Contents

Bivariate Lifetime Data Analysis Using Block and Basu Bivariate Distribution in Presence of Cure Fraction

NYA
Not Yet
Arranged

JORGE ALBERTO ACHCAR[*,§], EMÍLIO AUGUSTO COELHO-BARROS[*,†], JOSMAR MAZUCHELI[‡]

[*]University of São Paulo, Medical School, Ribeirão Preto, SP, Brazil,
[†]Federal Technological University of Paraná, Cornélio Procópio, PR, Brazil,
[‡]University of Maringá, Maringá, PR, Brazil
[§]email: achcar@fmrp.usp.br

1:JorgeAlbertoAchcar1.tex,session:NYA

In some areas of application, especially in medical and engineering studies, we could have two lifetimes T_1 and T_2 associated with each unit. Usually, these data are assumed to be independent, but in many cases, the lifetime of one component could affect the lifetime of the other component. This is the case as an example in the medical area of paired organs like kidneys, lungs, eyes, ears, dental implants among many others. In the literature we observe many papers related to bivariate lifetime parametric models. One of these bivariate lifetime distributions is very popular: the bivariate exponential distribution introduced by Block and Basu (1974). This is one of the most flexible bivariate exponential distributions and it was derived by Block and Basu by omitting the singular part of Marshall and Olkin distribution [4]. We develop a Bayesian analysis for bivariate lifetime data considering a generalization of the bivariate exponential distribution of Block and Basu in the presence of censored data, covariates and cure fraction. Posterior summaries of interest are obtained using standard MCMC methods as the popular Gibbs Sampling algorithm [2] or the Metropolis-Hastings algorithm [1]. Posterior summaries of interest for each model are simulated using standard MCMC methods. We also have used the *rjags* package [5] for *R* software [6]. Just Another Gibbs Sampler (JAGS) is a program for analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulation not wholly unlike BUGS. In this application, we consider a data set [3] with 197 patients were a 50% random sample of the patients with "high-risk" diabetic retinopathy as defined by the Diabetic Retinopathy Study (DRS).

References

[1] Chib, S.; Greenberg, E. *Undestanding the metropolis-hastings algorithm*, The American Statistician 49 (1995), pp. 327–335.

[2] Gelfand, A. E.; Smith, A. F. M. *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association 85 (1990), pp. 398–409.

[3] Huster, W. J.; Brookmeyer, R.; Self, S. G. *Modelling paired survival data with covariates*, Biometrics 45 (1989), pp. 145–156.

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

[5] Plummer, M. (2011), r package version 3-3.

[6] R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria (2011), ISBN 3-900051-07-0.

CS13B Envtl. & Biol. Stat.

Correlated Pharmacokinetic Bioequivalence: A Bayesian approach assuming multivariate Student t-distribution

ROBERTO MOLINA DE SOUZA[*,†], JORGE ALBERTO ACHCAR[*,§], JOSMAR MAZUCHELI[‡]

[*]University of São Paulo, Medical School, Ribeirão Preto, SP, Brazil,
[†]Federal Technological University of Paraná, Cornélio Procópio, PR, Brazil,
[‡]University of Maringá, Maringá, PR, Brazil
[§]email: achcar@fmrp.usp.br 2:JorgeAlbertoAchcar.tex,session:CS13B

In this paper, we introduce a Bayesian analysis for correlated pharmacokinetic parameters assuming a multivariate Student t-distribution. This modeling is compared with usual bioequivalence models assuming normal distributions for the error terms of the model and also for the random effects usually considered to model pharmacokinetic parameters. The Bayesian analysis is developed using standard MCMC (Markov Chain Monte Carlo) methods. The use of a heavy tail distribution given by the Student t-distribution could be a good alternative to analyse pharmacokinetic parameters especially in the presence of outlier observations. To assess the biological equivalence of two formulations (see for example, Chow et al., 2003), that is, if two formulations are bioequivalent, the plasma concentration data are used to assess key pharmacokinetics parameters such as the area under the curve, AUC and peak concentration, C_{max}. Also it is possible to consider the time where occurs the maximum concentration, T_{max}. The bioequivalence study usually is performed assuming independent normal distributions (see for example, Ghosh and Gönen, 2008) for the bioequivalence measures, usually in the logarithmic scale with unknown variances. Since the Student t-distribution has more heavy tails which can give more robustness and better fit for these discordant measures, we assume the multivariate Student t-distribution to jointly analyse the pharmacokinetic parameters AUC and C_{max}, as an alternative for the multivariate normal distribution [Souza et al. (2009)].

References

[Chow et al. (2003)] Chow, S.C., Shao, J., Wang, H., 2003: In vitro bioequivalence testing, *Statistics in Medicine*, **22**, 55 - 98.

[Ghosh and Gönen (2008)] Ghosh, P., Gönen, M., 2008: Bayesian modeling of multivariate average bioequivalence, *Statistics in Medicine*, **27**, 2402 - 2419.

[Souza et al. (2009)] Souza, R.M., Achcar, J.A., Martinez, E.Z., 2009: Use of Bayesian methods for multivariate bioequivalence measures, *Journal of Biopharmaceutical Statistics*, **19**, 42 - 66.

OCS9 Design of Experiments

Bayesian Optimal Design of Experiments for Prediction of Correlated Processes

MARIA ADAMOU[*,†], SUSAN LEWIS[*], SUJIT SAHU[*], DAVID WOODS[*]

[*]Southampton Statistical Sciences Research Institute, University of Southampton, UK
[†]email: ma5g10@southampton.ac.uk 3:Maria_Adamou.tex,session:OCS9

Data collected from correlated processes arise in many diverse application areas including studies in environmental and ecological science and in both real and computer experiments. Often, such data are used for predicting the process at unobserved points in some continuous region of interest. The design of the experiment from which the data are collected may strongly influence the quality of the model fit and hence the accuracy of subsequent predictions. The objective of this work is to obtain

an optimal choice of design to allow precise prediction of the response at unobserved points using an appropriate statistical model.

We consider Gaussian process models that are typically defined by a spatial correlation structure that may depend upon unknown parameters. This correlation structure may affect the choice of design points, and must be taken into account when choosing a design.

We illustrate a new approach for the selection of Bayesian optimal designs using a decision theoretic approach and an appropriate loss function. To avoid the computational burden usually associated with Bayesian design, we have developed a new closed form approximation that allows fast approximation of the objective function. The resulting designs are illustrated through a number of examples using applications from both spatial statistics and computer experiments.

Topological Inference: A Challenge for Probability and Statistics

SIL Spec. Invited Lecture

ROBERT J. ADLER*,†

*Technion – Israel Institute of Technology, Haifa, Israel
†email: robert@ee.technion.ac.il 4:RobertAdler.tex,session:SIL

The last few years have seen an exciting development in what is reputedly one of the most esoteric areas of pure mathematics: Algebraic Topology. A small but rapidly growing group of dedicated mathematicians is actually trying to apply it to real world problems, and, as a result, 'Applied Algebraic Topology' is no longer an oxymoron. Parts of this project are not totally new. For example, the brain imaging community has been using random fields and topology for quite some time, leading to the notion of 'topological inference'. However, what is completely new is the mathematical sophistication of the techniques now being applied to areas as widespread as data mining, dimension reduction, and manifold learning, all topics familiar to statisticians, as well as to areas classically outside of Statistics.

In the talk I shall describe some of the new ideas that have arisen in Applied Algebraic Topology, and discuss the challenges they raise for Statistics and Probability. I will give examples via easy to understand (but not so easy to prove) results about the persistent homologies of random fields and random complexes, describing both their theory and applications.

The main aim of this lecture will be to convince statisticians and probabilists that there is an entire new area crying out for probabilistic modelling and statistical analysis, and to maybe motivate some of them to participate in solving the exciting challenges it is already providing.

Branching Processes with Immigration in Random Environment

IS3 Branching Proc.

VALERIY I. AFANASYEV*,†

*Steklov Institute, Moscow, Russia
†email: viafan@mail.ru 5:ValeriyAfanasyev.tex,session:IS3

Let (p_i, q_i), $i \in \mathbb{N}$, be a sequence of independent and identically distributed random vectors such that $p_1 + q_1 = 1$, $p_1 > 0$, $q_1 > 0$. It is *the main assumption* that $\mathbf{E} \ln (q_1/p_1) = 0$, $\mathbf{E} \ln^2 (q_1/p_1) = \sigma^2$, $0 < \sigma^2 < \infty$.

Let $\{Z_i, i \in \mathbb{N}_0\}$ be a *branching process in random environment* with one immigrant in each generation. The reproduction distribution of a representative (or an immigrant) of i-th generation has the view $\{p_{i+1} q_{i+1}^k, k \in \mathbb{N}_0\}$.

Suppose that $S_0 = 0$, $S_n = \sum_{i=1}^{n} \varkappa_i$, $n \in \mathbb{N}$, where $\varkappa_i = \ln(q_i/p_i)$, $i \in \mathbb{N}$. Suppose also that $a_n = \exp(-S_n)$, $b_n = \sum_{i=0}^{n-1} a_i$, $n \in \mathbb{N}$. Introduce for $n \in \mathbb{N}$ a random process Y_n: $Y_n(0) = 0$, $Y_n(t) = a_{\lfloor nt \rfloor} Z_{\lfloor nt \rfloor} / b_{\lfloor nt \rfloor}$ for $t > 0$.

Let W be a Brownian motion and $\{\gamma_n, n \in \mathbb{N}\}$ be a sequence of independent random variables distributed by exponential law with parameter 1 (W and $\{\gamma_n, n \in \mathbb{N}\}$ are independent). Suppose that $L(t) = \inf_{s \in [0,t]} W(s)$, $t \geq 0$.

Describe finite-dimensional distributions of a random process $Y = \{Y(t), t \geq 0\}$. The random vector $(Y(t_1), \ldots, Y(t_m))$ for $0 < t_1 < t_2 < \ldots < t_m$, $m \in \mathbb{N}$, has the same distribution as the following random vector $\left(\widehat{Y}_1, \ldots, \widehat{Y}_m\right)$: $\widehat{Y}_1 = \gamma_1$ and, if $\widehat{Y}_k = \gamma_i$ for $k < m$, $i \in \mathbb{N}$ and $L(t_{k+1}) = L(t_k)$, then $\widehat{Y}_{k+1} = \widehat{Y}_k$, but if $L(t_{k+1}) < L(t_k)$, then $\widehat{Y}_{k+1} = \gamma_{i+1}$. The following theorem was proved in [Afanasyev (2012)].

Theorem 1. *If the main assumption is valid then, as* $n \to \infty$,

$$Y_n \Rightarrow Y,$$

where the symbol $\Rightarrow$ *means convergence of finite-dimensional distributions.*

Let W^* be an *inverse Brownian meander* W^* (this is a Brownian motion on $[0,1]$ conditioned to attain its minimum at moment 1) and $\{\gamma_n, n \in \mathbb{N}\}$ be a sequence of independent random variables distributed by exponential law with parameter 1 (W^* and $\{\gamma_n, n \in \mathbb{N}\}$ are independent). Suppose that $L^*(t) = \inf_{s \in [0,t]} W^*(s)$, $t \in [0,1]$. A process $Y^* = \{Y^*(t), t \in [0,1]\}$ is defined similarly the process Y with replacement $\{L(t), t \in [0,1]\}$ for $\{L^*(t), t \in [0,1]\}$.

Theorem 2. *If the main assumption is valid then, as* $n \to \infty$,

$$\{Y_n(t), t \in [0,1] \mid Y_n(1) = 0\} \Rightarrow Y^*.$$

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

References

[Afanasyev V.I. (2012)] Afanasyev, V.I., 2012: On the time of attaining a high level by a random walk in random environment, *Theory Probab. Appl.*, **57**, 625 - 648.

CS36A Graphical Methods

Deterministic Modelling of Gene Network via Parametric and Nonparametric Approaches

MELİH AĞLAZ*, BURAK KILIÇ*, VILDA PURUTÇUOĞLU*,†

*Department of Statistics, Middle East Technical University, Ankara, Turkey
†email: vpurutcu@metu.edu.tr 6:MelihAgraz.tex,session:CS36A

The major challenge in system biology is to understand the behaviour of actual biological networks which have many genes and dense interactions. In order to construct such complex systems and make inference about the model parameters, certain levels of summaries are needed to be done so that the final network can, at least, explain the main description of the actual biological events. There are two major approaches, namely, deterministic and stochastic methods, to model the complex systems. The former can describe the random nature of the system, via the number of molecules in the environment whereas, the latter can explain the network under stead-state conditions. In this study we investigate the possible modelling of moderately complex system via deterministic approaches under the assumption that the genes are measured at different levels of concentrations through times.

Thereby in deterministic models of the dynamic data, we consider nonparametric and parametric alternatives. In nonparametric approach, we apply MARS (Multivariate Adaptive Regression Splines) methods that are based on the nonlinear regression modelling via splines. On the other hand among parametric approaches, we implement GGM (Gaussian Graphical Models), Bernstein polynomial and Szász Mirakyan polynomial that depend on normal, binomial, and poisson distribution, respectively. In application we perform simulated datasets which have small and moderate dimensions. Finally in order to evaluate the estimated networks from each approach and under different network sizes, we select certain model selection criteria such as false positive rate, recall, and accuracy measures, and compare the results with the true network. We believe that the outcomes of this study can be helpful also for inference of large networks and real data application that we aim to extent in future studies.

Acknowledgment. The author would like to thank the EU FP-7 Collaborative Project (Project No: 26042) for their support.

References

[1] Copyright StatSoft, Inc., 2013: Multivariate Adaptive Regression Splines, http://www.statsoft.com/textbook/stmars.html (accessed 15 March2013).

