ranking mechanism but with repetitions in time order. A typical example are university league tables that are calculated each year. An institution providing such tables adheres to the same ranking criteria for sake of comparability along the time line. So far statistical analysis of such time course ranking data could only be performed under the insufficient assumption of independence. Most recently, Hall and Schimek (2012) have developed a nonparametric inference procedure which allows us, not only to estimate the length of a top-k sublist of high concordance under independence of the rankings, but also under moderate m-dependence of the rankings. In both instances their approach provides an estimate of the point of degeneration j_0, where $k = j_0 - 1$, for pairwise combinations of the input lists, even when they are highly irregular and very long.

The estimation of j_0 is achieved via a *moderate deviation*-based ap-

proach. In the theoretical analysis of the probability that an estimator (computed from a pilot sample size ν) exceeds a value z, the deviation above z is said to be a moderate deviation if its associated probability is polynomially small as a function of ν, and to be a large deviation if the probability is exponentially small in ν. In regular cases, the values of $z = z_\nu$ that are associated with moderate deviations are $z_\nu \equiv \bigl(C\,\nu^{-1}\,\log\nu\bigr)^{1/2}$, where $C > \frac{1}{4}$. The null hypothesis H_0 that $p_k = \frac{1}{2}$ for ν consecutive values of k, versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k, is rejected if and only if $\hat{p}_j^{\pm} - \frac{1}{2} > z_\nu$. The quantities $\hat{p}_j^{+}$ and $\hat{p}_j^{-}$ represent estimates of p_j computed from the ν data pairs I_i for which i lies to the right of j, or to the left of j, respectively. Under H_0, the variance of $\hat{p}_j^{\pm}$ equals $(4\nu)^{-1}$, hence we can evaluate the above inference procedure in practice.

Our understanding of time course ranking data is still quite limited. Therefore, we have developed a suitable data generating model that allows us to perform conclusive simulation experiments. As well as the simulation evidence we also present an example based on UK university league tables from 2008 to 2013 (source: The Complete University Guide).

Keywords: Consensus ranking, moderate deviations, m-dependence, ranked list, time course data.

References

Hall P., Schimek M.G. (2012): Moderate deviation-based inference for random degeneration in paired rank lists. *J. Amer. Statist. Assoc.*, 107, 498, 661-672.

Schimek M.G., Bloice M. (2012): Modelling the rank order of Web search engine results. In: Komarek A., Nagy S. (Eds). *Proceedings of the 27th International Workshop on Statistical Modelling.* Vol. 1, 303-308.

Exact One-Sided Tests for Semiparametric Binary Choice Models

Karl H. Schlag
Department of Economics, University of Vienna
karl.schlag@univie.ac.at

Francesca Solmi
I-BIOSTAT, University of Hasselt
francesca.solmi@uhasselt.be

A test is exact if its significance level can be proven for the given finite sample size. We present the first exact one-sided tests for binary choice models. The tests apply when there are sufficiently many pairs of observations that have the same attributes apart from the one to be investigated. Note that exact tests have not even been previously available for the parametric logit and probit models.

We consider a binary choice model in which m real valued attributes $z_1, .., z_m$ are used to explain the binary response $Y \in \{0, 1\}$. Unknown parameters $\beta_1, .., \beta_m$ determine how changes in the attributes influence the probability that $Y = 1$. Specifically,

$$P(Y = 1|z) = F(z, \beta)$$

where we assume that F is strictly increasing in z_j if and only if $\beta_j > 0$. A popular example is logistic choice where $F(z, \beta) = \frac{e^{z'\beta}}{1+e^{z'\beta}}$. We wish to test $H_0 : \beta_j \leq 0$ against $H_1 : \beta_j > 0$ based on n observations $(x_i, y_i) \in \mathbb{R}^{m+1}$, $i = 1, .., n$, where y_i is a realization of Y_i such that $P(Y_i = 1|x_i) = F(x_i, \beta)$ and $(Y_i)_{i=1}^n$ are independent random variables if one conditions on $x_1, .., x_n$.

Our tests do not require that the statistician knows F. However, they are only applicable if there are sufficiently many pairs of observations in which all attributes except for the jth one are identical. Formally a test is a mapping $\phi : \mathbb{R}^{n\times m} \times \mathbb{R}^n \to \{0, 1\}$ such that the null hypothesis

is rejected if $\phi(x, y) = 1$. The test has *level* α if $\beta_j \leq 0$ implies $P(\phi = 1|x) \leq \alpha$. The test is also called *exact* as the inequality defining the level can be proven for the given sample of attributes and is not based on asymptotic theory.

Note that there are no exact one-sided tests for the popular cases of logit and probit. Exact two-sided tests for $H_0 : \beta_j = 0$ can easily be constructed using permutation tests, however a rejection does not give any indication as to whether the jth attribute has a positive or a negative effect on the probability. The only known exact one-sided test is available for the linear probability model where $F(z, \beta) = z'\beta$, and is due to Gossner and Schlag (2012).

Our test is constructed as follows. We first collect observations into blocks where observations belonging to the same block have the same values of attributes $k \neq j$. We then pair observations within each block and investigate across all blocks, using the binomial test, how the response depends on changes in the value of the jth attribute. Additional randomization and derandomization techniques are used to make the test independent of how observations are indexed.

Numerical examples are used to demonstrate the power of our test and to show how standard tests do not control the type I error.

Keywords: binary choice, single index model, exact testing, logit, probit.

References

Gossner O., Schlag K.H. (2012): *Finite Sample Nonparametric Tests for Linear Regressions*, SSRN working paper, http://ssrn.com /abstract=1569483.

Schlag K.H. (2008): *A New Method for Constructing Exact Tests Without Making any Assumptions*, working paper 1109, Universitat Pompeu Fabra, Barcelona.

Exact P-value Computation for Correlated Categorical Data

Pralay Senchaudhuri
Cytel Software Corporation
pralay@cytel.com

Chris Corcoran, V.P. Chandran
Utah State University, Cytel Software Corporation
chris.corcoran@usu.edu, chandran@cytel.com

Analysis of correlated data, for categorical as well as continuous endpoints, is very common in statistical analysis. When data arise from cluster-correlated binomial populations, the dependence among observations within clusters can lead to what is known as extra-binomial variation or overdispersion. Methodologies based on random effect and marginal models, such as GEE and GMM, are widely available to analyze such data. These methods generally assume large-sample continuous distributions to approximate the distributions of test statistics used to evaluate model parameters. For categorical data, such approaches may work poorly when sample sizes are small or sparse. Permutation or exact methods are available in such cases (conditioning on the sufficient statistics from the exponential family formulation for correlated categorical data developed by Molenberghs and Ryan, 1999) but the algorithms for computing exact p-values are often memory and time intensive. We have developed new computational approaches using multicore processors to cut the execution time and to more efficiently manage memory. In this paper we discuss how these methods can be applied through Monte Carlo simulation, using a graphical network-based algorithm.

We first we create a set of loosely connected networks (See Mehta and Patel, 1983) to represent a set of all possible contingency tables that have the same correlation structure as the observed data. Instead of using one monolithic network as previously suggested (see Corcoran et

al., 2001), we create multiple networks, each representing a value of the sufficient statistic related to the observed correlation structure. We then develop a network-based sampling method to sample these contingency tables and to compute the exact p-value. To increase efficiency we use parallel processing during the network-building phase, as well as during the Monte Carlo sampling phase. We accomplish this by using multicore processors, and in this paper we show how we can employ MPI and grid structure to distribute the computation.

Keywords: exact test, correlated data, Monte Carlo, network algorithm, parallel processing

Acknowledgements: This work was supported by NIH grant R44 RR019052.

References

Corcoran C., Ryan L., Mehta C.R., Senchaudhuri P., Patel N., and Molenberghs G. (2001): An Exact Trend Test for Correlated Binary Data, *Biometrics*, Vol. 57, pp. 941-948.

Mehta C.R., and Patel N.R. (1983): A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *JASA*, Vol. 78, N. 382, pp. 427-434.

Molenberghs G. and Ryan L.M. (1999): An exponential family model for clustered multivariate binary data, *Environmetrics*, Vol. 10, N. 3, pp. 279-300

An efficient method for pseudo-random UDG graph generating

Vladimir Shakhov, Olga Sokolova, Anastasia Yurgenson
Institute of Computational Mathematics and Mathematical Geophysics
SB RAS, Russia
shakhov@rav.sscc.ru, olga@rav.sscc.ru, nastya@rav.sscc.ru

Simulation technique has always been the primary methodological framework for research and development of communication networks. For this purpose, the analytical approach based on queue theory or graph theory have been often used as well. However, its use requires quit simplifying assumptions since more realistic assumptions make comprehensive analysis extremely difficult. The main disadvantages of empirical methods include inability to test network sensitivity and tune performance. Thus, simulation technique is an indispensable tool for the research and development of wireless sensor networks. It allows to get significant comparative studies of different traffic-aggregation protocols or energy efficient routing algorithms, and hence, a researcher can determinate which protocol performs better from the quality-of-service point of view in a concrete situation.

One of the most important issue of a wireless sensor network simulation is the pseudo-random graphs generator required to describe the structure of links connecting pairs of sensors. A network topology strongly influence the choice and performance of routing and broadcasting protocols. For these reasons, it is very important to consider the nature and mechanism of network topology generating, and the manner in which it depends on the features of the wireless sensor network. In most investigations of wireless sensor networks, when topology-based algorithms are evaluated, a unit disk graph (UDG graph) is generally used in assumptions of the simulation framework. Let us provide the formal definition of UDG graph from the textbook (Wagner and Wattenhofer, 2007).

Definition. A graph $G = (V, E)$ is a UDG graph if and only if there is an embedding of the nodes in the plane such that $\{u, v\} \in E$, $u, v \in V$ if and only if the Euclidean distance between u and v is less than or equal to 1.

Taking into account characteristic of real networks, a reliable generator has to deliver UDG graphs with some additional properties, for example, graph connectivity, the limitations for graph diameter, desired nodes density etc. The common used approach of UDG graphs generating is as follows. A graph is formed by distributing nodes uniformly, randomly and independently from each other in some area (Choo, 2010). Next, for each pair of nodes the edge presence is computed. If the generated graph is not connected then one is rejected. Depending on desired properties of UDG graph, some additional actions can be made.

We provide a new method that can speedup simulation of wireless sensor networks or ad hoc networks. Performance improvement is achieved by generating pseudo-random UDG graphs with prescribed properties. Thus, additional treatment of the generated graphs is not required and the graph rejection procedure is not used. Numerical results demonstrate that the proposed method achieves an essential computational cost reduction in comparison with the standard approach.

Keywords: simulation, pseudo-random graph generator, UDG graphs

References

Wagner D., and Wattenhofer R. (2007): *Algorithms for Sensor and Ad Hoc Networks*, LNCS, vol. 4621. Springer, Heidelberg

Jaewan Seo, Moonseong Kim, In Hur, Wook Choi and Hyunseung Choo (2010): *DRDT: Distributed and Reliable Data Transmission with Cooperative Nodes for LossyWireless Sensor Networks*, Sensors 2010, 10(4), pp.2793-2811

Functional central limit theorem for integrals over level sets of Gaussian random fields

Alexey Shashkin
Moscow State University
ashashkin@hotmail.com

Level sets of Gaussian random fields have attracted much interest due to their applications to modeling complex stochastic structures. Starting from the classical Rice formula, a group of results concerning the behavior of the Hausdorff measure of these level sets has been established, see Shashkin (2013) and references there. There are also some limit theorems for the integrals over the level sets, provided that the underlying random field satisfies a mixing property, see Iribarren (1989). Our talk is devoted to the functional limit theorem for the integrals over Gaussian level sets where the integrand is a continuous random field independent of the underlying Gaussian one. Instead of mixing we employ the notion of (BL, θ)-dependence, which comprises both Gaussian and associated random fields with integrable covariance functions. For $\Delta > 0$ set $T(\Delta) = \{(k_1/\Delta, \dots, k_d/\Delta) \in \mathbb{R}^d : k \in \mathbb{Z}^d\}$.

Definition (Bulinski (2010)). A random field $\xi = \{\xi(t), t \in \mathbb{R}^d\}$ is called (BL, θ)-dependent if there exists a nonincreasing function $\theta_\xi : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\theta_\xi(r) \to 0$ as $r \to \infty$ and for all $\Delta > 0$ large enough, any pair of finite disjoint $I, J \subset T(\Delta)$ and any Lipschitz functions $f : \mathbb{R}^{|I|} \to \mathbb{R}$, $g : \mathbb{R}^{|J|} \to \mathbb{R}$ one has

$$|cov(f(\xi_I), g(\xi_J))| \le Lip(f)Lip(g)(|I| \wedge |J|)\Delta^d \theta_\xi(r). \qquad (1)$$

Here r is the distance between I and J, $|M|$ is the cardinality of a finite set M, the abbreviation $\xi_M = (\xi_i, i \in M)$ is used for a finite $M \subset \mathbb{R}^d$, and the Lipschitz constants are taken with respect to the l_1-norm.

Let $d \ge 3$ and $X = \{X(s), s \in \mathbb{R}^d\}$ be an isotropic Gaussian random field with realizations that are C^1 almost surely. For a bounded Borel $A \subset \mathbb{R}^d$ and $u \in \mathbb{R}$, let $B(u, A) = \{s \in A : X(s) = u\}$ and

let $H(u, A)$ denote the $(d-1)$-dimensional Hausdorff measure of this set. Assume that $Y = \{Y(s), s \in \mathbb{R}^d\}$ is a continuous random field independent from X. With probability one, we can define a normalized integral $Z_n(u) := n^{-d/2} \int_{[0,n]^d} Y(s) H(u, ds)$, $n \in \mathbb{N}$.

Theorem. Random processes $\{Z_n(\cdot), n \in \mathbb{N}\}$ have continuous trajectories almost surely. Moreover, suppose that the covariance function of X is integrable together with its derivatives of order one and two. Let Y also be centered, (BL, θ)-dependent and strictly stationary with integrable covariance function. Then the sequence $\{Z_n, n \in \mathbb{N}\}$ converges in distribution in $C(\mathbb{R})$, as $n \to \infty$, to a centered Gaussian random process with covariance function determined by X and Y.

Keywords: Gaussian random fields, level sets, Rice formula, functional central limit theorems, weak dependence.

Acknowledgements: This work was supported by RFBR, project 13-01-00612.

References

Bulinski A. (2010): Central limit theorem for random fields and applications. In: Skiadas, C.H. (Ed.) *Advances in Data Analysis*, Birkhäuser, Boston, pp. 141–150.

Iribarren I. (1989): Asymptotic behaviour of the integral of a function on the level set of a mixing random field, *Probability and Mathematical Statistics*, Vol. 10, No. 1, pp. 45–56.

Shashkin A. (2013): Functional central limit theorem for the level measure of a Gaussian random field, *Statistics and Probability Letters*, Vol. 83, No. 2, pp. 637–643.

Monte Carlo method for partial differential equations

Sipin Alexander
Vologda State Pedagogical University
cac1909@mail.ru

The unbiased estimators for solutions $u(x)$ of boundary value problems for PDE's are usually constructed on trajectories of Markov processes in domain $\mathcal{D}$ in R^n or on the boundary $\partial\mathcal{D}$. The transition funtion $P(x, dy)$ of such processes $\{x_i\}_{i=0}^{\infty}$ is usually the kernel of integral equation

$$u(x) = \int_Q u(y)P(x, dy) + F(x), x \in Q. \tag{1}$$

Here, $Q = \overline{\mathcal{D}}$ or $Q = \partial\mathcal{D}$, the function $F(x)$ is defined by boundary conditions and right part of differential equation.

Let K – integral operator in the equation (1). If $\|K\| < 1$, we may use von-Neumann-Ulam scheme to construct the unbiased estimators for $u(x)$. In case of $\|K\| = 1$ and $F(x) \geq 0$ for any bounded solution $u(x)$ of equation (1) we have representation

$$u(x) = \sum_{i=0}^{\infty} K^i F(x) + K^{\infty} u(x), x \in Q. \tag{2}$$

Now, the Markov chain $\{x_i\}_{i=0}^{\infty}$ must have additional properties :

1). $P_x(\tau = \infty) > 0$,

2). $P_x(x_i \to x_\infty, x_\infty \in \partial\mathcal{D} | \tau = \infty) = 1$ or

3). $P_x(\rho(x_i, \partial\mathcal{D}) \to 0 | \tau = \infty) = 1$,

4). $E_x min(\tau, \tau_\varepsilon) < \infty$ for $\tau_\varepsilon = \inf(i : \rho(x_i, \partial\mathcal{D}) < \varepsilon)$.

The properties permit us to obtain unbiased and ε-biased estimators for $u(x)$.

Using invariant and excessive functions for $P(x, dy)$, we can construct simple conditions, which yield properties 1) – 4). Markov chains, usually used in Monte Carlo algorithms for boundary value problems (Ermakov S.M., 1989), satisfy these conditions. These results are applied to the "walk in hemispheres" (Ermakov S.M.,2009) and the "walk on cylinders" (Sipin A.S., 2012) processes.

Keywords: boundary value problems, von-Neumann-Ulam scheme, Markov process, Monte Carlo method.

Acknowledgements: This work was supported by RFBR (11-01-00769)

References

Ermakov S.M., Nekrutkin V.V., Sipin A.S. (1989): *Random processes for classical equations of mathematical Physics*, Kluwer Academic Publishers, Dordrecht, Boston, London.

Ermakov S.M., Sipin A.S. (2009): The "walk in hemispheres" process and its applications to solving boundary value problems, *Vestnik St.Petersburg University: Mathematics*, Vol. 42, N. 3, pp. 155-163.

Sipin A.S., Bogdanov I.I. (2012): "Random Walk on Cylinders" for Heat Equation. In: Tchirkov M.K. (Eds.) *Mathematical models. Theory and applications.*, VVM, Sc.Res.Inst.Chem., St.Petersburg University, Issue. 13, pp. 25-36. (in Russion)

A change detection in high dimensions using random projection – simulation study

Ewa Skubalska-Rafajłowicz
Institute of Computer Engineering, Automation and Robotics, Technical University of Wrocław, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
ewa.rafajlowicz@pwr.wroc.pl

The method of change detection in high-dimensional time sequences is presented. We use normal random projections and random projection with sub-gaussian tails as a method of dimensionality reduction. Diagnostic properties of the Hotelling control chart applied to data projected onto random subspace of $\mathcal{R}^n$ are examined. We assume that the observations, d-dimensional vectors X, are independent random samples drawn from an unknown distribution normal distribution, let say $N_d(\mu_0, \Sigma)$. If the distribution mean changes from $\mu_0 = 0$ to μ_1 we should be able to detect this event with possibly highest probability.

The goal of this simulation study is to show that random projections (see Achlioptas (2003), (Dasgupta, 2003), (Matousek, 200), (Vempala, 2004), among many others) can be efficiently used for mean change detection during statistical monitoring of a high-dimensional normal process.

The common statistics used in monitoring individual multivariate observations is the Hotelling statistic Hotelling (1931), Rao (1973), Mason and Young (2002). The bases of the T^2 statistic is knowledge of the covariance matrix Σ or its good estimate. Here we assume that the variance-covariance matrix Σ is not known and due to a large dimensionality d its estimation is impossible.

A random projection from d dimensions to k dimensions is a linear transformation represented by a $d \times k$ matrix $S \in \mathcal{R}^{k\times\, d}$ – a matrix whose entries are i.i.d. samples of a certain random variable. But once these values (i.e., projection matrix entries) are obtained, it remains

fixed for the rest of simulation.

Thus, projected observations $V = SX$ are k-dimensional random vectors which follow normal distribution $N_k(0,\ S\,\Sigma\,S^T)$. As a next step we calculate T^2 statistics from projected observations (in k dimensions). Without loss of generality we consider the case when the covariance matrix Σ is a diagonal matrix $diag(l_1, \ldots, l_d)$, where $l_1 \geq \ldots \geq l_d$.

We discuss some properties of this test, while the goal of our simulations studies is to evaluate its power. In this simulation studies $d = 10000$ and k varies from 10 to 100. *Keywords:* change detection, multidimensional control charts, dimensionality reduction, random projections, Hotelling statistics

References

Achlioptas D. (2003): Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *Journal of Computer and System Sciences*, Vol. 66, N. 4, pp. 671–687.

Dasgupta, S., Gupta, A. (2003): An elementary proof of a theorem of Johnson and Lindenstrauss, *Random Structures and Algorithms,* Vol. 22, N. 1, pp. 60–65.

Mason, R.L. , Young, J.C. (2002) *Multivariate Statistical Process Control with Industrial Application*, ASA-SIAM Series on Statistics and Applied Mathematics, Philadelphia.

Matoušek, J. (2008): On Variants of the Johnson-Lindenstrauss Lemma, *Random Structures and Algorithms*, Vol. 33, N. 2, pp. 142–156.

Vempala, S.(2004): *The Random Projection Method*, American Mathematical Society, Providence, RI.

Rao, C. R. (1973): *Linear Statistical Inference and Its Applications*, John Wiley and Sons New York, London, Sydney, Toronto.

The probabilistic approximation of the one-dimensional initial boundary value problem solution

Smorodina N.V.
St.Petersburg State University, Department of Physics. Russia
nsmorodna@yahoo.com

We consider the equation

$$\frac{\partial u}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 u}{\partial x^2} + f(x)u,$$

where σ is a complex-valued parameter, such that $\mathrm{Re}\sigma^2 \geq 0$. When σ is a real number this equation corresponds to the Heat equation while when $\mathrm{Re}\sigma^2 = 0$ it corresponds to the Schrödinger equation. For this equation we consider the initial boundary value problem with Dirichlet condition $u(0,x) = \varphi(x),\ u(a,t) = 0,\ u(t,b) = 0$. In the case when σ is a real number there exists probabilistic representation of the solution (Freidlin (1985)) in a form of the mathematical expectation (so called Feynman - Kac formula), namely

$$u(t,x) = \mathrm{E}\left\{\varphi(\widetilde{\xi}_x(t\wedge\tau))e^{\int_0^{t\wedge\tau} f(\widetilde{\xi}_x(v))dv}\right\},$$

where $\widetilde{\xi}_x(t)$ is a Brownian motion with a parameter σ, killed at the exit time τ from the interval $[a,b]$. On the base of this representation one can approximate the solution using some suitable approximation of the Wiener process. This approach doesn't work if $\mathrm{Im}\sigma \neq 0$. It is known that when σ is not a real number there exists no analogue of the Wiener measure and hence one can not present the Feynman -Kac formula as an integral with respect to a σ-additive measure in a trajectory space. When $\mathrm{Re}\sigma^2 = 0$ (that corresponds to the Schrödinger equation) one can apply an integral with respect to the so called Feynman measure

that is a finitely-additive complex measure in the trajectory space which is defined as a limit over a sequence of partitions of an interval $[0, T]$. It should be mentioned that this approach is not a probabilistic approach in the usual sense since the very notion of a probability space does not appear in it. To get the stochastic approximation of the solution we use another approach based on the theory of generalized function. On a special probability space we define the sequence of probability measures $\{\mathrm{P}_n\}$ and limit object $\mathrm{L} = \lim_{n\to\infty} \mathrm{P}_n$ but this limit object is not a measure it is only generalized function. That means that the convergence $\int f d\mathrm{P}_n \to (\mathrm{L}, f)$ is valid only if f belongs to the class of test functions. On this probability space we define a complex-valued process so that the mathematical expectation with respect to the measure P_n of some functional of this process converges to the value of the generalized function applied to this test function that leads to a solution of the initial boundary value problem.