[2] Nakont, D., Neha, B., 2010: Some Approximations Theorems for Multivariate Bernstein Operators, *Southeast Asian Bulletin of Math.*, **34**, 1023-1034.

[3] Purutçuoğlu, V., Erdem, T., and Weber, G.W., (2011): Inference of the JAK-STAT Gene Network via Graphical Models, *Proceeding of the 23rd International Conference on Systems Research, Informatics and Cybernetics*, Baden, Germany, 46-50.

Parameter Estimation for EEG Distribution based on Progressively Censored Sample

OCS27 Infer. Censored Sample

AHMET ÇALIK[*,†], ILKAY ALTINDAĞ[*], COŞKUN KUŞ[*], YUNUS AKDOĞAN[*], ISMAIL KINACI[*]

[*]Selcuk University, Konya, Turkey
[†]email: ahmetcalik@selcuk.edu.tr 7:AhmetCalik.tex,session:OCS27

In this paper, we consider the parameter estimation for extension exponential-geometric (EEG) distribution based on progressive Type-II censored sample. The maximum likelihood estimates (MLEs) are obtained. Bayes estimates of parameters are also discussed by using Tierney-Kadane approximation under quadratic loss function. The approximate confidence intervals based on Fisher information matrix are obtained. The performance of MLEs and Bayes estimates are investigated in that their estimated risks by simulation study for different parameter values and censoring schemes. Simulation study is also perfomed to get coverage probabilities of approximate confidence intervals. Finally, a numerical example is given to illustrate the methodology.

Acknowledgment. This research was partially supported by Selcuk University BAP Office.

POSTER
Poster

Generalized Isotonized Mean Estimators for Judgement Post-stratification with Multiple Rankers

JOHAN LIM*, SOOHYUN AHN*,§, XINLEI WANG†, MIN CHEN‡

*Department of Statistics, Seoul National University, Seoul, Korea,
†Department of Statistical Science, Southern Methodist University, Dallas, Texas, U.S.A.,
‡Department of Clinical Science, U.T. Southwestern Medical Center, Dallas, Texas, U.S.A.
§email: puppy0603@snu.ac.kr 8:shahn203.tex,session:POSTER

This paper proposes a new set of estimators of population mean from JPS with multiple rankers. Suppose Y is the variable of interest that is absolutely continuous with population mean μ and finite variance σ^2. The JPS data with multiple rankers are constructed as follows. First, select a simple random sample of n units, on each of which the value of Y is measured. For each unit i ($1 \le i \le n$), an additional $H-1$ units are randomly selected and the judgment order of Y_i among its H comparison units by the k-th ranker, denoted by $O_{i,k}$, is determined subjectively without measuring the $H-1$ units. Thus, the JPS sample with m multiple rankers consists of data of the form $\{Y_i, O_i\}_{i=1}^n$ with $O_i = (O_{i,1}, \ldots, O_{i,m})$, and the n measured units fall into H^m post-strata formed by the orders. Let $Y_{[\mathbf{r}]}$ denote $Y|O = \mathbf{r}$, any observation falling in the $\mathbf{r}-$th post-stratum, where $\mathbf{r} = (r_1, \ldots, r_m)$ and each $r_k \in \{1, \cdots H\}$. Let $n_{\mathbf{r}}$, $\overline{Y}_{[\mathbf{r}]}$, $\mu_{[\mathbf{r}]}$, and $\sigma^2_{[\mathbf{r}]}$ denote the number, sample mean, mean and variance of the observations in the $\mathbf{r}$-th stratum.

Three nonparametric mean estimators have been proposed for JPS with multiple rankers in the literature, the estimator by MacEachern et al.(2004), the raking estimator and the best linear unbiased estimator by Stokes et al.(2007). All three estimators can be written as

$$\widehat{\mu} = \sum_{\mathbf{r}} \widehat{\pi}_{\mathbf{r}} \overline{Y}_{[\mathbf{r}]}, \tag{1}$$

for different estimators of $\widehat{\pi}$, where $\overline{Y}_{[\mathbf{r}]}$ is the sample mean of the $\mathbf{r}$-th stratum. In this paper, we propose a new set of mean estimators by extending the Wang et al.(2008) to JPS data with the multiple rankers. Wang et al.(2008) considered the monotonicity among means of judgement strata and proposed an isotonized mean estimator. In addition, they numerically showed that the isotonized mean estimator performed better than the original JPS estimator. For multiple ranker problem, we consider the matrix partial order among $\mu_{[\mathbf{r}]}$s. To be specific, we assume that $F_{[\mathbf{r}]}$s, the stratum CDFs, are stochastically ordered in the sense that, if $\mathbf{r}_1 \le \mathbf{r}_2$ in a component-wise manner, then, for any $y \in \mathcal{R}$, $F_{[\mathbf{r}_1]}(y) \ge F_{[\mathbf{r}_2]}(y)$. The stochastic order in stratum CDFs introduces the built-in ordering in the means of the strata; that is, if $\mathbf{r}_1 \le \mathbf{r}_2$, then $\mu_{[\mathbf{r}_1]} \le \mu_{[\mathbf{r}_2]}$. However, the sample means $\overline{Y}_{[\mathbf{r}]}$s, the most common empirical estimator of $\mu_{[\mathbf{r}]}$, often do not satisfy the order constraints. Therefore, we propose to modify the sample means $\overline{Y}_{[\mathbf{r}]}$s by solving the following generalized isotonic regression problem:

$$\begin{array}{ll} \text{minimize} & \sum_{\mathbf{r}} n_{\mathbf{r}} \big(\overline{Y}_{[\mathbf{r}]} - \mu_{[\mathbf{r}]}\big)^2 \\ \text{subject to} & \mu_{[\mathbf{r}_1]} \le \mu_{[\mathbf{r}_2]}, \quad \text{if} \quad \mathbf{r}_1 \le \mathbf{r}_2. \end{array} \tag{2}$$

To solve the algorithm, we use the pooled-adjacent-vaiolator algorithm iteratively. We finally suggest a set of new estimators

$$\widehat{\mu}^* = \sum_{\mathbf{r}} \widehat{\pi}_{\mathbf{r}} \widehat{\mu}^*_{[\mathbf{r}]}, \tag{3}$$

where $\widehat{\mu}^*_{[\mathbf{r}]}$s are the solution to the above optimization problem.

A Statistical Evaluation of Competing Mosquito Control Methods

POSTER
Poster

OSHO O. AJAYI*

*American University of Nigeria, Yola, Nigeria
†email: osho.ajayi@aun.edu.ng 9:OshoAJAYI.tex,session:POSTER

Mosquito investation and management has a number of social and economic implications in all societies. Because poorly managed single implement control strategies leading to an outbreak can be devastating, attempts are usually made to use more than one control methods regardless of environmental cost consideration. This consideration, among others, suggests the need for a quantitative evaluation of all the methods in use at a particular time so as to determine the benefit(s) which may be obtained from both the individual agents as well that of all inputs combined. In this work, we attempt to model the plausible benefit derivable from the application of a number of (single) inputs independently, as well as possible relative values whenever there are more than one inputs in play. The work is applied to a real dataset generated from a population with a good level of heterogeneity.

Network Completion Approach for Inference of Genetic Networks

OCS7
Comp.
Biology

TATSUYA AKUTSU*,†

*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan
†email: takutsu@kuicr.kyoto-u.ac.jp 10:Akutsu.tex,session:OCS7

Extensive studies have been done in these 10 years on inference of genetic networks using gene expression time series data, and a number of computational methods have been proposed. However, there is not yet an established or standard method for inference of genetic networks. One of the possible reasons for the difficulty of inference is that the amount of available high-quality gene expression time series data is still not enough, and hence it is intrinsically difficult to infer the correct or nearly correct network from such a small amount of data. Therefore, it is reasonable to try to develop another approach. For that purpose, we proposed an approach called *network completion* [1] by following Occam's razor, which is a well-known principle in scientific discovery. Network completion is, given an initial network and an observed dataset, to modify the network by the minimum amount of modifications so that the resulting network is (most) consistent with the observed data.

Using a Boolean model, we proved some computational complexity results in [1]. Since these result were not practical, we proposed a practical method named DPLSQ based on least-squares fitting and dynamic programming [2], where the former technique is used for estimating parameters in differential equations and the latter one is used for minimizing the sum of least-squares errors by integrating partial fitting results on individual genes under the constraint that the numbers of added and deleted edges must be equal to the specified ones. One of the important features of DPLSQ is that it can find an optimal solution (i.e., minimum squared sum) in polynomial time if the maximum indegree of a network is bounded by a constant. Although DPLSQ does not automatically find the minimum modification, it can be found by examining varying numbers of added/deleted edges, where the total number of such combinations is polynomially bounded. If a null network (i.e., a network having no edges) is given as an initial network, DPLSQ can work as a usual inference method for genetic networks.

Very recently, we extended DPLSQ for completion of time varying networks such that it can identify the time points at which the structure of gene regulatory network changes [3]. For this extension, a novel double dynamic programming method was developed in which the inner loop is used to identify static network structures and the outer loop is used to determine change points.

Both DPLSQ and its extension were shown to be effective via applications to completion of genetic networks by using artificially generated time series data and real gene expression time series data.

References

[1] Akutsu, T., Tamura, T., Horimoto, K., 2009: Completing networks using observed data, *Proc. 20th International Conference on Algorithmic Learning Theory*, 126 - 140.

[2] Nakajima, N., Tamura, T., Yamanishi, Y., Horimoto, K., Akutsu, T., 2012: Network completion using dynamic programming and least-squares fitting, *The Scientific World Journal*, **2012**, 957620.

[3] Nakajima, N., Akutsu, T., 2013: Network completion for time varying genetic networks, *Proc. 6th International Workshop on Intelligent Informatics in Biology and Medicine*, in press.

OCS2 space-time modeling

Spatio-Temporal Analysis for Bird Migration Phenology

ALI ARAB*, JASON COURTER†

*Georgetown University, †Taylor University

email: aa577@georgetown.edu 11:Ali_Arab.tex,session:OCS2

The study of changes in the patterns of bird arrival data is an important problem in phenology as the bird migratory processes can serve as potential bioindicators of climate change. To this end, several recent studies have analyzed trends of ecological processes such as bird migration patterns and provided evidence that pattern changes in these processes are correlated with climate effects. However, most studies ignore spatial and temporal variability in migration data and this often results in loss of important information across spatial and temporal scales. Critically, as the impact of climate change on natural processes may not be consistent over time and space, spatio-temporal analysis of bird migration processes allows climate scientists to better understand changes and shifts in migration patterns. We discuss a spatio-temporal modeling framework for analyzing birds spring arrival data which also allows for investigating potential links to climate indices. To demonstrate the methodology, we use two citizen science databases for Purple Martins; historic bird arrival data (1905-1940) from the North American Bird Phenology Program (BPP), and data for recent years (2001-2010) from the Purple Martin Conservation Association. Our results show significantly earlier spring arrivals for Purple Martins in most of the U.S. (South, Midwest, Mid-Atlantic, and North West) with link to Winter North Atlantic Oscillation. Finally, as most long-term phenological studies (including ours) heavily rely on data obtained from citizen science programs, we discuss potential sources of bias in these data sets and in general, sampling issues related to citizen science programs. Potential improvements will be discussed in order to obtain high quality data from these efforts.

IS6 Financial Time Ser.

Shortest-weight Paths in Random Graphs

HAMED AMINI

EPFL, Lausanne, Switzerland

email: hamed.amini@epfl.ch 12:HamedAmini.tex,session:IS6

Consider a random regular graph with degree d and of size n. Assign to each edge an i.i.d. exponential random variable with mean one. In the first part, we establish a precise asymptotic expression for the maximum number of edges on the shortest-weight paths between a fixed vertex and all the other vertices, as well as between any pair of vertices. This is a joint work with Yuval Peres [1]. We then analyze the impact of the edge weights on distances in sparse random graphs. Our main result consists of a precise asymptotic expression for the weighted diameter and weighted flooding time when the edge weights are i.i.d. exponential random variables. This is based on a joint work with Moez Draief and Marc Lelarge [2].

References

[1] H. Amini and Y. Peres. Shortest-weight paths in random regular graphs. arXiv: 1210.2657, 2012.

[2] H. Amini, M. Draief, and M. Lelarge. Flooding in weighted sparse random graphs. *SIAM Journal on Discrete Mathematics*, **27(1)**, 1 - 26, 2013.

Finite Sample Analysis of Maximum Likelihood Estimators and Convergence of the Alternating Procedure

CS37A Estim. Methods

ANDREAS ANDRESEN*,†, VLADIMIR SPOKOINY*

*Weierstraß Insitiute, Berlin, Germany
†email: andresen@WIAS-berlin.de 13:Andresen.tex,session:CS37A

This talk revisits the classical inference results for profile quasi maximum likelihood estimators (profile MLE) in the semiparametric estimation problem. We mainly focus on two prominent theorems: the Wilks phenomenon and Fisher expansion for the profile MLE are stated in a new fashion allowing for finite samples and for model misspecification. The method of study is also essentially different from the usual analysis of the semiparametric problem based on the notion of the hardest parametric submodel. Instead we apply the local bracketing and the upper function devices from Spokoiny (2011) [1]. This novel approach particularly allows to address the important issue of the effective target and nuisance dimension and it does not involve any pilot estimator of the target parameter. The obtained nonasymptotic results are surprisingly sharp and yield the classical asymptotic statements including the asymptotic normality and efficiency of the profile MLE.

The talk explains the main ideas of these results. Further it is shown how the local quadratic bracketing device and the bounds for parametric maximum likelihood estimators from Spokoiny (2011) [1] serve as tools to derive a general convergence result for the alternating procedure to approximate the profile maximum likelihood estimator.

Acknowledgment. This research was partially supported by the German Research Foundation (DFG), grant No.: 1735.

References

[1] Spokoiny V., 2012: Parametric estimation. Finite sample theory, *Ann. Statist.* Volume 40, Number 6, 2877-2909.

Quasi-Maximum Likelihood Estimation and Order Selection for Asymmetric ARMA-GARCH Models

IS19 Shape & Image

BETH ANDREWS*,‡, HUANHUAN WANG†

*Northwestern University, Evanston, Illinois, USA,
†BMO Harris Bank, Chicago, Illinois, USA
‡email: bandrews@northwestern.edu 14:BethAndrews.tex,session:IS19

We consider estimation and order selection for ARMA processes with asymmetric GARCH innovations. These series are correlated with conditional heteroscedasticity and can exhibit a leverage effect. Series of this nature are frequently observed in economics and finance, and are often quite heavy-tailed, so Gaussian maximum likelihood estimators are not relatively efficient. Hence, non-Gaussian quasi-maximum likelihood estimation is considered in this paper. Under mild conditions, estimators of the ARMA and GARCH model parameters are shown to be consistent and we give

a limiting distribution for the estimators. The limiting result can be used to develop chi-squared tests for ARMA-GARCH order selection, which can be used along with the AIC/BIC order selection statistics. When all parameters are in the interior of the parameter space, the limiting distribution is Gaussian. Simulation results show the asymptotic theory is indicative of finite, large sample behavior. We use our results to fit an ARMA-GARCH model to financial data.

Acknowledgment. This research was partially supported by the U.S. National Science Foundation, grant No.: DMS0806104.

OCS2 space-time modeling

New Classes of Nonseparable Space-Time Covariance Functions

TATIYANA V. APANASOVICH*,†

*George Washington University, Washington D.C, USA

†email: apanasovich@gwu.edu 15:Apanasovich.tex,session:OCS2

Space-Time data arise in many different scientific areas, such as environmental sciences, epidemiology, geology, marine biology, to name but a few. Hence, there is a growing need for statistical models and methods that can deal with such data. In this talk we address the need for parametric covariance models which account not only for spatial and temporal dependencies but also for their interactions. It has been argued by many researchers separability of the covariance function can be a very unrealistic assumption in many settings. Hence we propose nonseprable space-time covariance structures which have celebrated Matern family for their spatial margins. Our covariances possess many desirable properties as we demonstrate. For example, the proposed structures allow for the different degree of smoothness for the process in space and time. Moreover, our covariances are smoother along their axis than at the origin. We also describe a simple modification to our family to address the lack of symmetry.

CS7A Spatio-Temp. Stat I.

Estimation of the Parameters of the Matérn Model

N. MIKLÓS ARATÓ*,†

*Eötvös Loránd University, Budapest, Hungary

†email: arato@math.elte.hu 16:MiklosArato.tex,session:CS7A

The Matérn model plays an increasing role in the spatial statistic in recent years. For the Matérn model the corresponding isotropic spectral density is

$$f(u) = \frac{\sigma^2 \alpha^{2\nu}}{\pi^{d/2}(\alpha^2 + u^2)^{\nu + d/2}} ,$$

where d is the dimension, σ^2 is the variance, $\alpha > 0$ and $\nu > 0$ are scale and smoothness parameters, respectively. Previously it was proved that condition $d < 4$ guarantees the equivalence of the probability measures corresponding to different scale parameters. Similarly it was proved that the measures are orthogonal at the case $d > 4$. In the present talk we prove the orthogonality of the measures for the different scale parameters for $d = 4$. At the nonstationary case we calculate the Radon-Nikodym derivative of the measures corresponding to different scale parameters ($d < 4$).

Stochastic Order Applied to the Calculus of Ranking of Bayes Actions in the Exponential Family

POSTER
Poster

JOSÉ PABLO ARIAS-NICOLÁS[*,‡], JACINTO MARTÍN[*], ALFONSO SUÁREZ-LLORENS[†]

[*]University of Extremadura, Spain, [†]University of Cádiz, Spain
[‡]email: jparias@unex.es 17:AriasNicolas.tex,session:POSTER

In a Bayesian context, the set of nondominated actions is considered a first approximation of the solution. Sometimes, this set often is too big to take it as the solution of the problem. So, different criterions are considered to choose an optimal alternative inside the nondominated set. Some authors recommend choosing the conditional Γ-minimax, the posterior regret Γ-minimax or the least sensitive alternatives (see Arias-Nicolás et al. 2009). In this work, we apply stochastic order to study the rank of the Bayes actions considering the exponential family of prior distributions and a class of quantile loss functions defined as $\mathcal{L} = \{L_p : L_p(a,\theta) = |a-\theta| + a(2p-1),\ p \in [p_0, p_1]\}$. The Bayes actions under this model coincide with the quantiles of posterior distributions.

We consider X belong to a one parameter exponential family having a quadratic variance, which has been characterized by Morris (1982). The probability measures of X over an interval $(\theta_0, \theta_1) \subset \{\theta \in \mathbb{R} : a\theta^2 + b\theta + c \text{ and } g(\theta) > 0\}$ has a density of the form $f(x|\theta) = g(\theta)e^{q(\theta)x}, x \in \mathbb{R}$, where g and q are continuosly differentiable real-valued functions satisfying $q'(\theta) = 1/(a\theta^2 + b\theta + c)$ and $g'(\theta)/g(\theta) = -\theta/(a\theta^2 + b\theta + c)$.

We consider the class $\Gamma = \{\pi_{\alpha,\beta}(\theta), \text{ with } \beta \in [\beta_1, \beta_2]\}$ of conjugate prior distributions over the parameter θ, with density function $\pi_{\alpha,\beta}(\theta) = C^{-1}g^{\alpha}(\theta)e^{\beta q(\theta)}$.

The relationship between a prior distribution $\pi(\theta)$ and its corresponding posterior distribution $\pi(\theta|x)$, through Bayes Theorem, is not a simple relationship that allows to obtain properties of the posterior distribution just by observing properties of the prior distribution.

We find the ideal tool for this study in the theory of stochastic orders. The applications of such orders notably simplifies the calculus of the non-dominated set. In particular, we show that the prior distributions in Γ can be ordered in likelihood ratio sense and, therefore they are stochastically ordered.

If two random variables are stochastically ordered, this implies that all their location parameters are also ordered. Let us remember that, in many examples in decision theory, the Bayes alternatives are the location parameters of the posterior distributions. Besides, it is immediate, from the definition, that the stochastic order between two variables implies the order between their respective quantiles.

Based on Theorem 2.1 in Arias-Nicolás et. al 2006, we show a result to the calculus of ranking of Bayes actions when we have imprecision in Decision Maker's beliefs and preferences.

Acknowledgment. This research has been partially supported by *Ministerio de Ciencia e Innovación*, grant No.: MTM2011-28983-C03-02 and partially funded by the *European Regional Development Fund (FEDER)*.

References

[Arias Nicolás et al. (2006)] Arias-Nicolás, J.P., Martín J. and Suárez, A., 2006: The nondominated set in Bayesian decision problems with convex loss functions, *Communications in Statistics*, **34**, 593-607.

[Arias-Nicolás et al. (2009)] Arias-Nicolás, J. P., Martín J., Ruggeri, F. and Suárez-Llorens, A. 2009: Optimal actions in problems with convex loss functions, *International Journal of Approximate Reasoning*, **50**, 303-316.

CS13B Envtl. & Biol. Stat.

Improving the PMA Index by Accounting for Reference Population Variation

JOHANNA ÄRJE[*,¶], FABIO DIVINO[†], SALME KÄRKKÄINEN[*], JUKKA AROVIITA[‡], KRISTIAN MEISSNER[§]

[*]University of Jyväskylä, Finland,
[†]University of Molise, Italy,
[‡]Finnish Environment Institute, Oulu, Finland,
[§]Finnish Environment Institute, Jyväskylä, Finland
[¶]email: johanna.arje@jyu.fi 18:JohannaArje.tex,session:CS13B

In aquatic research and biomonitoring, benthic macroinvertebrates are often used to make statements of ecosystem quality and analysis on different ecological aspects. In Finland, ecological status assessment of rivers using benthic macroinvertebrates is based on a combination of three indices, aiming to incorporate the normative definitions of the European Union Water Framework Directive (WFD). One of the indices is the Percent Model Affinity (PMA, [Novak and Bode, 1992]), which measures the difference between observed macroinvertebrate assemblage and an ideal, "reference status" (i.e. model), assemblage.

The PMA index uses abundance data and first calculates the mean observed proportions of taxa over all reference site samples. Proportions are then standardized so that their sum is one. Second, for any new sample, observed proportions of taxa are compared to the corresponding reference status proportions by calculating their absolute difference. The PMA index produces values ranging from 0 to 1, with 1 being from a sample that has identical proportions to the standardized proportion mean of the reference status, and thus indicates excellent status of the communities. However, the original PMA does not account for the variability of proportions in reference sites. We propose to modify the PMA index, by weighting the distances between the observed proportions and the standardized proportion means so that high variability in reference sites will tend to reduce the distance and low variability in reference sites will increase it. The performance of the proposed version of the PMA index is compared to the original PMA in reference and impacted sites using data provided by the Finnish Environment Institute.

Acknowledgment. This research was partially supported by the Maj and Tor Nessling Foundation, grant No.: 2013138.

References

[Novak and Bode, 1992] M. A. Novak and R. W. Bode,1992: Percent model affinity: a new measure of macroinvertebrate community composition, *J. N. Am. Benthol. Soc.* **11**, 80-85.

IS12 Machine Learning

Eigen-adjusted FPCA for Brain Connectivity Studies

JOHN ASTON[*,§], CI-REN JIANG[†], JANE-LING WANG[‡]

[*]Dept of Statistics, University of Warwick, UK,
[†]Institute of Statistical Science, Academia Sinica, Taiwan,
[‡]Dept of Statistics, UC Davis, USA
[§]email: j.a.d.aston@warwick.ac.uk 19:JohnAston.tex,session:IS12

Functional connectivity as measured using brain imaging techniques such as functional magnetic resonance imaging (fMRI) is profoundly changing the understanding of how the brain works. However, given the size of brain imaging studies, where there are typically on the order of a million voxels

(volume elements) recorded per time point, it is infeasible to examine the full spatial covariance matrix to establish connectivity links. This often leads to Region-Of-Interest analysis being performed or correlations across the whole brain being determined to a very small number of "seed" voxels.

In this work we will propose a functional data analysis methodology for whole brain connectivity. Rather than examining the whole covariance structure explicitly, we define an eigen-adjusted covariance system, where the eigenvalues of the functional principal component analysis are assumed to depend on a covariate (namely spatial location within the image). The eigen-adjusted PC scores then provide a summary of the connectivity, for which a multivariate clustering procedure can be used to define brain locations with similar connectivities.

We will provide a theoretical statistical justification for general framework of eigen-adjusted FPCA and illustrate the application of the technique in a study of 200 resting state fMRI scans which are used to estimate functional connectivity.

A New Depth Function Based on Runs

CS37A Estim. Methods

JEAN-BAPTISTE AUBIN[*,†,‡], SAMUELA LEONI-AUBIN[*,†]

[*]Institut National des Sciences Appliquées de Lyon, Villeurbanne, France,
[†]Institut Camille Jordan, Lyon, France
[‡]email: jean-baptiste.aubin@insa-lyon.fr 20:JeanBaptisteAubin.tex,session:CS37A

The main goal of this contribution is to introduce a new statistical depth function: the LeL *depth* (the Length of the Longest run depth) function. Roughly speaking, the depth of a point $x \in \mathbb{R}^d$, $d > 1$, quantifies the degree to which x is centrally located in a dataset $X_1, \ldots, X_n$. The LeL *depth* of a point $x \in \mathbb{R}^d$, $d > 1$, noted $LeLD_n(x)$, is constructed as follows. Given an hyperplane containing x, consider the length of the longest run of observations consecutively (with respect to some ordering criterion) on the "same side" of the hyperplane. The *LeL depth* of the point x is then defined as a function of the maximum of these lengths over all the possible orderings and all the possible hyperplanes containing x.

Various properties of the *LeL depth* function are explored. First, some asymptotic properties of the LeL *depth* for large sample sizes are studied. Then, a new type of symmetry, the $L-$symmetry is defined and compared to others symmetries such as the C-symmetry and the H-symmetry.

It is shown that the LeL *depth* function possesses the four desirable properties of statistical depth functions introduced by Liu (1990) and Zuo and Serfling (2000). These properties are affine invariance (the depth of a point should not depend on the underlying coordinate system or on the scales of the underlying measurements), maximality at a center, monotonicity (as a point moves away from the "deepest point" -the point with the greatest depth-, the depth of this point should decrease) and vanishing at infinity.

Moreover, some bounds for the value of the *LeL depth* of the deepest point are given. Finally, the *LeL depth* is compared to other classical depth functions for some datasets.

References

[1] Liu, R., 1990: On a notion of data depth based on random simplices, *A*nn. Statist., **18**, 405-414.

[2] Serfling, R. and Zuo, Y., 2000: General notions of statistical depth function, *A*nn. Statist. Volume 28, **2**, 461-482.