Keywords: random process, evolution equation, limit theorem, probabilistic approximation.

Acknowledgements: This work was supported by Russian Foundation for Basic Research 12-01-00487a.

References

M.Freidlin. (1985): *Functional integration and partial differential equations.*, Princeton University Press, Princeton, New Jersey.

Kai Lai Chung, Zhongxin Zhao. (1995): *From Brownian motion to Schrodinger's equation.*, Springer-Verlag Berlin Heidelberg.

The calculation of effective electro-physical parameters for a multiscale isotropic medium

Soboleva O. N.
Computational Mathematics and Mathematical Geophysics SB RAS
olgasob@gmail.com

Kurochkina E.P.
Institute of of Thermophysics SB RAS
Kurochkina@itp.nsc.ru

Wave propagation in complex inhomogeneous media is an urgent problem in many fields of research. In electromagnetics, these problems arise in such applications as estimation of soil water content, well logging methods, etc. In order to compute the electromagnetic fields in an arbitrary medium, one must numerically solve Maxwell's equations. The large scale variations of coefficients as compared with wavelength are taken into account in these models with the help of some boundary conditions. The numerical solution of the problem with variations of parameters on all the scales require high computational costs. The small scale heterogeneities are taken into account by the effective parameters. In this case, equations are found on the scales that can be numerically resolved. It has been experimentally shown that the irregularity of electric conductivity, permeability, porosity, density abruptly increases as the scale of measurement decreases. The spatial positions of the small-scale heterogeneities are very seldom exactly known. It is customary to assume the parameters with the small scale variations to be random fields characterized by the joint probability distribution functions. In this case, the solution of the effective equations must be close to the ensemble-averaged solution of the initial problem. For such problems, a well-known procedure of subgrid modeling is often used. The effective coefficients using subgrid modeling in the quasi-steady Maxwell's equations for a multiscale isotropic medium are described in (Soboleva

2011). In the present paper we obtain formulas of effective coefficients for Maxwell's equations in the frequency domain when the following condition $\sigma(\mathbf{x})/(\omega\varepsilon(\mathbf{x})) < 1$ is satisfied. The correlated fields of electric conductivity and permittivity are approximated by a multiplicative continuous cascade:

$$\varepsilon(\mathbf{x})_{l_0} = \varepsilon_0 \exp\left(-\int_{l_0}^{L} \chi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right), \tag{1}$$

$$\sigma(\mathbf{x})_{l_0} = \sigma_0 \exp\left(-\int_{l_0}^{L} \varphi(\mathbf{x}, l_1)\frac{dl_1}{l_1}\right), \tag{2}$$

where ε_0, σ_0 are the constants, l_0, L are the minimum-scale and maximum-scale of measuring, respectively. It is assumed that the fields $\chi(\mathbf{x}, l)$, $\varphi(\mathbf{x}, l)$ are isotropic with a normal distribution and a statistically homogeneous correlation function. The fluctuations of these fields on different scales of heterogeneities do not correlate. This assumption is standard in the scaling models. To derive subgrid formulas to calculate effective coefficients, this assumption may be ignored. However, this assumption is important for the numerical simulation of the field. The theoretical results obtained in the paper are compared with the results from direct 3D numerical simulation.

Keywords: Maxwell's equations, multiscale isotropic medium, subgrid modeling, effective parameters.

Acknowledgements: This work was supported by the Russian Foundation for Basic Research, Grant No. 11-01-00641a

References

Kurochkina E.P., Soboleva O. N. (2011): Effective coefficients of quasi-steady Maxwell's equations with multiscale isotropic random conductivity, *Physica A.*, Vol. 390, N 2, pp. 231-244.

Simulation-Based Optimal Design Using MCMC

Antti Solonen
Lappeenranta University of Technology, Finland
solonen@lut.fi

In classical parameter estimation of nonlinear models, a Gaussian approximation of parameter uncertainty is usually obtained by linearizing the model around a point estimate. Classical optimal design methods are based on this approximation, and aim, for instance, at minimizing the covariance matrix of the parameters. These approaches suffer from two flaws. First, they depend on the linearization point and on the validity of the Gaussian approximation. Second, classical design methods are often unavailable in ill-posed estimation situations, where previous data lacks the information needed to properly construct the design criteria. With MCMC methods, nonlinear parameter estimation problems can be solved without using, e.g., Gaussian approximations, and MCMC has been intensively used for model fitting in many fields of science and engineering. Therefore, it is natural to study ways how the MCMC output samples from the parameter posterior can be used further in optimal design. A framework for this is given by the simulation-based optimal design concept introduced by Mller et al. (2004). In this talk, we discuss several aspects of simulation based optimal design based on MCMC parameter estimation. We illustrate the benefits of the approach by numerical examples arising from parameter estimation problems in mechanistic dynamical ODE models.

Keywords: Simulation-based optimal design, Markov chain Monte Carlo, dynamical ODE models.

References

Müller P., Sansó B., De Iorio M. (2004): Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation, *Journal of the American Statistical Association*, Vol. 99, N. 467, pp. 788-798.

Nonparametric change detection based on vertical weighting

Steland A.
Institute of Statistics, RWTH Aachen University, Germany
steland@stochastik.rwth-aachen.de

Pawlak M.
The University of Manitoba, Winnipeg, Canada
Miroslaw.Pawlak@ad.umanitoba.ca,

Rafajłowicz E.
Wrocław University of Technology, Wrocław, Poland,
ewaryst.rafajlowicz@pwr.wroc.pl

The idea of weighting observation not only along the horizontal axes (as it is done in a regression kernel smoothers), but also along the vertical axes can be formally derived (see Rafajłowicz (1996), Pawlak and Rafajłowicz (2000)). When a horizontal weight is uniform on an interval of length $h > 0$ and a vertical weight is uniform on $[-H,\ H]$, $H > 0$ say, then vertical weighting reduces to clipping observations that are outside an interval $[\hat{y}-H,\ \hat{y}+H]$, where $\hat{y}$ is selected by the statistician. Selecting $\hat{y}$ as the last available observation (see Fig.2), one can derive a number of jump detectors: by appropriately processing observations that are captured by a box of the length h and the height $2\,H$ moving in time. In particular, one average them (as it is done in the above cited paper) or calculate their median (as proposed by Pawlak et all (2004)) or just count their number and compare it with the expected number of observations, assuming the a jump is not present (as discussed recently by Rafajłowicz et all (2010)). In all the above mentioned cases we obtain detectors that are nonparametric in the sense that we do not impose restrictions on probability distributions of random errors in observations. We are interested in the quickest detection properties of these and related jump detectors (see Poor (2009) and the bibliography cited therein). In

particular, our aim is to provide sufficient conditions for the detection with conditionally zero delay. By the conditionally zero delay property we mean a detection that takes place at sample point that is closest to the jump occurrence, provided that there was not jump at h time instants before. We provide also the results of simulations.

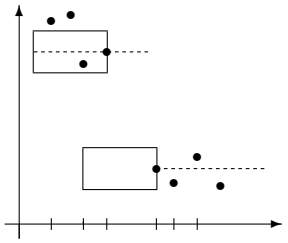

Figure 2. The idea of vertical clipping in jump detection.

Keywords: change detection, nonparametric setting, vertical weighting

References

Rafajłowicz E., Pawlak M., Steland A. (2010) Nonparametric Sequential Change-Point Detection by a Vertically Trimmed Box Method, IEEE Trans. Information Theory, Vol. 56, pp. 3621-3634.

Pawlak M., E. Rafajłowicz (2000) Vertically weighted regression. A tool for non-linear data analysis and constructing control charts, J. German Statist. Assoc., vol. 84, pp. 367-388.

Pawlak M., Rafajłowicz E., and A. Steland (2004) On detecting jumps in time series – Nonparametric setting, J. Nonparametric Statist., vol. 16, pp. 329-347.

Poor H. V. and O. Hadjiliadis (2009) Quickest Detection. Cambridge, U.K., Cambridge Univ. Press

E. Rafajłowicz (1996) Model Free Control Charts for Continuous Time Production Processes Wrocław University of Technology Tech. Rep. No. 56.

Some Thoughts About L-Designs for Parallel Line Assays

John Stufken
University of Georgia
jstufken@uga.edu

The focus in studying optimal parallel line assays for comparing two or more preparations has been on contrasts related to parallelism, a combined regression slope, and a comparison of the different preparations. The ultimate interest is often in a measure of relative potency. These experiments may be run in different design settings, such as completely randomized designs, randomized complete or incomplete block designs, row-column designs, and so on. A design that allows estimation of all of the aforementioned contrasts with full efficiency has been called an L-design.

We briefly review available results on the existence and construction of L-designs. We also present new results on the existence of row-column L-designs, leading to a simple necessary and sufficient condition for the existence of connected equireplicated L-designs. We close with some thoughts on additional criteria for identifying efficient designs in the context of parallel line assays.

For some of the available work we refer to Finney (1978), Gupta and Mukerjee (1996), and Chai (2002). The new results presented here are based on Chai and Stufken (2013).

Keywords: bioassay, optimal design, row-column design, connectedness.

Acknowledgements: This work was partially supported by NSF grant DMS–1007507. Support from the Institute of Statistical Science at Academia Sinica for a 3-week visit to collaborate with Dr. F.-S. Chai is also gratefully acknowledged.

References

Chai, F.-S. (2002): Block Designs for Asymmetrical Parallel Line Assays, *Sankhya B*, Vol. 64, N. 2, pp. 162-178.

Chai, F.-S., Stufken, J. (2013): Row-Column L-Designs for Symmetrical Parallel Line Assays, in preparation.

Finney, D.J. (1978): *Statistical Method in Biological Assay*, 3rd Edition, Charles Griffin, London.

Gupta, S,. Mukerjee, R. (1996): Developments in Incomplete Block Designs for Parallel Line Bioassays. In: Ghosh, S,. Rao, C.R. (Eds.) *Handbook of Statistics 13: Design and Analysis of Experiments*, Elsevier, Amsterdam, pp. 875-901.

Model selection approach for genome wide association studies in admixed populations

Piotr Szulc
Department of Mathematics and Computer Science
Wroclaw University of Technology, Poland
piotr.a.szulc@pwr.wroc.pl

The main purpose of Genome Wide Association Studies (GWAS) is the identification of genes responsible for quantitative traits (Quantitative Trait Loci, QTL) or disease causing genes in human populations. Localization of genes in such outbred populations is relatively difficult. The problem comes from the fact that due to cross-overs (exchange of genetic material among chromatids), which occur during production of reproductive cells, the statistical relation between a QTL genotype and a genotype of the neighbouring marker might be very small. Therefore the scientists need to use a huge number of densely spaced markers, which necessitates the application of stringent multiple testing corrections and results in a relatively low power of gene detection.

Modifications of Bayesian Information Criterion, mBIC and mBIC2, were successfully used in GWAS and results can be found e.g. in Frommlet et al. (2011). However, it turns out that we can find much more influential genes if we perform GWAS in *admixed population*, obtained as a result of interbreeding between previously separated ancestral populations. In that case, apart from genotypes, we have information about the origin of genome's fragments.

I will present the problem of localizing genes and show how to use modifications of model selection criteria in admixed population. Finally, I will present results of simulations.