POSTER
Poster

A Bayesian Non-Parametric Approach to Asymmetric Dynamic Conditional Correlation Model with Application to Portfolio Selection

M. CONCEPCION AUSIN*,†, AUDRONE VIRBICKAITE*, PEDRO GALEANO*

*Universidad Carlos III de Madrid, Spain
†email: concepcion.ausin@uc3m.es 21:Ausin.tex,session:POSTER

Modeling the dynamics of the assets' returns has been extensively researched for decades and the topic yet remains of great interest, especially in empirical finance setting. GARCH-family models, without doubt, are the most researched and used in practice to explain time-varying volatilities. When dealing with multivariate returns, one must also take into consideration the mutual dependence between them. The asymmetric behavior of individual returns has been well established in the financial literature. However, the use of models explaining asymmetric behavior of covariances is far less common, even though these effects exist. In this work, we consider an asymmetric dynamic conditional correlation GJR-GARCH model (ADCC) which allows for asymmetries not only in individual assets' returns, but also in their correlations. This specification provide a more realistic evaluation of the co-movements of the assets' returns than simple symmetric Multivariate GARCH models.

Whichever GARCH-type model is chosen, the distribution of the returns depends on the distributional assumptions for the error term. It is well known, that every prediction, in order to be useful, has to come with a certain precision measurement. In this way the agent can know the uncertainty of the risk she is facing. Distributional assumptions permit to quantify this uncertainty about the future. However, the traditional premises of Normal or Student-t distributions may be rather restrictive. Alternatively, in this work, we propose a Bayesian non-parametric approach for asymmetric MGARCH model avoiding the specification of a particular parametric distribution for the return innovations. More specifically, we consider a Dirichlet Process Mixture Model (DPM) with a Gaussian base distribution. This is a very flexible model that can be viewed as an infinite location-scale mixture of Gaussian distributions which includes, among others, the Gaussian, Student-t, logistic, double exponential, Cauchy and generalized hyperbolic distributions, among others.

The Bayesian approach also helps to deal with parameter uncertainty in portfolio decision problems. This is in contrast with the usual maximum likelihood estimation approach, which assumes a "certainty equivalence" viewpoint, where the sample estimates are treated as the true values, which is not always correct and has been criticized in a number of papers. This estimation error can gravely distort optimal portfolio selection. In this work, we propose a Bayesian method which provides the posterior distributions of the one-step-ahead optimal portfolio weights, which are more informative than simple point estimates. In particular, using the proposed approach, it is possible to obtain Bayesian credible intervals for the optimal portfolio weights.

Finally, we present some applications of the proposed Bayesian non-parametric approach using real data sets and solve various portfolio decision problems. We explore the differences in uncertainty between the proposed model and conventional restrictive distributional assumptions.

Acknowledgment. Research partially supported by grants ECO2011-25706 and ECO2012-38442 of the Spanish Ministry of Science and Innovation.

Generalized Least Squares Estimation of Panel with Common Shocks

CS9A Model Sel, Lin Reg

PAOLO ZAFFARONI*, MARCO AVARUCCI†,‡

*Imperial College Business School, London, UK,
†Adam Business School, University of Glasgow, UK
‡email: marco.avarucci@glasgow.ac.uk 22:MarcoAvarucci.tex,session:CS9A

This paper considers generalized least square estimation of linear panel models when the innovation and the regressors can both contain a factor structure. A novel feature of this approach is that preliminary estimation of the latent factor structure is not necessary. Under a set of regularity conditions here provided, we establish consistency and asymptotic normality of the feasible GLS estimator as both the cross-section and time series dimensions diverge to infinity. Dependence, both cross-sectional and temporal, of the idiosyncratic innovation is permitted. Monte Carlo experiments corroborate our results.

Inference of the Biological Systems via L_1-Penalized Lasso Regression

CS36A Graphical Methods

EZGİ AYYILDIZ*, VILDA PURUTÇUOĞLU*,†

*Department of Statistics, Middle East Technical University, Ankara, Turkey
†email: vpurutcu@metu.edu.tr 23:EzgiAyyildiz.tex,session:CS36A

The Gaussian Graphical Model (GGM) is one of the well-known deterministic inference methods which is based on the conditional independency of nodes in the system. In this study we consider to implement this approach in a complex JAK-STAT pathway which is one of the major signaling networks that is activated by the type I interferon (IFN) and regulates cytokine-dependent gene expression as well as growth factors of mammals. But here we, particularly, consider the description of the system under the IFN treatment which is developed against the Hepatitis C virus. In inference of the pathway, we perform GGM under the L1-penalized lasso regression method that enables us to estimate the structure of the network with the strength of their interactions. In estimation we initially consider different sizes of systems under distinct proportions of sparsity. Then we implement various penalty selection criteria such as false positive rate, precision, accuracy, and F-measure to detect the optimal choice of penalty constant used in the lasso regression. Then, we select the best criterion from the Monte Carlo runs and use this choice in inference of the JAK-STAT system whose time-course dataset is produced from the Gillespie algorithm. In data generation, firstly the system is run until it converges to its steady-state faze. Then the simulated numbers of molecules are converted to concentrations and taken as a set of time points to get discrete time-course observations. Once we infer the system via the L_1-lasso approach with selected criterion, we assess our findings by comparing them with the current knowledge about the pathway. Finally, in the application of the inference in GGM, we also perform the cross-validation and the gradient descent algorithm. We compare both methods in terms of the accuracy and computational efficiency.

Acknowledgment. The author would like to thank the EU FP-7 Collaborative Project (Project No: 26042) for its financial support.

References

[Maiwald et al. (2010)] Maiwald, T., Schneider, A., Busch, H., Sahle, S., Gretz, N., Weiss, TS., Kummer, U. & Klingmüller, U., 2010: Combining Theoretical Analysis and Experimental Data Generation Reveals IRF9 as a Crucial Factor for Accelerating Interferon α-Induced Early Antiviral Signalling, *The FEBS Journal*, 277 (22), 4741–54.

[Streib et al. (2008)] Streib, F. E. & Dehmer, M., 2008): Statistical Methods for Inference of Genetic Networks and Regulatory Modules, Chapter in: Analysis of Microarray Data: A Network-Based Approach, Weinheim, Wiley.

[Whittaker (1990)] Whittaker, J., 1990: Graphical Models in Applied Multivariate Statistics, New York, John Wiley and Sons.

[Wit et al. (2010)] Wit, E., Vinciotti, V., & Purutçuoğlu, V., 2010: Statistics for Biological Networks: Short Course Notes, *25th International Biometric Conference (IBC)*, Florianopolis, Brazil.

IS12 Machine Learning

Stochastic Gradient Methods for Large-Scale Machine Learning

FRANCIS BACH

INRIA - Ecole Normale Supérieure, Paris, France

*email: francis.bach@ens.fr 24:Bach_Francis.tex,session:IS12

Many machine learning and signal processing problems are traditionally cast as convex optimization problems. A common difficulty in solving these problems is the size of the data, where there are many observations ("large n") and each of these is large ("large p"). In this setting, online algorithms which pass over the data only once, are usually preferred over batch algorithms, which require multiple passes over the data. In this talk, I will present several recent results, showing that in the ideal infinite-data setting, online learning algorithms based on stochastic approximation should be preferred, but that in the practical finite-data setting, an appropriate combination of batch and online algorithms leads to unexpected behaviors, such as a linear convergence rate with an iteration cost similar to stochastic gradient descent.

References

[1] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Technical report, HAL-00804431, 2013.

[2] N. Le Roux, M. Schmidt, F. Bach. A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets. Advances in Neural Information Processing Systems (NIPS), 2012.

[3] F. Bach, E. Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. Advances in Neural Information Processing Systems (NIPS), 2011.

CS28A Random Graphs

The Asymptotic Degree Distribution of a Random Graph Model with Duplications

ÁGNES BACKHAUSZ*,†, TAMÁS F. MÓRI*

*Eötvös Loránd University, Budapest, Hungary

†email: agnes@cs.elte.hu 25:AgnesBackhausz.tex,session:CS28A

We examine a random graph model where the basic step is the duplication of a vertex. That is, we select a vertex randomly, and create a new vertex that is connected to the neighbours of this vertex. These kind of random graph models are often used in biological sciences, e.g. for modelling interactions of proteins, where it often happens that different proteins have similar role and therefore they have many common neighbours.

These models were examined from many points of view. One of them deals with the distribution of the degree of a randomly chosen vertex, and then the proportion of vertices of large degree. More precisely, if the proportion of vertices of degree d tends to some constand c_d almost surely as

the number of vertices tends to infinity, then we say that the sequence c_d is the asymptotic degree distribution. Moreover, if c_d is polynomially decaying, that is, $c_d d^\gamma$ is converges to a positive constant for some $\gamma > 0$, then the random graph model has the so called scale free property. Several real world networks were examined empirically and their asymptotic degree distribution was not far from polynomially decaying sequences.

In the last decades many scale free random graph models were constructed and analysed. However, for the duplication models the mathematical proofs of the scale free property were missing in the previously defined strong sense, when almost sure convergence is included instead of the convergence of expectations.

The result presented in the talk is the proof of the scale free property of an appropriately defined random graph model with duplications. The model is the following. We start from a single vertex. At each step a new vertex is born. We select an old vertex uniformly at random and we connect the new vertex to its neighbours and also to the selected old vertex itself. This is the duplication part. Then comes the erasure part: we choose an old vertex uniformly at random independently from the previous choices and delete all its edges, except the edge connecting it to the new vertex. This procedure is iterated.

The steps are defined to be independent. Note that at the duplication part the probability that a given old vertex gets a new edge depends on its degree, and vertices of larger degree have larger chance to be connected to the new vertex. Thus the model is a kind of preferential attachment model. If there is no erasure step, then the graph becomes dense, and there is no asymptotic degree distribution. This way the number of edges is bounded by a linear function of the number of vertices.

The proof is based on a coupling argument. We modify the model such that the new graph sequence has a very special structure. This enables us to use methods of martingale theory to determine the asymptotic degree distribution of the modified model. Then we prove that the original one has the same asymptotic degree distribution. In addition, this kind of argument shows that the graph has large complete subgraphs with a few edges between them.

Statistical Evaluation of Low-Template DNA Profiles

IS7 Forensic Stat.

DAVID BALDING*,†

*Institute of Genetics, University College London, UK
†email: d.balding@ucl.ac.uk

26:Balding.tex,session:IS7

Recently, forensic DNA profiling has been used with far smaller volumes of DNA than was previously thought possible. This “low template" (or LTDNA) profiling enables evidence to be recovered from the slightest traces left by touch or even breath, but brings with it serious interpretation problems that courts have not yet adequately solved. These problems have contributed to important cases collapsing or convictions being overturned, for example in R v Hoey in Northern Ireland, and the case of Knox and Sollecito in Italy. The most important challenge to interpretation arises when either or both of “dropout" and “dropin" create discordances between the crime scene DNA profile and that expected under the prosecution allegation. Stochastic artefacts affecting the peak heights in the electropherogram (epg) are also problematic, in addition to the effects of masking from the profile of a known contributor.

One of the major issues in modelling the “raw" LTDNA profiling results is whether or not to use the epg peak heights. These are potentially informative, but can be subject to substantial variation, and the pattern of variation is highly machine/protocol-specific. I will briefly review other approaches and then describe the approach that I have developed, based on likelihoods that do not use peak height but allow an “uncertain" category in addition to presence/absence for each allele.

This approach is implemented in R code `likeLTD` that is freely available on the author's website sites.google.com/site/baldingstatisticalgenetics.

I apply `likeLTD` to casework LTDNA examples and reveal deficiencies in some reported analyses. Courts increasingly demand "validation" of software, although they seem to be unclear about what that should mean. A problem with likelihood ratios for complex DNA evidence is that there is no "gold standard": there is no non-trivial profile for which it is apparent what the likelihood ratio (LR) should be. Even if the true contributors of DNA to a profile are known, that does not determine the appropriate LR for a noisy LTDNA profile. We overcome this problem by showing that the LR for multiple noisy replicates converges to that for a single, high-quality DNA profile.

CS7B Spatio-Temp. Stat II.

Probabilistic Temperature Forecasting with Statistical Calibration in Hungary

SÁNDOR BARAN[*,†,§], ANDRÁS HORÁNYI[‡], DÓRA NEMODA[*]

[*]University of Debrecen, Debrecen, Hungary,
[†]University of Heidelberg, Heidelberg, Germany,
[‡]Hungarian Meteorological Service, Budapest, Hungary
[§]email: baran.sandor@inf.unideb.hu

27:BaranS.tex,session:CS7B

Bayesian Model Averaging (BMA) is a statistical post-processing technique which produces calibrated probability density functions (PDF) of the predictable meteorological quantities from ensembles of forecasts using predictions and validating observations of the preceding days. The obtained PDFs are convex combinations of PDFs corresponding to the bias corrected ensemble members, and the weights express the predictive skill of an ensemble member over the training period.

In the present work we show our experiences with the application of the BMA normal model for exchangeable forecasts [Fraley et al., 2010] to temperature data of the Hungarian Meteorological Service (HMS). The data file contains 11 member ensembles of 42 hour forecasts for temperature for 10 major cities in Hungary generated by the Limited Area Model Ensemble Prediction System called ALADIN-HUNEPS of the HMS [Horányi et al., 2011], together with the corresponding validating observations for a six months period. An ensemble consists of 10 exchangeable forecasts started from perturbed initial conditions and one control member started from the unperturbed analysis. We remark, that BMA method has already been successfully applied for wind speed ensemble forecasts of the ALADIN-HUNEPS [Baran et al., 2013] – BMA post-processing of these forecasts significantly improved the calibration and the accuracy of point forecasts as well.

As a first step we found the optimal length of the training period by comparing the goodness of fit of models with training periods ranging from 10 to 60 days. Then we checked the fit of the optimal BMA model and using appropriate scoring rules we showed the advantage of the BMA post-processed forecasts compared to the predictions calculated from the raw ensembles. The BMA post-processing and the analysis of the models was performed with the help of the `ensembleBMA` package of R.