Keywords: linear regression, model selection criteria, GWAS, admixed population

References

Frommlet F., Ruhaltinger F., Twarog P., Bogdan M. (2011): A model selection approach to genome wide association studies, *Computational Statistics and Data Analysis*

Chronological bias in randomized clinical trials

Miriam Tamm
Department of Medical Statistics
RWTH Aachen University
mtamm@ukaachen.de

In accrual clinical trials patients are often recruited over a long period of time. This may be caused by a limited number of diseased eligible patients per time. As a result of the lengthy recruitment period, time trends are suspected to occur. For example, this could be a result of changes in recruitment policy (ICH E9, 1998) or a learning effect in the application of the new method (Altman, 1988).
Randomization is used to balance out time trends between treatment groups. Nevertheless, bias can occur if a long run of patients is assigned to the same treatment group (Matts, 1978). To account for this, the randomization of subjects in blocks (Rosenberger, 2002) is explicitly recommended by the ICH guidelines (ICH E9, 1998) in order to get comparable groups. However, one major drawback of permuted block randomization is the increase in the risk of selection bias (Kennes, 2011; Tamm, 2012). Therefore, the benefit of permuted block randomization should be investigated in more detail.
To our knowledge, there exist no unified standard to evaluate the influence of chronological bias. Several aspects need to be examined to study the impact of chronological bias on the results of clinical trials. Hence, different methods to quantify chronological bias are discussed taking into account some developments in literature (Berger, 2003; Rosenkranz, 2011) and the influence of the chosen randomization procedure (random allocation rule, permuted block randomization, maximal procedure) is examined. Theoretical results regarding worst case sceanarios as well as results based on simulations for type I error and power are given.

Keywords: chronological bias, time trends, restricted randomization, permuted block design, accrual clinical trial.

References

Altman D.G., Royston J.P. (1988): The hidden effect of time, *Statistics in Medicine*, Vol. 7, N. 6, pp. 629-637.

Berger V.W., Ivanova A., Knoll M.D. (2003): Minimizing predictability while retaining balance through the use of less restrictive randomization procedures, *Statistics in Medicine*, Vol. 22, N. 19, pp. 3017-3028.

ICH E9. Statistical Principles for Clinical Trials. Website, 1998. Available at: http://www.ich.org/ [Accessed 20 February 2013].

Kennes, L. N., Cramer, E., Hilgers, R. D., Heussen, N. (2011): The impact of selection bias on test decisions in randomized clinical trials, *Statistics in Medicine*, Vol. 30, N. 21, pp. 2573-2581.

Matts J.P., McHugh R.B. (1978): Analysis of accrual randomized clinical trials with balanced groups in strata, *Journal of Chronic Diseases*, Vol. 31, N. 12, pp. 725-740.

Rosenkranz G.K. (2011): The impact of randomization on the analysis of clinical trials, *Statistics in Medicine*, Vol. 30, N. 30, pp. 3475-3487.

Rosenberger W.F.,Lachin J.M.(2002): *Randomization in Clinical Trials: Theory and Practice*, Wiley, New York.

Tamm M., Cramer E., Kennes L. N., Heussen N. (2012): Influence of selection bias on the test decision: A simulation study, *Methods of Information in Medicine*, Vol. 51, N. 2, pp. 138-143.

Multichannel Queuing Systems in a Random Environment

Tkachenko Andrey
Lomonosov Moscow State University
tkachenko.av.87@gmail.com

We consider queuing systems with r heterogeneous channels. Service times of customers are independent random variables. The service time η_n^i of the n-th customer by the i-th server has distribution function $B_i(x)$ with finite mean β_i^{-1}. Let $\beta = \sum_{i=1}^{r} \beta_i$. Customers are served in order of their arrivals at the system.

The input flow $X(t)$ is assumed to be regenerative. Let θ_i be the i-th regeneration point of $X(t)$, $\tau_i = \theta_i - \theta_{i-1}$, $\xi_i = X(\theta_i) - X(\theta_{i-1})$ ($i = 1, 2, \ldots; \theta_0 = 0$). Then τ_i is the regeneration period, ξ_i is the number of customers arrived during the i-th regeneration period. Assume that $a = E\xi_i < \infty$, $\tau = E\tau_i < \infty$, and $\lambda = \lim\limits_{t\to\infty} \frac{X(t)}{t} = a\tau^{-1}$ $a.s.$.

The random environment can destroy all the servers simultaneously. After breakdown the servers are repaired during the random time. If the customer's service was interrupted by the breakdown then it is continued after repair from the point at which it was interrupted. We assume that the working periods $\{u_n^1\}_{n=1}^{\infty}$ of the system have exponential distribution with mean a_1, and the periods of reconstruction $\{u_n^2\}_{n=1}^{\infty}$ are i.i.d.r.v's with distribution function $G(x)$ and mean a_2. Let $q(t)$ be the number of customers in the system at time t. Under some additional assumptions $q(t)$ is a regenerative process and θ_i is it's point of regeneration if $q(\theta_i - 0) = 0$ and the system is in the working state.

Theorem 1. The process $q(t)$ is ergodic iff $\rho = \frac{\lambda(a_1+a_2)}{a_1\beta} < 1$.

First we give the following result concerning so called super-heavy traffic situation.

Theorem 2. If $\rho > 1$ ($\rho = 1$) and for some $\delta > 0$

$$E\tau_1^{2+\delta} < \infty,\ E\xi_1^{2+\delta} < \infty,\ E(u_1^2)^{2+\delta} < \infty,\ E(\eta_1^i)^{2+\delta} < \infty,\ i = \overline{1,r}, \tag{$\star$}$$

then the normalized process $\hat{q}_n(t) = \frac{q(nt)-\beta(\rho-1)nt}{\hat{\sigma}\sqrt{n}}$ weakly converges on any finite interval $[0,t]$ to Brownian motion (absolute value of Brownian motion) as $n \to \infty$. Here

$$\hat{\sigma}^2 = \sigma_X^2 + \lambda\beta^2\sigma_\beta^2 + \beta^2\sigma_S^2,\ \sigma_X = \frac{\sigma_\xi^2}{\tau} + \frac{a^2\sigma_\tau^2}{\tau^3} - \frac{2acov(\xi_1,\tau_1)}{\tau^2},$$

$$\sigma_\beta^2 = \sum_{i=1}^{r} Var(\eta_1^i),\ \sigma_\tau^2 = Var(\tau_1),\ \sigma_\xi^2 = Var(\xi_1),$$

$$\sigma_S^2 = \frac{a_1^2\sigma_2^2 + a_2^2\sigma_1^2}{(a_1+a_2)^3},\ \sigma_i^2 = Var(u_1^i),\ i = 1,2.$$

For heavy traffic situation we consider time-compression asymptotic. Namely the input flow is given by the relation

$$X_n(t) = X\left(\rho^{-1}\left(1 - \frac{1}{\sqrt{n}}\right)t\right),$$

so that the traffic coefficient depends on the parameter n and $\rho_n \uparrow 1$ as $n \to \infty$. Let $q_n(t)$ be the process $q(t)$ for the system with input flow $X_n(t)$. Then under conditions $(\star)$ the normalized process $\tilde{q}_n(t) = \frac{q_n(nt)}{\sqrt{n}}$ weakly converges on any finite interval $[0,t]$ as $n \to \infty$ to the diffusion process with reflecting at the origin and coefficients $(-\sqrt{\alpha}, \tilde{\sigma}^2)$. Here $\tilde{\sigma}^2 = \alpha\beta\sigma_\beta^2 + \frac{\alpha}{\lambda\beta}\sigma_X^2 + \sigma_S^2$ and $\alpha = \frac{a_1}{a_1+a_2}$.

Keywords: multichannel queuing system, regenerative process, random environment, heavy traffic.

Acknowledgements: this work was partially supported by RFBR grant 13-01-00653.

Algorithm of Approximate Solution of Traveling Salesman Problem

Tatiana M. Tovstik
St. Petersburg State University
peter.tovstik@mail.ru

Ekaterina V. Zhukova
catich06@mail.ru

The salesman problem is to find the closed path x with the minimal length $H(x)$, which crosses N given points exactly one time. As a first step the good initial approximation is proposed. The initial points are reflected into a unit square. The polar co-ordinates r, ϕ with the origin in the square center are introduced, and in the initial approximation (see the left side in Fig. 1) all points are numbered according the growth of the angle ϕ. The following approximations are constructed by using the simulation of dynamic fields by Metropolis (1953).

If in the k-th approximation $x = x^{(k)}$, then the path $x^{(k)}$ is $x^{(k)} = \{j_1^{(k)} \to j_2^{(k)} \to \cdots \to j_N^{(k)} \to j_{N+1}^{(k)} = j_1^{(k)}\}$. In partial, $x^{(0)} = \{1 \to 2 \to \cdots \to N \to 1\}$ defines the initial path.

The energy function $H(x^{(k)})$ is equal $H(x^{(k)}) = \sum_{i=1}^{N} r_i^{(k)}$, where $r_i^{(k)} = r(j_i^{(k)}, j_{i+1}^{(k)})$ is the distance between points $j_i^{(k)}$ and $j_{j+1}^{(k)}$.

The next approximation is performed by the two-change, at which accidentally two numbers of the previous path are chosen, and the direction of motion between them is changed to inverse. The obtained path y is considered as a test path.

If $\Delta H = H(y) - H(x^{(k)}) \le 0$, then we put $x^{(k+1)} = y$. If $\Delta H > 0$, then $x^{(k+1)} = y$ with the positive probability $P = e^{-\beta \Delta H}$, and in the opposite case the path y is rejected $x^{(k+1)} = x^{(k)}$.

If the annealing coefficient $\beta \to \infty$, then the Metropolis method converges to the path with minimum $H(x)$.

The following expression for the probability P is proposed

$$P = \exp\left(-\frac{\beta_* \, \Delta H}{\sqrt{(\hat{\sigma}^{(k)})^2 + (\hat{\sigma}^{(k+1)})^2}}\right),$$

where $(\hat{\sigma}^{(k)})^2 = \frac{1}{N}\sum_{i=1}^{N}(r_i^{(k)})^2 - (\frac{1}{N}\sum_{i=1}^{N} r_i^{(k)})^2$, and β_* is the dimensionless annealing coefficient (we take $2 \leq \beta_* \leq 7$).

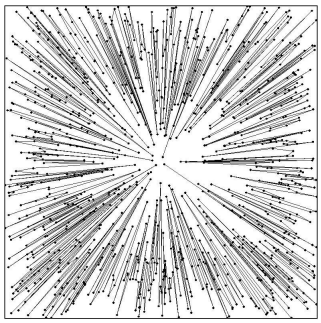

Fig. 1. The initial path (left) and the final path (right)

At the large N we decrease the general path length by optimizing the separate parts of path. Also it is important to delete the cross-sections.

For the random independent homogeneous distribution of points the criteria of the path optimality is the normalized length $\gamma = H/\sqrt{NA} \approx 0.749$, where A is the area, containing all points (Beardwood, 1959).

The proposed algorithm gives the path, close to optimal, and sometimes optimal. This conclusion is supported by some examples. Consider one of them. The $N = 1000$ random independent uniformly distributed points in a unite square are studied. The final path with $\gamma = 0.757$ is shown in Fig. 1. The next examples of the points disposition are taken from the Internet library $TSPLIB$, and sometimes the proposed algorithm gives the better result than in this library. In partial, for $TS225$ we get $\gamma = 0.6978$.

Keywords: traveling salesman problem, Metropolis annealing.

Acknowledgements : This work was supported by RFBR, grant 11-01-00769-a.

References

Metropolis N., et al. (1953): Equations of state calculations by fast computing machines, *J. Chem.Phys.* Vol. 21, pp. 1087-1092.

Beardwood J., et al. (1959): The shortest path through many points, *Proc. Cambridge Phil. Soc.* Vol. 55, pp. 299-327.

Flexible Parametric and Semiparametric Inference for Longitudinal Data with a Censored Covariate

John V. Tsimikas
University of the Aegean, Dept. of Statistics and Actuarial-Financial Mathematics
tsimikas@aegean.gr

Leonidas E. Bantis
University of the Aegean, Dept. of Statistics and Actuarial-Financial Mathematics
lbantis@aegean.gr, lbantis@aegean.gr

Censoring of the covariate can occur in many applications. In the Biostatistics field common examples are 1) when interest lies in assessing the time-dependent accuracy of a biomarker where survival time is considered a covariate and 2) when a covariate is subject to a limit of detection. We address the problem of inference in a repeated measures setting when a covariate is subject to censoring via an estimating function approach. Our method does not necessarily assume a parametric form for the distribution of the response given the regressors. In the linear regression case, the proposed approach implies the use of mean imputation of the censored regressor accompanied by a simple adjustment made to the working covariance matrix. We show how the use of flexible parametric models for the distribution of the covariate can be employed. When survival time is considered as the covariate subject to censoring, the use of the generalized gamma distribution is explored, since it is considered as a platform distribution covering a wide variety of hazard rate shapes. The method is further robustified with the use of a constrained natural spline (CNS) that allows us to obtain a smooth monotone estimate of the cumulative hazard function for the distribution of the censored covariate. For models involving additional,

fully observed, covariates we propose either the use of the generalized gamma accelerated failure time regression or an extension of the CNS method under the proportional hazard model framework. The proposed approach is in some cases broader than likelihood based multiple imputation techniques. Moreover, even in cases with a known parametric form for the response distribution, the method can be considered a feasible alternative to likelihood based estimation. We present applications involving biomarkers and discuss ROC estimation based on our modeling approach.

Keywords: censored covariate, estimating equations, generalized linear model, natural cubic spline, restricted least squares, survival estimation.

The Supertrack Approach as a Classical Monte Carlo Scheme

Egor Tsvetkov
Moscow Institute of Physics and Technology
tsvetkov_egor@mail.ru

Boltzmann tallies are the linear functionals on the solutions of Boltzmann equations (Booth, 1994). Monte Carlo methods based on the Neumann-Ulam scheme can be used to estimate the Boltzmann tallies. However, not all real-world calculations can be represented as such tallies. The pulse height tally is example of non-Boltzmann tally.

Thomas E. Booth introduced variance reduction techniques to estimate non-Boltzmann tallies (Booth, 1994). The main idea is to consider the branching trajectory as indivisible collection of tracks that is referred as supertrack. When we want to split the particle we have to split the whole supertrack. This rule leads us to the idea that supertrack approach can be derived from general theoretical probability approach if we consider the supertrack as the elementary event.

Original article (Booth, 1994) discusses importance sampling, splitting, Russian Roulette and DXTRAN. In this paper we discuss also forced collisions technique.

We depart from describing the probability space $(\mathbb{S}, \mathcal{H}, P)$ on the set of all branching trajectories $\mathbb{S}$. Then we apply the variance reduction techniques to estimate the Lebesgue integral

$$I = \int_{\mathbb{S}} q\,(S)\,P\,(dS). \tag{1}$$

The applied techniques are importance sampling, Russian Roulette game, splitting and stratified sampling.

When applied, the importance sampling, Russian Roulette game and splitting immediately lead us to the formulation of the rules how to calculate the weight of the trajectory. These rules are identical to those in

supertrack approach.

To apply the stratified sampling method we have to partition the set of branching trajectories $\mathbb{S}$ into countable number of subsets. Depending on what kind of partition is chosen we derive the forced collisions technique or DXTRAN.

Let us consider the forced collision technique. Given the trajectory, we can say how many times the trajectory enters the selected region where the forced collision is applied. Each time the trajectory enters the region we can say if the particle runs throughout the region without collision or not. So, we can partition set $\mathbb{S}$ into subsets $\mathbb{T}_k$ where k indicates how many times trajectory enters the region. Each $\mathbb{T}_k$ we partition into 2^k subsets $\mathbb{T}_{ki}$, where i indicates what branch was selected at each intersection of the region (collided or non-collided). All $\mathbb{T}_{ki}$ generate the countable partition of $\mathbb{S}$. When we apply the stratified sampling with this partitioning we derive the formulation of the forced collision technique within the supertracks approach.

Let us consider the DXTRAN technique. Given the trajectory we can say after which collision it enters the DXTRAN sphere. This possibility shows how to partition $\mathbb{S}$ in case of DXTRAN. We denote the subset of trajectories that enters DXTRAN sphere after k-th collision as $\mathbb{T}_k$, and non-DXTRAN trajectories as $\mathbb{T}_0$. With this kind of splitting we derive the DXTRAN technique.

Keywords: supertracks, DXTRAN, forced collisions

References

Booth Th.E.(1994): Monte Carlo variance reduction approaches for non-Boltzmann tallies, *Nuclear Science and Engineering*, Vol. 116, pp. 113-124.

The dependence of the ergodicity on the time effect in the repeated measures ANOVA with missing data based on the unbiasedness recovery

Anna Ufliand
Saint Petersburg State University
anna.uflyand@gmail.com

Nina Alexeyeva
Saint Petersburg State University
ninaalexeyeva@mail.ru

The Ergodic missing data analysis(Alexeyeva, 2012) is based on repeated measures ANOVA and was created for studying longitudinal data, containing gaps with full data at least in one point.

The model
The following form of analysis of variance is taken into consideration:

$$x_{ijk} = \mu + \alpha_i + \varepsilon^1_{ij} + \beta_k + \gamma_{ik} + \varepsilon_{ijk},$$

where:
μ – general mean
α_i – fixed effect of group
β_k – the effect of time
γ_{ik} – interaction effect of time and group
$\varepsilon^1_{ij}, \varepsilon_{ijk}$ – independent normally distributed unbiased individual and total errors, $i = 1 \dots r, j = 1 \dots \nu_i, k = 1 \dots \mathbf{T}$.

By introducing individual displacements $\mathbf{H}_{ij}$ as sums of series, based on infinite sequences of cross-averages, it became possible to split it into two unbiased models:

- $x_{ij\cdot} - \mathbf{H}_{ij} = \mu + \alpha_i + \Theta^1_{ij}$
- $x_{ijk} - x_{ij\cdot} + \mathbf{H}_{ij} = \beta_k + \gamma_{ik} + \Theta_{ijk}$, where:

$\Theta^1_{ij}, \Theta_{ijk}$ – new unbiased, but correlated errors.

To estimate the parameters and to test the hypotheses about the significance of the effects one needs only to construct the correlation matrices of these errors.

Ergodic property

The main problem was to investigate the fact of the improving of the ergodic property of data by subtraction of the individual displacements $\mathbf{H}_{ij}$ from initial data.
The ergodic property here should be understood in its physical meaning of the lack of the difference between time and space counted means.
As the result of this work, it became known that the improvement of the ergodic property depends on the significance of time effect.
In other words, the presence of the trend as an increasing(decreasing) monotonic function in data mostly allows to say about improving the ergodic properties in case of positive(negative) covariance between the number of presented sample values and the value of the displacement. This condition often occurs for example with the growth of the amount of missing data in time.
The rate of trend's increase(decrease) affects the degree of confidence with which we can claim about this fact.

Keywords: repeated measures ANOVA, missing data, unbiasedness, ergodic property

References

Alexeyeva N.P. (2012): *Analysis of biomedical systems. Reciprocity. Ergodicity. Synonymy*, Publishing of the Saint-Petersburg State University, Saint-Petersburg.

Afifi A.A, Azen S.P. (1972): *Statistical analysis: a computer oriented approach*, Academic Press, New York.

A contribution review In Memoriam of Professor Reuven Rubinstein

Slava Vaisman
Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology, Haifa , Israel
slava@tx.technion.ac.il

In this report, we cover major contributions made by Prof. Reuven Rubinstein (1938-2012) in the field of Monte Carlo Simulation. During his scientific career Reuven authored more than one hundred papers and six books. His research focused on various fields of applied probability, such as adaptive importance sampling, rare event simulation, stochastic optimization, sensitivity analysis and counting in NP-complete problems.

In 2010 Prof. Rubinstein won the INFORMS Simulation Society highest prize - the Lifetime Professional Achievement Award (LPAA), which recognizes scholars who have made fundamental contributions to the field of simulation that persist over most of a professional career.
In 2011 Reuven Rubinstein won the Operations Research Society of Israel (ORSIS) highest prize - the Lifetime Professional Award (LPA), which recognizes scholars who have made fundamental contributions to the field of operations research over most of a professional career.

In this report we concentrate on his fundamental results:

- The *score function* (SF) method. While analyzing complex discrete-event systems, we are interested not only in performance (estimator) evaluation, but also in sensitivity analysis of the later. The purpose of *score function* method is to estimate the gradient and higher derivatives of the corresponding estimator (Rubinstein 1986).

- The *Cross-Entropy* (CE) method. CE is an efficient procedure for

the estimation of rare-event probabilities. The method proved to be a very powerful technique capable to handle hard rare-event and combinatorial optimisation problems. During the algorithm execution, samples of random data are repeatedly generated, then, the parameters of random generation mechanism are updated according to the previous samples in order to produce "better" ones in the next generation. (Rubinstein and Kroese 2004,2008).

- The *Stochastic Enumeration* (SE) Algorithm. SE is a new generic sequential importance sampling scheme for counting $\sharp P$ complete problems such as the number of satisfiability assignments and the number of perfect matchings in the graph (permanent). The SE provides a natural generalization of the classic one-step-look-ahead algorithm in sense that it runs many trajectories in parallel instead of one and employs a polynomial time decision making oracle. (Rubinstein, 2011)

Keywords: Simulation, rare events, applied probability, stochastic optimization, sensitivity analysis, counting.

References

Rubinstein R.Y. and Kroese D.P. (2008): *Simulation and the Monte Carlo Method, 2nd Edition*, Wiley, New York.

Rubinstein R.Y. and Kroese D.P. (2004): *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, Springer-Verlag New York,

Rubinstein, R.Y. (1986): The Score Function Approach for Sensitivity Analysis of Computer Simulation Models", *Mathematics and Computers in Simulation*, Vol.28, pp.351-379

Rubinstein R. Y. (2011): Stochastic Enumeration Method for Counting NP-hard Problems, *Methodology and Computing in Applied Probability* , pp. 1-43

Sequential Monte Carlo Method for counting vertex covers

Radislav Vaisman
Faculty of Industrial Engineering and Management,
Technion, Israel Institute of Technology, Haifa, Israel
slava@tx.technion.ac.il

In graph theory, a *vertex cover* of a graph is a set of vertices such that each edge of the graph is incident to at least one vertex of the set. In this article we are interested in counting all vertex covers in a graph. The area of counting, and in particular the definition of $\sharp P$ complete class introduced by Valiant (1979) received much attention in the Computer Science community. For example, Karp and Lubby (1983) introduced a *FPRAS* (A fully polynomial randomized approximation scheme) for counting the solutions of *DNF* satisfiability formula. Similar results were obtained for the *Knapsack* and *Permanent* problems. See Jerrum, Sinclair, Vigoda (2004) and (Dyer, 2003).On the other hand, there are many 'negative' results. For example, Dyer, Frieze and Jerrum (1999) showed that there is no FPRAS for $\sharp IS$ (Independent Set) if the maximum degree of the graph is 25 unless RP = NP. Counting the number of vertex covers remains hard even when restricted to planar bipartite graphs of bounded degree, or regular graphs of constant degree. See Vadhan (1997) for details.

We propose a Sequential Importance Sampling procedure for counting the number of Vertex Covers in a graph. In spite of the fact that counting algorithms based on Importance Sampling seems to perform very well in practice, it is also known that their performance depends heavily on the closeness of the proposal distribution to the uniform one. Our algorithm introduces probabilistic relaxation technique combined with Dynamic Programming, in order to obtain efficient estimate of this distribution. At each step, the procedure decides whether to include the given vertex in the cover set or not. The decision is based on the approximated number of covers in each case. During the algorithm execution, we guarantee that the sequential procedure always produces a valid ver-

tex cover while saving computation effort. Moreover, the algorithm can supply a probabilistic lower bound that is calculated online.

Our numerical results indicate that the proposed scheme compares favorably with other existing methods. We were able to handle large graph problems in reasonable time. In particular we compare with *cachet* - an exact model counter, and, with the state of the art *SampleSearch* that is based on Belief Networks and Importance Sampling.

Keywords: Counting, Sequential Importance Sampling, Dynamic Programming, Relaxation, Random Graphs.

References

Dyer M. (2003): Approximate counting by dynamic programming. *In Proceedings of the 35th ACM Symposium on Theory of Computing.* pp. 693-699.