Acknowledgment. This research has been supported by by the Hungarian Scientific Research Fund under Grants No. OTKA T079128/2009 and OTKA NK101680/2012, by the Hungarian –Austrian intergovernmental S&T cooperation program TÉT_10-1-2011-0712 and partially supported the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[Baran et al., 2013] Baran, S., Horányi, A., Nemoda, D., 2013: Statistical post-processing of probabilistic wind speed forecasting in Hungary, *Meteorol. Z.*, to appear.

[Fraley et al., 2010] Fraley, C., Raftery, A. E. and Gneiting, T., 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging, *Mon. Wea. Rev.* **138**, 190 - 202.

[Horányi et al., 2011] Horányi, A., Mile, M., Szűcs, M., 2011: Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary, *Tellus A* **63**, 642 - 651.

A New Estimator in the Regression Setting

IS4 Empirical Proc.

YANNICK BARAUD[*,‡], LUCIEN BIRGÉ[†]

[*]Université de Nice Sophia Antipolis, Laboratoire J.A. Dieudonné,
[†]Université Paris 6, Laboratoire de Probabilités et Modèles Aléatoires
[‡]email: baraud@unice.fr 28:Baraud_Yannick.tex,session:IS4

In the regression setting, it is well-known that the distribution of the noise strongly influences the rates of convergence toward the regression function and that these rates can eventually be much faster than the usual ones obtained under a Gaussian noise assumption.

In order to tackle this problem, we build a new estimator which can not only adapt to the regularity of the regression function, but also (possibly) to the singularity of the noise distribution in order to achieve these faster rates. Our approach is based on model selection and amounts to choosing from the data a suitable model among a collection of candidate ones for both the regression function and the noise distribution.

Our results are not asymptotic and lead to oracle-type inequalities. The proof relies on the control of the supremum of a suitable empirical process over an Hellinger-type ball. In particular, we show how the expectation of this supremum governs the properties of our estimator from a non-asymptotic point of view. We also present some simulations in view of illustrating the theoretical part of talk.

Parameter Estimation for Affine Processes

OCS28 Stat. Affine Proc.

MÁTYÁS BARCZY[*,¶], LEIF DÖRING[†], ZENGHU LI[‡], GYULA PAP[§]

[*]University of Debrecen, Debrecen, Hungary,
[†]Universität Zürich, Zürich, Switzerland,
[‡]Beijing Normal University, Beijing, People's Republic of China,
[§]University of Szeged, Szeged, Hungary
[¶]email: barczy.matyas@inf.unideb.hu 29:MatyasBarczy.tex,session:OCS28

Let us consider the affine diffusion process (affine two-factor model) given by the SDE

$$\begin{cases} \mathrm{d}Y_t = (a - bY_t)\,\mathrm{d}t + \sqrt{Y_t}\,\mathrm{d}W_t, \\ \mathrm{d}X_t = (m - \theta X_t)\,\mathrm{d}t + \sqrt{Y_t}\,\mathrm{d}B_t, \end{cases} \qquad t \geq 0, \tag{1}$$

where $a > 0$, $b, \theta, m \in \mathbb{R}$, and $(W_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ are independent standard Wiener processes. The first coordinate $(Y_t)_{t\geq 0}$ is known to be a square-root process. We study the asymptotic behaviour of the maximum likelihood and (conditional) least squares estimators of the parameters m and θ based on discrete and continuous time observations of the process.

In the critical case $b = 0, \theta = 0$, we examine the asymptotic behaviour of the (conditional) least squares estimators of m and θ based on discrete time observations of the process X. For the proofs, we derive a simple set of sufficient conditions for the weak convergence of scaled affine processes with more general state space $\mathbb{R}_+ \times \mathbb{R}^d$. We specialize our scaling theorem to one-dimensional continuous state branching processes with immigration.

In the subcritical case $b > 0$, $\theta > 0$, we show strong consistency and asymptotic normality of the maximum likelihood and least squares estimators of m and θ based on continuous time observations of the process (Y, X). The existence of a unique stationary distribution and ergodicity are also shown for the model, which turn out to be crucial for the proofs.

Finally, if we replace the square root in the first SDE of (1) by α-root with $\alpha \in (1, 2)$, then the existence of a unique stationary distribution for the new model is proved.

The full presentation of our results can be found in [Barczy et al. (2013a)], [Barczy et al. (2013b)] and [Barczy et al. (2013c)].

Acknowledgment. M. Barczy, Z. Li and G. Pap have been supported by the Hungarian Chinese Intergovernmental S & T Cooperation Programme for 2011-2013 under Grant No. 10-1-2011-0079. M. Barczy and G. Pap have been partially supported by the Hungarian Scientific Research Fund under Grant No. OTKA T-079128. L. Döring has been supported by the Foundation Science Matématiques de Paris. Z. Li has been partially supported by NSFC under Grant No. 11131003 and 973 Program under Grant No. 2011CB808001.

References

[Barczy et al. (2013a)] Barczy, M., Döring, L., Li, Z., Pap, G., 2013: On parameter estimation for critical affine processes, *Electronic Journal of Statistics*, Volume **7**, 647 - 696.

[Barczy et al. (2013b)] Barczy, M., Döring, L., Li, Z., Pap, G., 2013: Ergodicity for an affine two factor model, arXiv: 1302.2534.

[Barczy et al. (2013c)] Barczy, M., Döring, L., Li, Z., Pap, G., 2013: Parameter estimation for an affine two factor model, arXiv: 1302.3451.

NYA
Not Yet
Arranged

A New Fuzzy Time Series Forecasting Method Based on Fuzzy Rule Based Systems and OWA Operator

MURAT ALPER BASARAN[*,†], HILMI UYAR[*]

[*]Akdeniz University, Antalya, Turkey
[†]email: muratalper@yahoo.com 30:MuratAlperBasaran.tex,session:NYA

Fuzzy time series models have found several application areas when classic one has failed to generate forecasts. Fuzzy time series models have been investigated since the first paper of Song and Chissom. Several aspects of them have been examined since then. Recently two- factors high order fuzzy time series have been introduced and applied to some data sets with a high forecasting accuracy. In this paper, a new method taking into account the membership degrees, using fuzzy rules-based systems and OWA operator as aggregation function for two factor- high order fuzzy time series method is proposed in order to forecast temperature data and TAIFEX data. Its results are promising.

CS19E
Lim.
Thms.

Limit Theorems for some Inelastic Kinetic Models

BASSETTI FEDERICO[*,‡], LUCIA LADELLI[†]

[*]University of Pavia, Italy,
[†]Politecnico of Milan, Italy
[‡]email: federico.bassetti@unipv.it 31:FedericoBassetti.tex,session:CS19E

Homogeneous kinetic equations model the redistribution of energy and momentum in an ensemble of particles which interact in binary collisions. In the original Boltzmann equation, these collisions are fully elastic, i.e., they conserve the total energy and momentum of the two interacting

particles precisely. It is well known that, in this case, the velocity distribution converges weakly to a Gaussian law. Over the last decade, inelastic models of various kinds have been introduced. See, e.g., [Bobylev et al. (2003)]. In those models, collisions are not fully elastic and this leads to a statistical gain and/or loss of the total energy in binary collisions. The stationary velocity distributions turn out to be scale mixtures of stable laws.

One dimensional versions of these models (extensions of the well-known Kac caricature) have been studied, from an essentially analytic viewpoint, in [Bobylev et al. (2009)]. We show how these models can be studied by probabilistic methods, based on techniques pertaining to the classical central limit problem and to the fixed-points of smoothing transformations. See [Bassetti et al. (2011), Bassetti et al. (2012)]. An advantage of resorting to these methods is that the same results —relative to self-similar solutions— as those obtained in [Bobylev et al. (2009)], are here deduced under weaker conditions. In particular, it is shown how convergence to a self-similar solution depends on the belonging of the initial datum to the domain of attraction of a specific stable distribution of index $0 \leq \alpha \leq 2$. In addition, if $\alpha \neq 2$ and under suitable assumptions on the collisional kernel, the precise asymptotic behavior of the large deviations probability is deduced. See [Bassetti et al. (2013)].

The probabilistic approach to study the long time behavior of the Kac-type models can be extended to some multidimensional inelastic Bolztmann equations. In this case, a key tool is a probabilistic interpretation of the solution, which has been introduced recently by [Dolera et al. (2012)].

References

[Bassetti et al. (2011)] Bassetti, F., Ladelli, L., Matthes, D., 2011: Central limit theorem for a class of one-dimensional kinetic equations, *Probab. Theory Related Fields*, **150**, 77 - 109.

[Bassetti et al. (2012)] Bassetti, F., Ladelli, L., 2012: Self similar solutions in one-dimensional kinetic models: a probabilistic view, *Ann. App. Prob.*, **22**, 1928 - 1961.

[Bassetti et al. (2013)] Bassetti, F., Ladelli, L., 2013: Large Deviations for the solution of a Kac-type kinetic equation, *Kinet. Relat. Models*, **6**, 245 - 268.

[Bobylev et al. (2003)] Bobylev, A. V., Cercignani, C., 2003: Self-similar asymptotics for the Boltzmann equation with inelastic and elastic interactions, *J. Statist. Phys.*, **110**, 333 - 375.

[Bobylev et al. (2009)] Bobylev, A.V., Cercignani, C., Gamba, I.M., 2009: On the self-similar asymptotics for generalized nonlinear kinetic maxwell models, *Comm. Math. Phys.*, **291**, 599 - 644.

[Dolera et al. (2012)] Dolera, E., Regazzini, E., 2012: Proof of a McKean conjecture on the rate of convergence of Boltzmann-equation solutions, arXiv: 1206.5147.

Approximate Bayesian Computation to Improve Dynamical Modelling of Dietary Exposure

CS12B Bayesian computing

CAMILLE BÉCHAUX*,‡, AMÉLIE CRÉPET*, STÉPHAN CLÉMENÇON†

*ANSES, French Agency for Food, Environmental and Occupational Health Safety, Maisons-Alfort, France,
†Telecom ParisTech, Paris, France
‡email: camille.bechaux@anses.fr

32:BECHAUX.tex,session:CS12B

Exposure assessment is an essential step in human health risk assessment. Since the true level of exposure is never known, exposure is usually indirectly assessed from exposure models which combine food contamination data and food consumption data. Recently available biomonitoring data allow for a certain measurement of the exposure through measurement of the concentration of a chemical in biological substances (blood or urine for example). However, biomonitoring data alone are difficult to interpret and accurate exposure assessment from indirect data remains a necessary tool

to develop effective health risk management strategies. Thus using information from biomonitoring data to improve modeling of exposure has become a new challenge in risk assessment.

This paper proposes a method to integrate the information provided by biomonitoring data to fit a dynamical exposure model which takes into account the accumulation of the chemical in the body and the evolution of the past exposure.

A Kinetic Dietary Exposure Model (KDEM) is used to describe the temporal evolution of the total body burden of a chemical present in a variety of foods involved in the diet. The dynamic of this model is governed by two components, a marked point process (MPP) which describes the dietary behaviour and a linear differential equation which takes into account the accumulation of the chemical in the human body and its physiological elimination. Moreover, the exposure to contaminants present in the environment for a while is often known to have changed over time. Therefore a function λ determined by a vector of parameters θ which describes the evolution of the exposure over the last decades is added to the model and θ is fitted using biomonotoring data.

Since the likelihood of the proposed dynamical exposure model is very complex and thereby neither analytically solvable nor computationnally tractable, a likelihood free method named Approximate Bayesian Computation (ABC) is used to fit this model to biomonitoring data. Original summary statistics and distance are suggested to suit to the data and the risk assessment context. In order to decrease the number of simulations required to estimate the posterior distribution comparing to a basic rejection algorithm, an ABC-MCMC algorithm is implemented. The historical variation of the exposure over time modeled by λ can be described with varying degrees of accuracy and a model selection has to be performed. A two stages hierarchical algorithm is proposed to include the length of θ as a new parameter. Thus, parameter estimation for each model is performed simultaneously with model selection.

Applied to blood concentration of dioxins data in a population of French fisherman families, this method results in the calibration of the function λ which describes the unknown evolution of the exposure to dioxins during the last few decades. It results also in an exposure model able to predict the body burdens of the population from current dietary intakes. In this application, the different chains have converged after 150,000 iterations and the acceptance rate is close to 38%. Therefore, the chosen ABC-MCMC algorithm seems efficient enough in this case. In situations where a more complex model of exposure is needed or where it is necessary to be more restrictive on the tolerance δ, it would be relevant to opt for a more efficient algorithm, sequential importance sampling algorithm for instance.

CS14A Stat. Neuronal Data

A Non Parametric Estimator for The Hazard Rate Function in Presence of Dependent Sample Data

ELISA BENEDETTO[*,†], FEDERICO POLITO[*], LAURA SACERDOTE[*]

[*]Mathematics Department of Torino University, Italy
[†]email: elisa.benedetto@unito.it 33:ElisaBenedetto.tex,session:CS14A

In point process literature there are many examples of hazard rate estimators. When the point process is a Poisson or a general renewal process, there exist various estimators for the hazard rate (see e.g. [2], [3], [4], [5]). However such assumptions are too strong to model data from several instances, as for example neural spike trains and learning processes. Our aim is to relax the hypothesis of independence and provide a uniform strong consistent estimator for the hazard rate function of a point process. In [1], a maximum likelihood estimator of the hazard rate function is obtained for a point process with a given inter-event interval distribution. Here we focus our attention on a non-parametric estimation of the hazard rate. Hence we assume that the inter-event intervals belong

to a Markov and ergodic stochastic process. Under these hypotheses we propose a non-parametric uniform strong consistent hazard rate estimator. Then we validate our estimate with a specific statistical test. Finally, we illustrate our method on some examples with simulated data, including an application to neuroscience.

References

[1] E.N. Brown, R. Barbieri, V. Ventura, R.E. Kass and L.M. Frank, The time rescaling theorem and its application to neural spike train data analysis., Neural Computation 14, 325-346 (2001).