Dyer M., Frieze A. ,Jerrum M. (1999): On counting independent sets in sparse graphs. *In 40th Annual Symposium on Foundations of Computer Science.* pp. 210-217.

Jerrum M., Sinclair A., Vigoda E. (2004): A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries, *Journal of the ACM*, pp. 671-697.

Karp R.M., Luby M. (1983): Monte-carlo algorithms for enumeration and reliability problems, *In Proceedings of the 24th Annual Symposium on Foundations of Computer Science, SFCS'83.*, pp. 56-64.

Vadhan S.P. (1997): The complexity of counting in sparse, regular, and planar graphs. *SIAM Journal on Computing.* Vol. 31, pp. 398-427.

Valiant L.G. (1979): The complexity of enumeration and reliability problems, *SIAM J. Comput.*, Vol. 8, N. 3, pp. 410-421.

MCMC estimation of directed acyclic graphical models in genetics

Stéphanie M. van den Berg
Department of Research Methodology, Measurement, and Data Analysis
University of Twente, the Netherlands
stephanie.vandenberg@utwente.nl

Genetic models are used to estimate the extent to which genetic variation explains variation in an observed trait. General models in both human and animal genetics decompose total trait variance into genetic variance and non-genetic (environmental) variance. This is relatively straightforward for traits that are assessed by manifest measures like height in humans and milk yield in cows, but not for traits that are more indirectly measured, like personality or intelligence. Such traits are often measured using test items. The reliability of the measurements is then related to the number of test items and the quality of the items in the test. To avoid the confounding of measurement error with true environmental variance, such traits assessed with measurement error can be analysed with a combination of a measurement error model and a structural model. The structural model is a genetic model that defines the phenotype, and the measurement model is an Item-Response Theory (IRT) model. For examples of this approach, refer to van den Berg, Glas and Boomsma (2007) and van den Berg and Service (2012). The combined models can be estimated in a Bayesian framework using MCMC sampling. In more advanced applications, this often requires recasting existing models into directed acyclic graphical (DAG) models. Several more complex genetic models will be discussed, such as models involving gene-environment interaction, assortative mating (non-random mating where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating) and inbreeding (mating of parents which are closely related genetically). The models will be illustrated using computational

examples with empirical data.

Keywords: Bayesian Statistical Methods, directed acyclic graphical (DAG) models, Genetic models, Item Response theory, MCMC.

References

Van den Berg, S.M., Glas, C.A.W., Boomsma, D. I. (2007). Variance Decomposition Using an IRT Measurement Model. *Behavior Genetics, 37,* 604-616.

Van den Berg, S.M., Service, S.K. (2012). Power of IRT in GWAS: Successful QTL Mapping of Sum Score Phenotypes Depends on Interplay Between Risk Allele Frequency, Variance Explained by the Risk Allele, and Test Characteristics. *Genetic Epidemiology, 36,* 882-889.

Use of Doehlert Designs for second-order polynomial models

L. Rob Verdooren
Danone Research, Centre for Specialised Nutrition ,Wageningen, The Netherlands
rob.verdooren@danone.com

The most popular designs for fitting the second-order polynomial model are the central composite designs of Box and Wilson (1951) and the designs of Box and Behnken (1960). For k = 2, 4, 6 and 8, the uniform shell designs of Doehlert (1970) require fewer experimental runs than the central composite or Box-Behnken designs. In analytic chemistry the Doehlert designs are widely used. The uniform shell designs are based on a regular simplex, this is the geometric figure formed by k + 1 equally spaced points in a k dimensional space; an equilateral triangle is a two-dimensional regular simplex. The shell designs are used for fitting a response surface to k independent factors over a spherical region. Doehlert (1930 1999) proposed in 1970 the design for k = 2 factors starting from an equilateral triangle with sides of length 1, to construct a regular hexagon with a centre point at (0,0). The n = 7 experimental points are (1,0), (0.5 , 0.866) , (0, 0), (-0.5, 0.866), (-1, 0), (-0.5, -0.866) and (0.5, -0.866).The 6 outer points lie on a circle with a radius 1 and centre (0, 0). This Doehlert design has an equally spaced distribution of points over the experimental region, a so-called uniform space filler, where the distances between neighboring experiments are equal. Response surface designs are usually applied by scaling the coded factor ranges to the ranges of the experimental factors. The first factor covers the interval [-1, + 1], the second factor covers the interval [-0.866 , +0.866].
A Doehlert design for four factors needs 21 trials with the intervals for the factors respectively [-1, +1], [-0.866 , + 0.866], [-0.816, + 0.816] and [-0.791, +0.791]. The 21 design points are (0, 0, 0, 0), (1, 0, 0,

0), (0.5, 0.866, 0, 0), (-0.5,0.866, 0, 0), (-1, 0, 0, 0), (-0.5, -0.866, 0,0), (0.5, -0.866, 0 , 0), (0.5, 0.289, 0.816, 0), (-0.5, 0.289, 0.816, 0), (0,-0.577, 0.816, 0), (0.5, -0.289, -0.816, 0), (-0.5, -0.289, -0.816,0), (0, 0.577, -0.816, 0), (0.5, 0.289, 0.204, 0.791), (-0.5, 0.289, 0.204, 0.791), (0, -0.577, 0.204, 0.791), (0, 0, -0.612, 0.791), (0.5, -0.289, -0.204, -0.791), (-0.5, -0.289, -0.204, -0.791), (0, 0.577, -0.204, -0.791), (0, 0, 0.612, -0.791).

Doehlert and Klee (1972) show how to rotate the uniform shell designs to minimize the number of levels of the factors. Most of the rotated uniform shell designs have no more than five levels of any factor; the central composite design has five levels of every factor.

The determinant criterion of the variance matrix of Doehlert designs will be compared with central composite designs and Box-Behnken designs, see Rasch et al. (2011).

Keywords: second-order polynomial designs, quadratic response designs, Doehlert designs.

References

Box, G.E.P. and Behnken, D.W. (1960): *Some new three-level designs for the study of quantitative variables*, Tehnometrics, 2, 455 475.

Box, G.E.P. and Wilson, K.B. (1951): *On the experimental attainment of optimum conditions*, Journal of the Royal Statistical Society, Ser. B, 13, 1 45.

Doehlert, D.H. (1970): *Uniform shell designs*, Journal of the Royal Statistical Society, Ser. C, 19, 231 239.

Doehlert, D.H. and Klee, V.L. (1972): *Experimental designs through level reduction of the d-dimensional cuboctahedron*, Discrete Mathematics, 2, 309 334.

Rasch, D., Pilz, J., Verdooren, R. and Gebhardt, A. (2011): *Optimal experimental design with R*, Chapman & Hall /CRC, Boca Raton FL, USA.

Assessing errors in CMM measurements via Kriging and variograms: a simulation study

Grazia Vicario
Department of Mathematical Sciences, Politecnico di Torino, Torino
grazia.vicario@polito.it

Suela Ruffa
Department of Production Systems and Economics, Politecnico di Torino, Torino
suela.ruffa@polito.it

Giovanni Pistone
Collegio Carlo Alberto, Moncalieri, Torino
giovanni.pistone@carloalberto.org

Industrial parts are routinely affected by dimensional and geometric errors due to the manufacturing processes used for their production. These errors, that usually have a typical pattern related to the employed manufacturing process, are limited by means of dimensional and geometrical tolerances (such as straightness, roundness, flatness, profile) that have to be verified on the manufactured parts. Coordinate Measuring Machines (CMM) are the most common equipment for 3D measurement because of their accuracy and flexibility.
In the present paper we focus on the inference on the error of different planar surfaces whose tolerances are verified using a CMM. For this purpose we suggest the prediction of the surface model using a Kriging model on a set of measured points. Kriging is a stochastic linear interpolation technique that predicts the response values at untried locations with weights assigned to the tried locations. The weights are selected so that the estimates are unbiased (repeatedly Kriging we expect the correct result on average) and they have minimum variance. The fundamentals is the rate at which the variance between points changes over space. This is expressed as a variogram which shows how the average

difference between values at points changes; it is a function of the distance and of the corresponding direction of any pair of points depicting their correlation extent. Theoretically, it is defined as the variance of the difference between the response values at two locations and it is equivalent to the correlation function for stationary processes. The use of the variogram instead of the correlation function is recommended by the geostatisticians even if the process is not stationary.
In this paper we resort to variograms to detect possible manufacturing signatures, i.e. systematic pattern that characterizes all the features manufactured with a particular production process, and systematic errors of the CMM measurement process. We simulate different, and most common, manufacturing signatures of a planar surface and possible errors of a measurement process with CMM. The variograms are estimated using the most robust empirical estimator in the case at hand and the likelihood (or restricted likelihood) estimator. The behavior of the omnidirectional variogram suggests the spatial correlations, giving evidence of possible non isotropy.

Keywords: Kriging Model, Spatial Correlation, Variogram, Anisotropy, Geometric Errors.

References

Cressie, N. A, (1997): Spatial Prediction and ordinary kriging, *Mathematical Geology*, 20(4), 407-421.

Haslett, J., (1997): On sample variogram and the sample autocovariance for non-stationary time series. *The Statistician*, 46, 475-485.

Jin, N., Zhou, S., (2006): *Signature construction and matching for fault diagnosis in manufacturing processes through fault space analysis*, IIE Transactions, 38:4, 341-354.

Estimating power grid reliability using a splitting method

Wander Wadman
CWI Amsterdam
w.wadman@cwi.nl

Daan Crommelin, Jason Frank
CWI Amsterdam
Daan.Crommelin@cwi.nl, J.E.Frank@cwi.nl

Contemporary western societies have grown accustomed to a very reliable electricity supply by power transmission grids. However, substantial implementation of intermittent renewable generation, such as photovoltaic power or wind power, may threaten grid reliability. Power imbalances caused by generation intermittency may force grid operators to curtail power to ensure grid stability. As they are obliged to keep reliability at a prescribed level, grid operators must be able to perform a quantitative reliability analysis of the grid.

For this purpose, various grid reliability indices exist (Billinton et al., 1994), and most of them depend on the probability $P(C)$, where C denotes the event of a power curtailment during the time interval $[0, T]$ of interest. We model the uncertain energy sources as stochastic processes, discretized in time. At each time step, the mapping of these sources to the occurence of a curtailment C requires solving a nonlinear algebraic system. Since this mapping is only implicitly defined, we can not derive $P(C)$ directly, and we estimate it by a Monte Carlo simulation.

However, as power curtailments are undesirable (for T equal to one week, $P(C) < 10^{-4}$ could typically be desired), we may expect their occurence to be rare. Crude Monte Carlo (CMC) estimators for rare event probabilities require a large number of samples to achieve a fixed accuracy level (Rubino et al., 2009). Since one CMC sample already involves solving a large number of nonlinear systems, CMC estimation is computationally too intensive for general grid reliability analyses.

To reduce the computational burden, we apply rare event simulation

techniques. We use the insight that we can write most important reliability indices as expectations $\mathbb{E}[I]$, which we can decompose as

$$\mathbb{E}[I] = P(C)\mathbb{E}\left[I|C\right].$$

This decomposition suggests unbiased estimation of $\mathbb{E}[I]$ by $\hat{I} := \hat{P}\hat{I}_C$, a product of independent unbiased estimates for $P(C)$ and $I_C := \mathbb{E}\left[I|C\right]$, respectively. To reduce the relative variance of the estimator

$$\frac{\mathrm{Var}(\hat{I})}{\mathbb{E}^2[I]} = \frac{\mathrm{Var}(\hat{P})}{P(C)^2} + \frac{\mathrm{Var}(\hat{I}_C)}{I_C^2} + \frac{\mathrm{Var}(\hat{P})}{P(C)^2}\frac{\mathrm{Var}(\hat{I}_C)}{I_C^2},$$

we reduce the first two terms on the right-hand side using an appropriate splitting technique called Fixed Number of Successes (FNS) (Amrein et al., 2011). This technique controls the relative variance of $\hat{P}$ by fixing the number of hits per level. By performing one additional splitting when the rare event set is hit, we can control the second term as well. In this way, we control the precision of the reliability index estimate $\hat{I}$. Simulations show that for different levels of desired precision, FNS requires less computational effort than CMC.

Keywords: Rare event simulation, splitting methods, reliability analysis, power grids, renewable energy.

References

Amrein M., Künsch, H. (2011): A Variant of Importance Splitting for Rare Event Estimation: Fixed Number of Successes, *ACM Trans. on Modeling and Computer Simulation*, Vol. 21, N. 2.

Billinton R., Li W. (1994): *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*, Plenum Press, N.Y.

Rubino, G., Tuffin, B. (2009): *Rare event simulation using Monte Carlo methods*, John Wiley & Sons Ltd., Chichester.

Optimal designs for hierarchical generalized linear models

Tim Waite and Dave Woods
Statistical Sciences Research Institute, University of Southampton, UK
tww1g08@soton.ac.uk, D.Woods@soton.ac.uk

Peter Van de Ven
p.vandeven@vumc.nl
VU University Medical School, Amsterdam, The Netherlands

In recent years there has been much interest in experiments with two features: (i) a non-normally distributed response variable, and (ii) grouping of the experimental units into groups, or *blocks*, within which the experimental units are more homogeneous. The first feature is exemplified by experiments with discrete responses, for instance in pharmaceutical applications where the response may correspond to the formation, or not, of a crystalline product. Blocking occurs in many situations, being particularly common in industrial process research where there are often batch effects.

There are two main approaches which can address both of these features simultaneously: generalized linear mixed models (GLMMs), and the less popular hierarchical generalized linear models (HGLMs). Both of these extend generalized linear models by allowing random effects in the predictor. The latter models are fitted by maximization not of the marginal likelihood, but of the h-likelihood. The introduction of this estimation method sparked some controversy, with critics raising doubts about the range of applicability of asymptotic results. Nonetheless, we are able to demonstrate that the claimed theoretical properties of the procedure hold adequately in the examples we consider. One advantage of h-likelihood is the easy availability of estimates and standard errors for the individual block effects. A second benefit is its computational simplicity when compared to marginal likelihood.

We discuss the algorithmic construction of designs for HGLM experiments. A design optimality criterion is developed which is appropriate when the analysis is to be conducted using h-likelihood. A coordinate optimization algorithm is applied to several examples of multifactor experiments. The case when some factors are 'hard-to-change' is considered, yielding both split-plot and whole-plot design structures. We make observations on the differences between HGLM and GLMM designs. A pseudo-Bayesian procedure is considered for obtaining designs which perform effectively when a priori the parameter values are uncertain.

Keywords: blocking, discrete response, optimal design, random effects.

Acknowledgements: This work was supported by PhD studentship funding for the first author from the UK Engineering and Physical Sciences Research Council, and the Statistical Sciences Sciences Research Institute, University of Southampton, as well as a Research Fellowship from the School of Mathematics, University of Southampton.

Quadrature rules for polynomial chaos expansions using the algebraic method in the design of experiments

Henry P Wynn
London School of Economics
h.wynn@lse.ac.uk

Jordan Ko
London School of Economics
jordan.ko@me.com

A general method for quadrature for uncertainty quantification (UQ) is introduced based on the algebraic method in experimental design see, Pistone, Riccomgano and Wynn (2001) and the review Maruri-Aguilar and Wynn (2012). This is a method based on the theory of zero dimensional algebraic varieties. It allows quadrature for polynomials $p(x),\ x \in R^d$, or polynomial approximands, for quite general sets of quadrature points, here called "designs". Thus, let ξ be a probability measure on R^d with finite moments: $\mu_\alpha = \mathrm{E}_\xi(x^\alpha)$. Also, let $D = \{z_1, \ldots, z_n\}$ be a finite set of distinct quadrature points, the "design", in R^d. Let $G = \{g_1, \ldots, g_m\}$ be a Gröbner basis of $I(D)$ with respect to the chosen monomial ordering, $\prec$, and L the corresponding set of exponents for the basis of the quotient ring. Following the algebraic theory, any polynomial $p(x) = p(x_1, \ldots, x_d)$ decomposed as

$$p(x) = \sum_{i=1}^{m} s_i(x) g_i(x) + r(x)$$

where $r(x)$ is a member of the quotient ring with basis elements $x^\alpha, \alpha \in L$, and recall that $|L| = |D|$. The remainder $r(x)$ can be written as $r(x) = \sum_{z \in D} p(z) l_z(x)$ where the $l_z(x)$ is the polynomial interpolator unity which takes the value unity $z \in D$ and zero at the other points in D . The quadrature weights are given by $w_z = \mathrm{E}_\xi(l_z(X)), z \in D$. From this we can derive necessary and sufficient condition for exact

quadrature of any polynomial $p(x)$ with respect to ξ. We might then call $\{p, D, \prec, \xi\}$ a "good" 4-tuple. The paper covers the following.

(i) Conditions for exact quadrature for moments. In particular all the moments $\mu_\alpha, \alpha \in L$ have exact quadrature. Classical quadrature is often expressed in terms of quadrature for moments. A key issue is how good the quadrature is for moments, when it is not exact.

(ii) Conditions for when the quadrature weights are positive, which need not always be the case.

(iii) Orthogonal polynomials. The relationship between the G-basis of the design and the orthogonal polynomials remains a hard problem but the present work goes some way in explanation.

(iv) Comparisons, using some special metrics, with some known methods of Gauss type and methods used in UQ such as Smolyak grids and in general simulation such as Sobol' sequences.

Keywords: Quadrature, design of experiments, Gröbner basis, orthogonal polynomial

Acknowledgements: The first author would like to acknowledge the UK EPSRC Basic Translation award, MUCM2, grant number EP/H007377/1.

References

1 Pistone, G., Riccomagno, E. and Wynn, H.P. (2001). Algebraic Statistics, volume 89 of Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, 2001.

2 Maruri-Aguilar, H and Wynn, H.P. (2012). The algebraic method in experimental design. To appear in Handbook of Design and Analysis of Experiments, ed. D. Bingham et al. Chapman and Hall/CRC, Boca Raton. Available at arXiv:1207.2968

Clustering of Longitudinal Data Based on Mixture of ELMM with Autoregressive Errors

ChangJiang Xu
McGill University
changjiang.xu@mail.mcgill.ca

Vicky Tagalakis, Celia M. T. Greenwood, Antonio Ciampi
McGill University
vicky.tagalakis@mcgill.ca, celia.greenwood@mcgill.ca,
antonio.ciampi@mcgill.ca

Identifying subtypes of an illness from the time evolution of an index is an important clinical objective, which may be met by appropriate statistical methods. We model the longitudinal data obtained from a heterogeneous population as samples from a mixture of Extended Linear Mixed Models (ELMM) with autoregressive errors (Pinheiro and Bates, 2000). Specifically, we consider the following mixed-effects model for individual i:

$$y_{it} = x'_{it}\beta + z'_{it}b_{it} + \epsilon_{it},$$

where β is a vector of fixed effects, b_{it} is a vector of random effects, and ϵ_{it} is an autoregressive error of order r:

$$\epsilon_{it} = \phi_1\epsilon_{i,t-1} + \ldots + \phi_{t-r}\epsilon_{i,t-r} + \eta_{it},$$

where η_{it} is an innovation, $t = 1, \ldots, m_i$. In matrix form, $L\epsilon = \eta$, where $\epsilon = (\epsilon_{i1}, \ldots, \epsilon_{i,m_i})'$, $\eta = (\eta_{i1}, \ldots, \eta_{i,m_i})'$, and

$$L = L(\phi, m_i) = \begin{bmatrix} 1 & & & & & \\ -\phi_1 & 1 & & & & \\ \vdots & \ddots & \ddots & & & \\ -\phi_r & \ldots & -\phi_1 & 1 & & \\ \vdots & \ddots & & \ddots & \ddots & \\ 0 & \ldots & -\phi_r & \ldots & -\phi_1 & 1 \end{bmatrix}.$$

We have developed a new algorithm for estimating the parameters of the mixture model described above. The algorithm is based on previous work (Ciampi et al. 2012) and on the modified Cholesky decomposition of the variance-covariance matrix (Pourahmadi 1999, McNicholas and Murphy 2010). We present an evaluation of the approach through limited simulations, and the analysis of two clinical data sets, one on warfarin initiation, the other on the time evolution of a delirium index.

Keywords: Clustering, mixture of Gaussian distribution, linear mixed-effects model, longitudinal data, time series.

References

Pinheiro, J. C., Bates, D. (2000): *Mixed-Effects Models in S and S-PLUS*, New York: Springer.

Ciampi A., Campbell H., Dyachenko A., Rich B., McCusker J., Cole M. G. (2012): Model-Based Clustering of Longitudinal Data: Application to Modeling Disease Course and Gene Expression Trajectories, *Communications in Statistics - Simulation and Computation*, 41:7, 992-1005

Pourahmadi M. (1999): Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika*, Vol. 86, N. 3, pp. 677-690.

McNicholas, P. D. and Murphy, T. B. (2010): Model-based clustering of longitudinal data, *The Canadian Journal of Statistics*, Vol. 38, N. 1, pp. 153-168.

An algorithm approach of constructing optimal/efficient crossover designs

Min Yang
University of Illinois - Chicago
myang2@uic.edu

Crossover designs are used in many clinical trials and other studies. Although there exist excellent publications on deriving optimal/efficient crossover designs, the existence of such optimal designs depends on the constraints of the combination of n (number of subjects), p (number of periods), and t (number of treatments), which may not be satisfied for most practical applications. In this talk, we shall propose an algorithm approach of constructing an optimal/efficient crossover design for any arbitrary combination of n, p and t. It can be demonstrated that the algorithm is fast and can be applied under variety of optimality criterion, regardless of parameters of interest.

Keywords: A-optimal; D-optimal; Balanced designs; Carryover effect.

Acknowledgements: This work was supported by NSF grants DMS-07-48409.

References

Yang, M. and Biedermann, S. (2012): On optimal designs for nonlinear models: a general and efficient algorithm, *under review*

Yang, M. and Stufken, J. (2008): Optimal and efficient crossover designs for comparing test treatments to a control treatment under various models, *Journal of Statistical Planning and Inference*, 138, 278-285, 2008

Limit distributions in branching random walks with finitely many centers of particle generation

Elena Yarovaya
Moscow State University
yarovaya@mech.math.msu.su

Previous models of continuous-time branching random walks (BRWs) on $\mathbf{Z}^d$ were studied for a medium with one source of branching (see, e.g., Bogachev and Yarovaya, 1998). The offspring reproduction law was defined by the intensities of a Markov birth-and-death process at the source, and the underlying random walk was usually assumed to be symmetric. One of the main problems in models of BRWs is a study of limit distributions for the number of the particles. The principal results of (Vatutin et. all 2003, Yarovaya 2010) were extended to BRWs containing a few sources of branching situated at arbitrary lattice points under the assumptions that underlying random walks may be nonsymmetric (Yarovaya, 2012). General methods were proposed to obtain conditions of an exponential growth for the number of particles in BRWs with several sources. It is known (Molchanov and Yarovaya, 2012), in the theory of BRWs the problem on the spectrum of a evolutionary operator of mean particle numbers plays an important role. Resolvent analysis of such operators (Cranston M. et al., 2009) has allowed to investigate BRWs with large deviation (Molchanov and Yarovaya, 2012). The limit theorems on asymptotic behavior of the Green's function for transition probabilities were established. The obtained results expand the previous studies in such direction as the structure of the population inside of the front and near to its boundary. A special attention was paid to the case when the spectrum of the evolutionary operator contains only one positive isolated eigenvalue.

Keywords: Branching random walks, Green's functions, large deviations.

Acknowledgements: This work was supported by RFBR grant 13-01-00653.