[2] B.L.S. Prakasa Rao, and J. Van Ryzin, Asymptotic theory for two estimators of the generalized failure rate., Statistical Theory and Data Analysis, ed. K. Matusita, North Holland, Amsterdam, 547-563 (1985).

[3] J. Rice and M. Rosenblatt, Estimation of the log survivor function and hazard function., Sankhya Ser. A, 38, 60-78 (1976).

[4] G.S. Watson and M.R. Leadbetter, Hazard analysis I, Biometrika, 51, 175-184 (1964).

[5] G.S. Watson and M.R. Leadbetter, Hazard analysis II, Sankhya Ser. A, 26, 110-116 (1964).

Local Asymptotic Mixed Normality in a Heston Model

OCS28 Stat. Affine Proc.

JÁNOS MARCELL BENKE*,†, GYULA PAP*

*University of Szeged, Hungary

†email: jbenke@math.u-szeged.hu 34:BenkeJanosMarcell.tex,session:OCS28

In this paper we consider a Heston model

$$\begin{cases} \mathrm{d}Y_t = (a - bY_t)\,\mathrm{d}t + \sigma_1\sqrt{Y_t}\,\mathrm{d}W_t, \\ \mathrm{d}X_t = (m - \beta Y_t)\,\mathrm{d}t + \sigma_2\sqrt{Y_t}\big(\varrho\,\mathrm{d}W_t + \sqrt{1-\varrho^2}\,\mathrm{d}B_t\big), \end{cases} \qquad t \geq 0,$$

where $a > 0$, $b, m, \beta \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, $\varrho \in (-1, 1)$ and $(W_t, B_t)_{t\geq 0}$ is a 2-dimensional standard Wiener process. We study the local asymptotic properties of the likelihood function for this model in the sense of Le Cam. There are 3 cases. In the subcritical case ($b > 0$) local asymptotic normality (LAN) holds. In the supercritical case ($b < 0$) if we fix the parameters a and m, local asymptotic mixed normality (LAMN) holds. Finally in the critical case ($b = 0$) if we fix the parameters b and β, we assert the LAN property. In these cases the maximum likelihood estimation is asymptotically efficient in the sense of the convolution theorem.

Acknowledgment. This research was partially supported by the Hungarian National Research Fund OTKA, grant No.: T-079128.

References

[1] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in statistics: some basic concepts.* Springer.

Simple Fractional Dickey-Fuller Test

CS4B Time Series I.

AHMED BENSALMA*,†

*Ecole National Supérieure de la Statistique et de l'Economie Appliquée (ENSSEA), Alger, Algiers

†email: bensalma.ahmed@gmail.com; bensalma.ahmed@enssea.dz

35:BENSALMA.tex,session:CS4B

This paper proposes a new testing procedure for the degree of fractional integration of a time series inspired on the unit root test of Dickey-Fuller (1979). The composite null hypothesis is that

of $d \geq d_0$ against $d < d_0$. The test statistics is the same as in Dickey-Fuller test using as output $\Delta^{d_0} y_t$ instead of Δy_t and as input $\Delta^{-1+d_0} y_{t-1}$ instead of y_{t-1}, exploiting the fact that if y_t is $I(d)$ then $\Delta^{-1+d_0} y_t$ is $I(1)$ under the null $d = d_0$. If $d \geq d_0$, using the generalization of Sowell's results (1990), we propose a test based on the least favorable case $d = d_0$, to control type I error and when $d < d_0$ we show that the usual tests statistics diverges to $-\infty$, providing consistency.

By noting that $d - d_0$ can always be decomposed as $d - d_0 = m + \delta$, where $m \in \mathbb{N}$ and $\delta \in]-0.5, 0.5]$, the asymptotic null and alternative of the Dickey-Fuller, normalized bias statistic $n\widehat{\rho}_n$ and the Dickey-Fuller t-statistic $t_{\widehat{\rho}_n}$ are provided by the theorem 1.

Theorem 1. *Let $\{y_t\}$ be generated according DGP $\Delta^d y_t = \varepsilon_t$. If regression model $\Delta^{d_0} y_t = \widehat{\rho}_n \Delta^{-1+d_0} y_{t-1} + \widehat{\epsilon}_t$ is fitted to a sample of size n then, as $n \uparrow \infty$,*

1. *$n\widehat{\rho}_n$ verifies that*

$$\begin{aligned}
&\widehat{\rho}_n = O_p(\log^{-1} n) && \text{and} && (\log n)\,\widehat{\rho}_n \xrightarrow{p} -\infty, && \text{if } d - d_0 = -0.5,\\
&\widehat{\rho}_n = O_p(n^{-1-2\delta}) && \text{and} && n\widehat{\rho}_n \xrightarrow{p} -\infty, && \text{if } -0.5 < d - d_0 < 0,\\
&\widehat{\rho}_n = O_p(n^{-1}) && \text{and} && n\widehat{\rho}_n \Rightarrow \frac{\frac{1}{2}\left\{\mathbf{w}^2(1) - 1\right\}}{\int_0^1 \mathbf{w}^2(r)dr}, && \text{if } d - d_0 = 0,\\
&\widehat{\rho}_n = O_p(n^{-1}) && \text{and} && n\widehat{\rho}_n \Rightarrow \frac{\frac{1}{2}\mathbf{w}^2_{\delta,m+1}(1)}{\int_0^1 \mathbf{w}^2_{\delta,m+1}(r)dr}, && \text{if } d - d_0 > 0.
\end{aligned}$$

2. *$t_{\widehat{\rho}_n}$ verifies that*

$$\begin{aligned}
&t_{\widehat{\rho}_n} = O_p(n^{-0.5}\log^{-0.5} n) && \text{and} && t_{\widehat{\rho}_n} \xrightarrow{p} -\infty, && \text{if } d - d_0 = -0.5,\\
&t_{\widehat{\rho}_n} = O_p(n^{-\delta}) && \text{and} && t_{\widehat{\rho}_n} \xrightarrow{p} -\infty, && \text{if } -\frac{1}{2} < d - d_0 < 0,\\
&t_{\widehat{\rho}_n} = O_p(1) && \text{and} && t_{\widehat{\rho}_n} \Rightarrow \frac{\frac{1}{2}\left\{\mathbf{w}^2(1) - 1\right\}}{\left[\int_0^1 \mathbf{w}^2(r)dr\right]^{1/2}}, && \text{if } d - d_0 = 0,\\
&t_{\widehat{\rho}_n} = O_p(n^{\delta}) && \text{and} && t_{\widehat{\rho}_n} \xrightarrow{p} +\infty, && \text{if } 0 < d - d_0 < 0.5,\\
&t_{\widehat{\rho}_n} = O_p(n^{0.5}) && \text{and} && t_{\widehat{\rho}_n} \xrightarrow{p} +\infty, && \text{if } d - d_0 \geq 0.5,
\end{aligned}$$

where $\mathbf{w}_{\delta,m}(r)$ is $(m-1)$-fold integral of $\mathbf{w}_\delta(r)$ recursively defined as $\mathbf{w}_{\delta,m}(r) = \int_0^r \mathbf{w}_{\delta,m-1}(s)ds$, with $\mathbf{w}_{\delta,1}(r) = \mathbf{w}_\delta(r)$ and $\mathbf{w}(r)$ is the standard Brownian motion.

These properties and distributions are the generalization of those established by Sowell (1990) for the cases $-\frac{1}{2} < d - 1 < 0$, $d - 1 = 0$ and $0 < d - 1 < \frac{1}{2}$.

References

[1] Bensalma, A. 2012: Unified theoretical framework for unit root and fractional unit root, arXiv: 1209.1031.

OCS25 Long-mem. Time Ser.

On Estimating Higher Order Derivatives and Smooth Change Points for Locally Stationary Long-Memory Processes

JAN BERAN*,†

*Department of Statistics, University of Konstanz, Germany
†email: jan.beran@uni-konstanz.de

36:JanBeran.tex,session:OCS25

Locally stationary processes with long memory based on discrete or continuous time Gaussian or more general subordination provide feasible models in situations where time series are subject to

smooth local changes. A particular application is the analysis of climatological time series where fast changes are of special interest. In contrast to a classical change point setting changes are expected to be smooth. Moreover, often historic data are not equidistant. This leads to the estimation of higher order derivatives of nonparametric curves from noisy, nonequidistant observations. Starting with a summary of results for equidistant stongly dependent stationary and locally stationary time series, extensions to nonequidistant subordinated processes are presented. Part of the work reviewed in this talk is based on joint research with Yuanhua Feng, Sucharita Ghosh and Patricia Menendez.

References

[Beran (2009)] Beran, J., 2009: On parametric estimation for locally stationary long-memory processes. *J. Statistical Planning and Inference*, **Vol. 139, No. 3**, 900-915.

[Beran and Feng (2002)] Beran, J. and Feng, Y., 2002: Local polynomial fitting with long-memory, short-memory and antipersistent errors. *Annals of the Institute of Statisitcal Mathematics*, *Vol. 54, No. 2*, 291-311.

[Beran and Feng (2007)] Beran, J. and Feng, Y., 2007: Weighted averages and local polynomial estimation for fractional linear ARCH processes. *Journal of Statistical Theory and Practice*, **Vol. 1, No. 2**, 149-166.

[Beran et al. (2013)] Beran, J., Feng, Y., Ghosh, S. and Kulik, R., 2013: Long-memory processes - Probabilistic Properties and Statistical Methods. *Springer, New York/Heidelberg*.

[Menendez et al. (2010)] Menendez, P., Ghosh, S. and Beran, J., 2010: On rapid change points under long memory. *Journal of Statistical Planning and Inference*, **Vol. 140, No. 11**, 3343-3354.

[Palma and Ferreira (2013)] Palma, W. and Ferreira, G., 2013. Regression estimation with locally stationary long-memory errors. *Journal of Multivariate Analysis*, **116**, 14 - 24.

[Palma and Olea (2010)] Palma, W. and Olea, R., 2010: An efficient estimator for locally stationary Gaussian long-memory processes. *Ann. Statist.*, **Vol. 38, No. 5**, 2958 - 2997.

[Roueff 2011] Roueff, F. and von Sachs, R., 2011: Locally stationary long memory estimation. *Stochastic Processes and their Applications*, **Vol. 121, Issue 4**, 813 - 844.

Recent Results in St. Petersburg Theory

StPburgS St. Petersburg Mem. Sess.

ISTVÁN BERKES*,†

*Graz University of Technology, Austria

†email: berkes@tugraz.at

37:Berkes.tex,session:StPburgS

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $P(X_1 = 2^k) = 2^{-k}$, $k = 1, 2, \ldots$ and let $S_n = \sum_{k=1}^{n} X_k$. A remarkable property of the St. Petersburg process $(X_k)_{k\geq 1}$ is that X_1 does not belong to any domain of attraction, i.e. $(S_n - a_n)/b_n$ has no nondegenerate limit distribution for any (a_n), (b_n), but the asymptotic behavior of S_n can be described precisely. Specifically, after an exponential time transformation, $Z_n = (S_n - n\log_2 n)/n$ is (asymptotically) periodic with a continuous set of subsequential limit distributions. As it turns out, this periodicity holds for many nonlinear functionals of the process $(X_k)_{k\geq 1}$ as well and leads to interesting limit theorems such as strong approximation with semistable Lévy processes, unusual a.s. central limit theorems, "St. Petersburg type" versions of classical sampling theorems, trimming theorems, Edgeworth expansions, etc. The purpose of this talk is to survey this field and to discuss recent developments.

CS12B Bayesian computing

Quantification of Estimation Instability and its Application to Threshold Selection in Extremes

THOMAS BERNING[*,†]

[*]Stellenbosch University, Stellenbosch, South Africa
[†]email: tberning@sun.ac.za 38:ThomasBerning.tex,session:CS12B

The main goal of this paper is to propose a measure which quantifies the instability of estimates over a range of chosen values of some other parameter. The measure is applied in an extreme value analysis context, where the perturbed Pareto distribution (PPD) is fitted to observed relative excesses. The estimates of the extreme value index (EVI) are considered over a range of choices for the number of excesses k. Methods are then developed to identify which range of values of k yields the lowest instability measure. The resulting region is then considered the *stable region*.

An estimator of the EVI is proposed as the mean of the EVI estimates in the stable region. For the purpose of inference, k has to be specified. The concept of *implied threshold* is introduced, where the choice of k is taken as the value of k inside the stable region which yields an EVI estimate closest to the mean EVI value over the stable region.

The instability measure is also applied to second order parameter estimation. The estimator by Gomes and Martins (2001) is employed as external estimator of the second order parameter of the PPD. This estimator is also modified and used to adaptively choose between methods of identifying the stable region. An extensive simulation study shows that the aforementioned approaches lead to more accurate estimators of the extreme value index.

References

[Beirlant et al. (2004)] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.L. 2004: Statistics of extremes. Chichester: John Wiley and Sons.

[Drees et al. (1998)] Drees, H., Kaufmann, E., 1998: Selecting the optimal sampling fraction in univariate extreme value estimation, *Stochastic Processes and their Applications*, **75**, 149-172.

[Gomes et al. (2001)] Gomes, M.I., Martins, M.J., 2001: Generalizations of the Hill estimator - asymptotic versus finite sample behaviour, *Journal of Statistical Planning and Inference*, **93**, 161-180.

[Gomes et al. (2002)] Gomes, M.I., Martins, M.J., 2002: Asymptotically unbiased estimators of the tail index based on external estimation of the second order parameter, *Extremes*, **5**, 5-31.

[Guillou et al. (2001)] Guillou, A., Hall, P., 2001: A diagnostic for selecting the threshold in extreme value analysis, *Journal of the Royal Statistical Society*, **63**, 293-305.

POSTER Poster

Asymptotically Optimal Method for Manifold Estimation Problem

ALEXANDER KULESHOV[*], ALEXANDER BERNSTEIN[†,*,‡], YURY YANOVICH[*]

[*]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia,
[†]Institute for System Analysis, Russian Academy of Sciences, Moscow, Russia
[‡]email: a.bernstein@mail.ru 39:BernsteinAlexander.tex,session:POSTER

Let $\mathbf{X}$ be an unknown nonlinear smooth q-dimensional Data manifold (D-manifold) embedded in a p-dimensional space ($p > q$) covered by a single coordinate chart. It is assumed that the manifold's condition number is positive so $\mathbf{X}$ has no self-intersections. Let $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\} \subset \mathbf{X} \subset \mathbf{R}^p$ be a sample randomly selected from the D-manifold $\mathbf{X}$ independently of each other according to an unknown probability measure on $\mathbf{X}$ with strictly positive density. The Manifold Estimation problem

(ME) is to construct sample-based q-dimensional manifold (Estimated Data Manifold, ED-manifold) $\mathbf{X}_\theta \subset \mathbf{R}^p$ covered by a single coordinate chart which is close to the D-manifold $\mathbf{X}$.

The problem solution $\theta = (h, g)$ consists of two sample-based interrelated mappings: an Embedding mapping $h\colon\ \mathbf{X}_h \to \mathbf{R}^q$ defined on the domain $\mathbf{X}_h \supseteq \mathbf{X}$, and a Reconstruction mapping $g\colon\ \mathbf{Y}_g \subset \mathbf{R}^q \to \mathbf{R}^p$ defined on the domain $\mathbf{Y}_g \supseteq h(\mathbf{X}_h) \supset h(\mathbf{X})$. The solution θ determines the ED-manifold $\mathbf{X}_\theta = \{X = g(y) \in \mathbf{R}^p\colon\ y \in \mathbf{Y}_\theta \subset \mathbf{R}^q\}$ embedded in $\mathbf{R}^p$-dimensional space and covered by a single coordinate chart g defined on space $\mathbf{Y}_\theta = h(\mathbf{X})$ and must ensuring the accurately reconstruction $\mathbf{X}_\theta \approx \mathbf{X}$ of the D-manifold (Manifold proximity property which implies the approximate equalities $X \approx g(h(X))$ for all $X \in \mathbf{X}$).

In [Bernstein & Kuleshov, 2012] an amplification of the ME called the Tangent Bundle Manifold Learning (TBML) is proposed. The TBML problem is to estimate a tangent bundle $\{(X, L(X)), X \in \mathbf{X}\}$ consisting of the points X from the D-manifold $\mathbf{X}$ and the tangent spaces $L(X)$ at these points. The TBML solution (θ, G) includes additionally the sample-based $p \times q$ matrices $G(y),\ y \in \mathbf{Y}_g$, such that the linear space $\mathrm{Span}(G(h(X)))$ accurately reconstructs the tangent space $L(X)$ for all $X \in \mathbf{X}$. A new geometrically motivated algorithm called Grassmann&Stiefel Eigenmaps (GSE) that solves the TBML problem and gives a new solution for the ME problem is proposed also in this paper.

We present also some asymptotic properties of the GSE. In particular, it is proven that under the appropriate chosen GSE parameters there exists a number C_{GSE} such that with high probability (whp) the inequalities

$$\|X - g(h(X))\| \le C_{GSE} \times n^{-\frac{2}{q+2}}$$

for all $X \in \mathbf{X}$ hold true. Here the phrase "an event occurs whp" means that the event occurs with probability at least $(1 - c_\alpha/n^\alpha)$ for any $\alpha > 1$ and c_α depends only on α.

The achieved convergence rate $O(n^{-\frac{2}{q+2}})$ coincides with a minimax lower bound for Hausdorff distance between the manifold and its estimator in the Manifold estimation problem obtained in [Genovese et al. (2012)] under close assumptions. So, GSE has optimal rate of convergence.

Acknowledgment. This work is partially supported byLaboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

References

[Bernstein & Kuleshov, 2012] Bernstein A.V., Kuleshov A.P., 2012: Tangent Bundle Manifold Learning via Grassmann & Stiefel Eigenmaps. arXiv: 1212.6031v1 [cs.LG].

[Genovese et al. (2012)] Genovese Christopher R., Perone-Pacifico Marco, Verdinelli Isabella, Wasserman Larry, 2012: Minimax Manifold Estimation. *JMLR*, **13**, 1263 - 1291.

Empirical Processes in Survey Sampling

CS16A Empirical processes

PATRICE BERTAIL[*,†,¶], STEPHAN CLÉMENÇON[‡], EMILIE CHAUTRU[§]

[*]Université Paris-Ouest-Nanterre-La Défense, Nanterre, France,
[†]CREST-LS, INSEE, Paris, France,
[‡]LTCI-Institut Telecom, Paris, France,
[§]ENSAI, Rennes, France
[¶]email: patrice.bertail@gmail.com

40:PatriceBertail.tex,session:CS16A

This work is devoted to the study of the limit behavior of extensions of the (i.i.d. sample based) empirical process, when the data available have been collected through an explicit survey sampling scheme. Indeed, in many situations, statisticians have at their disposal not only data but also weights arising from some survey sampling plans. These weights correspond either to true inclusion probabilities, as is often the case for institutional data, or to some calibrated or post-stratification weights.

Our main goal is here to investigate how to incorporate the survey scheme into the inference procedure dedicated to the estimation of a probability distribution function $P(dx)$ (viewed as a linear operator acting on a certain class of functions $\mathcal{F}$), in order to guarantee its asymptotic normality. This problem has been addressed in the particular case of a stratified survey sampling, where the individuals are selected at random (with replacement) in each stratum, by means of bootstrap limit results. Our approach is different and follows that of [Hajek (64)] and is applicable to more general sampling surveys, namely those with unequal first order inclusion probabilities which are of the Poisson type or sequential/rejective or close to this sampling plans according to a L1 or an entropy metric. The main result of the paper is a Functional Central Limit Theorem (FCLT) describing the limit behavior of an adequate version of the empirical process indexed by a class of function (referred to as the *Horvitz-Thompson empirical process* throughout the article) in a superpopulation statistical framework. The key argument involved in this asymptotic analysis consists in approximating the distribution of the extended empirical process by that related to a much simpler Poisson sampling plan. The particular case when the inclusion probabilities depend on some auxilliary variable is studied in detail. Statistical applications are considered, by considering Kolmogorov-Smirnov tests in the context of survey data or some Hadamard differentiable functionals including quantiles. We will present some simulation results on the Kolmogorov-Smirnov statistics as well as on confidence intervals for quantiles.

Acknowledgment. This research was partially supported by the LABEX MME-DII.

References

[Hajek (64)] Hajek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population, *Ann. Math. Statist.* **35**,4,1491-1523.

CS2A Stat. Genetics

Approximate Maximum Likelihood Inference for Population Genetics

JOHANNA BERTL[*,§,¶], GREG EWING[†], ANDREAS FUTSCHIK[*,§], CAROLIN KOSIOL[‡,§]

[*]Department of Statistics and Operations Research, University of Vienna, Austria,
[†]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland,
[‡]Institute of Population Genetics, Vetmeduni Vienna, Austria,
[§]Vienna Graduate School of Population Genetics, Austria
[¶]email: johanna.bertl@univie.ac.at 41:JohannaBertl.tex,session:CS2A

The goal of population genetics is to understand how evolutionary forces like mutation, recombination, selection and genetic drift shape the genetic variation within a population. A major challenge in the application of population genetic theory is to infer parameters of the evolutionary history of a population from DNA sequences of present-day samples only: Usually, the number of possible ancestral histories is tremendous and makes the likelihood analytically intractable. However, simulating data under a population genetic model is often straightforward and simulation software is at hand for many different models and scenarii. Therefore, simulation based inference methods such as approximate Bayesian computation (ABC) became popular in recent years.

Here, we propose an algorithm for maximum likelihood estimation that is based on stochastic gradient methods. By moving through the parameter space along an estimated gradient (or alternatively ascent direction) of the likelihood, the algorithm produces a sequence of estimates that will eventually converge to the maximum likelihood estimate (or, equivalently, to the posterior mode). This approach avoids sampling from the whole parameter space that can be very inefficient in high-dimensional problems. The gradient estimates are based on simulated summary statistics. Thus, the method can be applied to various problems that allow for data simulations and efficient computation of informative summary statistics, also outside of the field of population genetics.

To ensure a good performance of the algorithm, we propose tools for tuning and for assessing the convergence of a sequence of estimates. Finally, we complete the method by a strategy to obtain bootstrap confidence intervals. Apart from testing it on simulated coalescent data, we use it to estimate the ancestral history of Bornean and Sumatran orang-utans from DNA sequence data.

Optimal Prediction-Based Estimating Function for COGARCH(1,1) Models

CS11A
SDE-s

ENRICO BIBBONA*, ILIA NEGRI†,‡

*University of Torino,
†University of Bergamo
‡email: ilia.negri@unibg.it 42:IliaNegri.tex,session:CS11A

The COGARCH (COntinuous Generalized Auto-Regressive Conditional Heteroschedastic) model were introduced as a continuous version of the GARCH models by Klüppelberg and others authors in 2004. Indeed the modern treatment of stochastic volatility models is mostly in continuous time aiming at the analysis of high-frequency data. The COGARCH(1,1) model is driven by a general Lévy process trought the equation $dG_t = \sigma_{t-}dL_t$ and the resulting volatility process satisfies the stochastic differential equation $d\sigma_t^2 = (\beta - \eta\sigma_{t-}^2)dt + \phi\sigma_{t-}^2 d[L]_t^d$ where the parameters satisfy $\beta > 0$, $\eta \geq 0$ and $\phi \geq 0$ and $[L]_t^d$ is the discrete part of the quadratic variation of the Lévy process $L = (L_t)_{t\geq 0}$. The main difference between COGARCH models and other stochastic volatility model is that there is only one source of randomness (the Lévy process). These models captures some empirical observed facts about the volatility: that it changes randomly in time, has heavy tails, has cluster on high levels. In Klüppelberg *et al.* (2007), estimation of the model parameters from a sample of equally spaced returns $G_{ir}^{(r)} = G_{ir} - G_{r(i-1)}$ was presented. The methods consists in matching the empirical autocorrelation function and moments to their theoretical counterparts. Under some conditions on the driven Lévy process it turns out that the parameters are identifiable by this estimation procedure and the obtained estimators are strong consistent and asymptotically Normal. In this work prediction-based estimation functions are applied for drawing statistical inference about the COGARCH(1,1) model from discrete observed data. The prediction based estimating functions were introduced by Sørensen (2000) as a generalization of martingale estimation functions. They are based on linear predictors, have some of the most attractive properties of the martingale estimating functions but they only involve unconditional moments, and in general estimators based on conditional moments are more efficient than estimators based on unconditional moments. Choosing suitable predictor spaces it turns out that the proposed estimators can also have high efficiency and moreover an optimal prediction function can be found. To find the optimal prediction function to estimate the parameters of a COGARCH(1,1) model the moments till the order eight are necessary. In this work a general method to calculate the moment of higher order of this process is presented. This method is based on the iterative application of the Itô product formula. With the same method all the joint moments and the conditional moments can be calculated and this is interesting by itself. Under the same conditions assumed in Klüppelberg *et al.* (2007) it is proved that the estimators based on OPBE are consistent and optimal in the sense discussed in Sørensen (2000). Some simulation studies are presented to investigate the empirical quality of the proposed estimator (based on optimal prediction function) compared with the one obtained with the method of moment by Klüppelberg *et al.* (2007).

Acknowledgment. This research was supported by the Italian National Research Fund MIUR09.

References

[Klüppelberg et al. (2007)] Haug, S. and Klüppelberg, C. and Lindner, A. and Zapp, M. Method of moment estimation in the COGARCH(1,1) model, *Econom. J.*, **2**, 320 – 341.

[Sørensen (2000)] Sørensen, M. Prediction-based estimating functions. *Econom. J.* **2**, 123–147.

CS19D Lim. Thms. Processes

Inference on the Covariation of Multi-Dimensional Semimartingales from Discrete Noisy Observations

MARKUS BIBINGER*,†, MARKUS REISS*

*Humbolt-Universität zu Berlin
†email: bibinger@math.hu-berlin.de 43:Bibinger.tex,session:CS19D

We discuss lower and upper bounds for the semiparametric estimation of the integrated covolatility matrix of a multi-dimensional semimartingale in a discrete high-frequency latent observation setup. We propose a local parametric approach to design an asymptotically efficient locally adaptive spectral estimator. A feasible limit theorem is derived for complex models with stochastic volatilities and leverage accounting for market microstructure and non-synchronous observations. We shed light on the form of the asymptotic variance dependent on the covariance structure and for different observation frequencies and noise corruption. A nonparametric estimator for the instantaneous covolatility matrix is obtained implicitly.

OCS32 Valuation in Stoch. Fin.

Evaluating Securitization Portfolios—A Practical Constrained Optimization Problem

ZSOLT BIHARY*

*Mathematical Modeling Center, Morgan Stanley, Budapest, Hungary
44:BihariZsolt.tex,session:OCS32

Securitization takes a pool of risky assets, tranches it up, and issues bonds based on the cash-flow of the pool. The value of the deal depends on the bond sizes and on the bond ratings. High ratings are only possible if the structure passes regulatory stress-tests. Optimal structuring is thus a constrained optimization problem. We will discuss some practical approaches used in tackling such problems. These include surrogate modeling, support vector machines, and the expected improvement method.

CS6E Function Est.