References

Bogachev L., Yarovaya E. (1998): Moment analysis of a branching random walk on a lattice with a single source. *Doklady Akademii Nauk*, Vol. 363, pp. 439–442.

Vatutin V., Topchii V., Yarovaya E. (2003): Catalytic branching random walk and queueing systems with random number of independent servers. *Teoriya Jmovirnostej ta Matematichna Statistika* Vol. 69, pp. 158–172.

Yarovaya E. (2010): Criteria of Exponential Growth for the Numbers of Particles in Branching Random Walks, *Teor. Veroyatn. Primen.*, Vol. 55, N. 4, pp. 705–731.

Yarovaya E. (2012): pectral properties of evolutionary operators in branching random walk models, *Mathematical Notes*, Vol. 92, N. 1, pp. 115–131.

Cranston M., Koralov L., Molchanov S., Vainberg B. (2009): Continuous model for homopolymers, *J. Funct. Anal.*, Vol. 256, pp. 2656–2696.

Molchanov S., Yarovaya E. (2012): Branching Processes with Lattice Spatial Dynamics and a Finite Set of Particle Generation Centers, *Doklady Mathematics*, Vol. 86, N. 2, pp. 638–641

Molchanov S., Yarovaya E. (2012): Population Structure inside the Propagation Front of a Branching Random Walk with Finitely Many Centers of Particle Generation, *Doklady Mathematics*, Vol. 86, N. 3, pp. 787–790.

Convergence of adaptive allocation procedures

Maroussa Zagoraiou
Department of Economics, Statistics and Finance, University of Calabria
maroussa.zagoraiou@unical.it

Alessandro Baldi Antognini
Department of Statistical Sciences, University of Bologna
a.baldi@unibo.it

In this talk we provide a general convergence result for adaptive designs for treatment comparison, both in the absence and presence of covariates. By combining the concept of downcrossing and stopping times of stochastic processes, we demonstrate the almost sure convergence of the treatment allocation proportion for a vast class of adaptive procedures, even in the absence of a prefixed target, also including designs that have not been formally investigated but mainly explored through simulations, such as Atkinson's optimum biased coin design (Atkinson, 1982) and Pocock and Simon's minimization method (Pocock and Simon, 1975). Although the large majority of the proposals are based on continuous allocation rules, updated step by step on the basis of the current allocation proportion and some estimates of the unknown parameters, the recent literature tends to concentrate on discontinuous randomization functions because of their low variability. Our results allow to prove via a unique mathematical framework the convergence of adaptive allocation methods based on both continuous and discontinuous randomization functions, like for instance the Reinforced Doubly adaptive Biased Coin Design (Baldi Antognini and Zagoraiou, 2012), the Efficient Randomized-Adaptive Design (Hu et al., 2009) and Hu and Hu's Covariate-Adaptive rule (Hu and Hu, 2012). Our approach takes also into account designs based on Markov chain structures, such as the Adjustable Biased Coin Design (Baldi Antognini and Giovagnoli,

2004) and the Covariate-Adaptive Biased Coin Design (Baldi Antognini and Zagoraiou, 2011), that can be characterized by sequences of allocation rules. Moreover, by removing some unessential conditions usually assumed in the literature, our approach provides suitable extensions of several existing procedures.

Keywords: Biased Coin Design, CARA Procedures, Minimization methods, Response-Adaptive Designs, Sequential Allocations.

References

Atkinson A.C. (1982): Optimum biased coin designs for sequential clinical trials with prognostic factors, *Biometrika*, Vol. 69, pp. 61-67.

Baldi Antognini A., Giovagnoli A. (2004): A new "biased coin design" for the sequential allocation of two treatments, *Journal of the Royal Statistical Society C*, Vol. 53, pp. 651-664.

Baldi Antognini A., Zagoraiou M. (2011): The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors, *Biometrika*, Vol. 98, pp. 519-535.

Baldi Antognini A., Zagoraiou M. (2012): Multi-objective optimal designs in comparative clinical trials with covariates: the reinforced doubly adaptive biased coin design, *The Annals of Statistics*, Vol. 40, pp. 1315-1345.

Hu F., Zhang L. X., He X. (2009): Efficient randomized adaptive designs, *The Annals of Statistics*, Vol. 37, pp. 2543-2560.

Hu Y., Hu F. (2012): Asymptotic properties of covariate-adaptive randomization, *The Annals of Statistics*, Vol. 40, pp. 1794-1815.

Pocock S.J., Simon R. (1975): Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial, *Biometrics*, Vol. 31, pp. 103-115.

Response surface prediction from a spatial montoring process

Diego Zappa
Department of statistical sciences
diego.zappa@unicatt.it

Riccardo Borgoni, Luigi Radaelli
Department of economics, quantitative methods and managment sciences
riccardo.borgoni@unimib.it, luigi_radaelli@libero.it

In semiconductor processes devices (chipsets) are developed over the surface of a circular-shaped working substrate called wafer by a well-integrated sequence of several steps, called ”technological steps”; each step is studied by data acquisition in the development phase. This is usually done by collecting data over an assigned grid. In the production phase, the number of wafers to be measured/monitored becomes enormous. Data collected in this phase are generally used to check whether both target values are matched and homogeneity exists over the whole production surface. Since the collection of measures is highly time consuming and very expensive it is necessary to ascertain whether the number of sampling points selected in the development phase could be reduced by keeping at an acceptable level the degree of representativeness of the wafer surface and the predictability of the response surface. Since the sampling points of the reduced grid must be a subset of theoriginal monitoring grid they cannot be simply allocated by some experimental design. However, the number of possible configuration of sampling locations is huge even when the size of the starting grid is moderately large rising a formidable combinatorial problem. Borgoni et al. (2012) proposed a simulated annealing (SA) algorithm combined with a geostatistical model to select the sub map. The strength of their proposal is that it is fully nonparametric, data driven and naturally selects those

regions that are the most effective in predicting the response variable. As a consequence the resulting map may not be regular over the surface and can return maps that are concentrated in subregions. In this paper we present a different approach that, even without the availability of starting experimental data, selects a sub-grid according to the criterion of spatial optimal coverage of the wafer surface (see also Walvoort, 2010). This approach may also include expert knowledge about those areas where production is less precise because of unavoidable technical reasons and hence may indicate where a higher sampling density must be assured. If sampling measures are available, a validation procedure can be used to select the best sub-map based for instance on the prediction error, by comparing the results obtained using the full and the reduced grid.

Keywords: Statistics in microelctronics, subgrid selection, response surface

References

Borgoni R., Radaelli L., Tritto V., Zappa D. (2013): Optimal Reduction of a Spatial Monitoring Grid: Proposals and Applications in Process Control, *Computational Statistics & Data Analysis* , Vol. 58, N. 4, pp. 407-419.

Walvoort, D. J. J., Brus, D. J. and de Gruijter, J. J. (2010): An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means, *An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means*, Vol. 36, pp. 1261-1267.

The Study of the Laplace Transform of Marshall-Olkin Multivariate Exponential Distribution

Igor V. Zolotukhin
Russian Academy of Sciences, Institute of Oceanology, St.Petersburg Department
igor.zolotukhin@gmail.com

Let us consider a class of random vectors $\boldsymbol{Z}$ with the Marshall-Olkin multivariate exponential distribution (**MVE**) with reliability function

$$\bar{F}(z_1,\, z_2,\, ...,\, z_k) = \exp\left[-\sum_{\varepsilon\in\mathcal{E}} \lambda_\varepsilon \max_{1\le i\le k}\{\varepsilon_i z_i\}\right],\ z_i \ge 0;$$

here $\lambda_\varepsilon \ge 0$, $\mathcal{E} = \{\varepsilon\}$ is a set of $k-$dimensional indices $\varepsilon = (\varepsilon_1,\, \ldots,\, \varepsilon_k)$, and each component of ε_i is 0 or 1. Vector ε will be used for the coordinate hyperplane selection in k-dimensional space.

Let us introduce the following notation: $\mathbf{1} = (1,\, ...,\, 1)$; (ε, s) is the scalar product of vectors ε and s; $\varepsilon\, s$ is their coordinate-wise product. The sign "$\bullet$" denotes summation on some coordinate, for example:

$\lambda_{\bullet\ldots\bullet} = \sum\limits_{\varepsilon\in\mathcal{E}} \lambda_\varepsilon,\ \lambda_{1\bullet\bullet} = \lambda_{11\bullet} + \lambda_{10\bullet} = \lambda_{111} + \lambda_{110} + \lambda_{101} + \lambda_{100}.$

Theorem 1.

For any $\boldsymbol{Z} \in MVE(\lambda_\varepsilon,\, \varepsilon \in \mathcal{E})$ Laplace transform of its distribution is given by

$$\psi(s) = \mathsf{E}\, e^{-s\boldsymbol{Z}} = \frac{1}{(\mathbf{1},\, s) + \lambda_{\bullet\ldots\bullet}} \sum_{\varepsilon\in\mathcal{E}} \lambda_{\bar{\varepsilon}}\, \psi(\varepsilon s).$$

Remark. Let the random variable X has an exponential distribution with parameter $\lambda_{\bullet\ldots\bullet}$, and vector $\boldsymbol{X} = (X,\, ...,\, X)$. Laplace transform of such vector $\boldsymbol{X}$ is

$$\Psi(s) = \mathsf{E}\, e^{-s\boldsymbol{X}} = \frac{\lambda_{\bullet\ldots\bullet}}{(\mathbf{1},\, s) + \lambda_{\bullet\ldots\bullet}}.$$

Assuming that $p_\varepsilon = \dfrac{\lambda_\varepsilon}{\lambda_{\bullet\ldots\bullet}}$, this formula can be rewritten as

$$\psi(s) = \Psi(s) \sum_{\varepsilon \in \mathcal{E}} p_\varepsilon \, \psi(\bar{\varepsilon} s).$$

Hence a **MVE** distribution is the discrete mixture of the distribution $\boldsymbol{X}$ and its convolutions with projections of this **MVE** distribution on all its coordinate hyperplanes.

Now let's set the partial order relation in the set $\mathcal{E}$: $\forall \varepsilon, \delta \in \mathcal{E}$ $\delta \leq \varepsilon$, if for all i $\delta_i \leq \varepsilon_i$; $\delta < \varepsilon$ if $\delta \leq \varepsilon$ and $\delta \neq \varepsilon$.

Let take in consideration also the vectors ε, whose coordinates can take three values: 0, 1, $\bullet$. For these vectors we define the $\bar{\varepsilon}$ as follows:

$\bar{\varepsilon}_j = 0$ if $\varepsilon_j = 1$; $\bar{\varepsilon}_j = 1$ if $\varepsilon_j = 0$; $\bar{\varepsilon}_j = \bullet$ if $\varepsilon_j = \bullet$.

The vector $\varepsilon \oplus \delta$ is a vector with coordinates determined by the following rule: $1 \oplus 0 = 1$; $0 \oplus 1 = 1$; $1 \oplus 1 = 0$; $0 \oplus 0 = \bullet$.

Theorem 2.

$$\forall\, \varepsilon \in \mathcal{E} \quad \psi(\varepsilon s) = \frac{1}{\sum\limits_{\delta < \varepsilon} \lambda_{\delta \oplus \varepsilon} + (\varepsilon,\ s)} \sum_{\delta < \varepsilon} \lambda_{\delta \oplus \varepsilon} \psi(\delta s).$$

Hence the Laplace transform of the vector $\boldsymbol{Z}$ projection on the coordinate hyperplane ε can be found by **Theorem 1**, but to do this we need to replace the zeros on the "bullets" in the indices of all parameters λ. Furthermore, the distribution $\boldsymbol{Z}$ can be uniquely determined by its one-dimensional marginal distributions.

Keywords: multivariate exponential distribution, Laplace transform.

References

Marshall A.W., Olkin I. (1967): A multivariate exponential distribution, *J. Amer. Statist. Assoc.*, vol. 62, pp. 30-44.