Significance Testing in Quantile Regression

STANISLAV VOLGUSHEV†, MELANIE BIRKE*,§, HOLGER DETTE†, NATALIE NEUMEYER‡

*Universität Bayreuth, Germany,
†Ruhr-Universität-Bochum, Germany,
‡Universität Hamburg, Germany
§email: melanie.birke@uni-bayreuth.de 45:MelanieBirke.tex,session:CS6E

We consider the problem of constructing a test for the hypothesis of the significance of the predictor Z, i. e.

$$\Delta = E[(P(Y \leq q_\tau(X)|X, Z) - \tau)^2 f_Z(Z)],$$

in the nonparametric quantile regression model, which can detect local alternatives converging to the null hypothesis at a parametric rate and at the same time does not depend on the dimension of the predictor Z, such that smoothing with respect to the covariate Z can be avoided. To be precise, the test proposed in this paper can detect alternatives converging to H_0 at any rate $a_n \to 0$ such

that $a_n\sqrt{n} \to \infty$, where n denotes the sample size. Our approach is based on an empirical process $T_n(x, z)$, which estimates the functional

$$T(x,z) = E[(P(Y \leq q_\tau(X)|X,Z)) - \tau)I_{\{X\leq x\}}I_{\{Z\leq z\}}] = E[(I_{\{Y\leq q_\tau(X)\}} - \tau)I_{\{X\leq x\}}I_{\{Z\leq z\}}]$$

for all (x, z) in the support of the distribution of the predictor (X, Z), where the inequality $X \leq x$ between the vectors X and x is understood as the vector of inequalities between the corresponding coordinates and I_A denotes the characteristic function of the event A. We use a quantile estimator proposed in Dette and Volgushev (2008) and calculate a stochastic expansion of the process $T_n(x, z)$ which allows us to obtain the weak convergence of an appropriately scaled and centered version of $T_n(x, z)$ under the null hypothesis, fixed and local alternatives. As a result we obtain a Kolmogorov-Smirnov or a Cramer von Mises type statistic for the hypothesis of the significance of the predictor Z in the nonparametric quantile regression model. Moreover, we are also able to extend the result to the case, where the dimension q of the predictor Z is growing with the sample size, that is $q = q_n \to \infty$ as $n \to \infty$. Finally we investigate a corresponding bootstrap test and show it's finite sample behavior in a simulation study.

Acknowledgment. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1, C4) of the German Research Foundation (DFG).

References

[Dette and Volgushev (2008)] Dette, H. and Volgushev, S., 2008. Non-crossing nonparametric estimates of quantile curves. *Journal of the Royal Statistical Society*, Ser. B **70**, 609-627

[Volgushev et al. (2013)] Volgushev, S., Birke, M., Dette, H. and Neumeyer, N., 2013. Significance Testing in Quantile Regression. *Electronic Journal of Statistics* **7**, 105-145

Detection of Local Network Motifs

IS23 Stat. Genetics, Biol.

ETIENNE BIRMELÉ*

*Laboratoire Statistique et Génome, Evry, France
†email: etienne.birmele@genopole.cnrs.fr 46:EtienneBirmele.tex,session:IS23

Studying the topology of biological networks by using statistical means has become a major field of interest in the last decades. One way to deal with that issue is to consider that networks are built from small functional units called *motifs*, which can be found by looking for small subgraphs which are over-represented in the network.

The first issue in the motif detection problem is the choice of the null model for random graphs. The model used will be the mixture of Erdős-Rényi random graphs, which shows a better fit to the topological properties of biological networks then the scale-free models.

The second issue is the choice of the test statistic. Existing methods all use the overall count of subgraph occurrences in the network. Those relying on sampling algorithms and Z-scores may lead to many false positives as the distribution of the number of small subgraphs is more heavy-tailed than a gaussian. This issue can however be adressed by fitting a Polya-Aeppli distribution to the subgraph occurrence distribution. However, a small graph can appear as over-represented in this framework because it contains an over-represented subgraph, which is in fact the biological relevant structure.

Relying on the fact that motifs in the yeast transcriptional regulatory network are known to aggregate, we will consider another statistic and hence introduce a new definition of a motif: given a small graph $\mathbf{m}$ and an occurence in the network of one of its subgraphs $\mathbf{m}'$, we will look for an over-representation of the number of occurences of $\mathbf{m}$ extending the given occurence of $\mathbf{m}'$. In other words, a motif will be defined by a local over-representation rather than by a global one.

That approach allows us to develop a statistic for the local over-representation of a motif, and to upper-bound its distribution tail using a Poisson approximation. Moreover, a filtering procedure can then be applied to avoid the selection of redundant motifs.

We apply it to the Yeast gene transcriptional regulation network and show that the known biologically relevant motifs are found again and that our method gives some more informations than the existing ones.

Acknowledgment. This research was supported by the French National Research Fund NeMo, grant No.: ANR-08-BLAN-0304-01.

IS5 Envtl. Epidem. Stat.

Using Propensity Score to Adjust for Unmeasured Confounders in Small Area Studies.

WANG YINGBO*, MARTA BLANGIARDO*,‡, NICKY BEST*, SYLVIA RICHARDSON†

*MRC-HPA Centre for Environment and Health, Imperial College London, UK,
†MRC Biostatistics Unit, Cambridge, UK
‡email: m.blangiardo@imperial.ac.uk 47:MartaBlangiardo.tex,session:IS5

Small area studies are commonly used in epidemiology to assess the impact of risk factors on health outcomes when data are available at the aggregated level. However the estimates are often biased due to unmeasured confounders which are not taken into account. Integrating individual level information into area level data in ecological studies may help reduce bias. To investigate this, we develop an area level propensity score (PS) for ecological studies derived from simulated individual-level data and evaluate its performance through several simulation scenarios. This methodological work comprises three steps:

1. We simulate individual level data (e.g. as if from surveys) to obtain information on the potential confounders, which are not measured at the area level. We synthesize these variables accounting for their uncertainty through a Bayesian hierarchical framework and calculate the PS at the ecological level. We conduct a simulation study to assess the impact of the number of areas and individuals on the ecological PS, taking into the account the correlation among the potential confounders.

2. As real survey data are typically characterized by a limited coverage of the territory compared to small area studies, we imputed the ecological PS for the areas with unmeasured confounders. Following Rubin's definition of missing data mechanisms, we assume the missing PS are either MCAR (Missing Complete At Random) or MAR (Missing At Random). For MCAR, we specify a flexible distribution on the PS, by means of a Dirichlet Process. For MAR, we introduce a novel REP (Random Effect Prediction) method to predict the missing PS. MNAR is considered (Missing Not At Random) to be unlikely, given the surveys are typically designed to collect data randomly from the population.

3. We include the imputed and observed PS as a scalar quantity in the regression model linking potential exposure and health outcome. As the PS has no epidemiological interpretation, we specify a spline parameterization to allow for non-linear effects. We compare the most common fixed knots splines with the adaptive splines underpinned by Reversible Jump MCMC.

In the talk, we illustrate each of the three steps, discuss their methodological challenges and show how these affect the performance of PS in the regression model using several simulation scenarios. We conclude that integrating individual level data via PS is a promising method to reduce the bias intrinsic in ecological studies due to unmeasured confounders.

References

[1] McCandless, L., Richardson, S., Best, N. (2012). Adjustment for Missing Confounders Using External Validation Data and Propensity Scores. *JASA*, **107(497)**, 40-51.

Improved Estimation of the Covariance Matrix and the Generalized Variance of a Multivariate Normal Distribution: Some Unifying Results

CS37A Estim. Methods

PANAYIOTIS BOBOTAS[*,‡], STAVROS KOUROUKLIS[†]

[*]Institute of Statistics, RWTH Aachen University, Germany,
[†]Department of Mathematics, University of Patras, Greece
[‡]email: bompotas@stochastik.rwth-aachen.de 48:Panayiotis_Bobotas.tex,session:CS37A

Let X, S_0 be independent statistics, where $X \sim N_p(\mu, \sigma^2 I)$, $S_0 \sim \sigma^2 \chi^2_m$ and $\mu \in \mathbb{R}^p$, σ^2 are unknown parameters. The best affine equivariant (b.a.e.) estimator of σ^2 under the entropy loss $L_1(\delta, \sigma^2) = \delta/\sigma^2 - \ln \delta/\sigma^2 - 1$ is $\delta_0^{\sigma^2} = S_0/m$, which is, however, inadmissible. In the literature there are available various improved scale equivariant estimators of σ^2, i.e., estimators of the form $\delta_\varphi^{\sigma^2} = \varphi(W_0)S_0$, for $W_0 = \|X\|^2/S_0$ and $\varphi(\cdot)$ a positive function on $[0, \infty)$, such as Stein (1964)-type and Brewster and Zidek (1974)-type.

A natural extension of the above problem is to consider estimation of the covariance matrix of a multivariate normal distribution. To this end, let $X_1, \ldots, X_n$ be a random sample from a multivariate normal distribution $N_p(\theta, \Sigma)$, where $\theta \in \mathbb{R}^p$ and $\Sigma > 0$ are unknown and $n \geq p + 1$. Under the entropy loss $L_1(\delta, \Sigma) = \mathrm{tr}(\delta\Sigma^{-1}) - \ln|\delta\Sigma^{-1}| - \mathrm{p}$, the b.a.e. estimator of Σ is $\delta_0 = (n-1)^{-1}S$, where $S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$ and $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. As in the univariate case, δ_0 is inadmissible. Towards improving on δ_0, Kubokawa et al. (1993) considered estimators of the form $\delta_\psi = n^{-1}S + \psi(W)(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'$ for a function $\psi(\cdot)$ and $W = n\bar{X}'S^{-1}\bar{X}$, and extended Stein's (1964) and Brewster and Zidek's (1974) techniques to the multivariate context.

In this work we construct in a very simple way improved estimators of Σ. Let $\delta_\varphi^{\sigma^2} = \varphi(W_0)S_0$ be an arbitrary estimator of σ^2 corresponding to dimension p for the mean vector μ and degrees of freedom $m = n - p$ for S_0 . Set $\psi(t) = \varphi(t) - 1/n$, $t \geq 0$. Then, for this $\psi(\cdot)$, we show that δ_ψ dominates δ_0 if and only if $\delta_\varphi^{\sigma^2}$ dominates $\delta_0^{\sigma^2}$. Consequently, *any* improved estimator of σ^2, no matter what type it is (such as Stein-type, Brewster and Zidek-type, Strawderman-type and Maruyama-type), directly leads to a corresponding improved estimator of Σ, and vice versa. Analogous results are also established for the estimation of Σ under the quadratic loss $L_2(\delta, \Sigma) = \mathrm{tr}(\delta\Sigma^{-1} - \mathrm{I})^2$.

Finally, in a similar way we construct improved estimators of the generalized variance $|\Sigma|$ using improved estimators $\delta_\varphi^{\sigma^2}$ of σ^2 under a general loss satisfying a certain condition.

These novel results reduce, in a specific way, the problems of estimating Σ or $|\Sigma|$ to the univariate problem of estimating σ^2. Also, this work unifies and extends previously obtained results on the estimation of Σ and $|\Sigma|$.

References

[Brewster and Zidek(1974)] Brewster, J. F., Zidek, J. V., 1974: Improving on equivariant estimators. *Ann. Statist.*, **2**, 21–38.

[Kubokawa et al.(1993)] Kubokawa, T., Honda, T., Morita, K., Saleh A.K.Md.E., 1993: Estimating a covariance matrix of a normal distribution with unknown mean. *J. Japan. Statist. Soc.*, **23**, 131–144.

[Stein(1964)] Stein, C., 1964. Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, **16**, 155–160. 190–198.

IS3 Branching Proc.

Conditional Limit Theorems for Intermediately Subcritical Branching Processes in Random Environment

VALERY AFANASYEV*, CHRISTIAN BÖINGHOFF†,‡, GÖTZ KERSTING†, VLADIMIR VATUTIN*

*Steklov Mathematical Institute Moscow, Russia,
†Goethe-University Frankfurt am Main, Germany
‡email: boeinghoff@math.uni-frankfurt.de 49:ChristianBoeinghoff.tex,session:IS3

Branching processes in random environment (BPRE) are a model for the development of a population of individuals, which are exposed to a random environment. More precisely, it is assumed that the offspring distribution of the individuals varies in a random fashion, independently from one generation to the other. Given the environment, all individuals reproduce independently according to the same mechanism.

As it turns out, there is a phase transition within the subcritical regime, i.e. the asymptotics of the survival probability and the behavior of the branching process, conditioned on survival, changes fundamentally. In the talk, we will describe the behavior of the BPRE in the *intermediately subcritical* case, which is at the borderline of the phase transition. In this case, the BPRE conditioned on survival consists of periods with supercritical growth, alternating with subcritical periods of small population sizes. This kind of 'bottleneck' behavior appears under the annealed approach only in the intermediately subcritical case. Our results include the asymptotics of the survival probability and conditional limit theorems for the (properly scaled) random environment and the BPRE, conditioned on non-extinction.

The proofs rely on a construction of the conditioned branching tree going back to Geiger (1999) and Lyons, Pemantle and Peres (1995). Using this construction and a conditional limit theorem for the (properly scaled) random environment, simulations are feasible in the case of geometric offspring distributions. These are used to illustrate our results.

Acknowledgment. Part of this work has been supported by the German Research Foundation (DFG) and the Russian Foundation of Basic Research (RFBF, Grant DFG-RFBR 08-01-91954)

References

[1] V. I. Afanasyev, Ch. Böinghoff, G. Kersting, and V. A. Vatutin. Limit theorems for weakly subcritical branching processes in random environment. *J. Theor. Probab.*, **25** (2012) 703-732

[2] V. I. Afanasyev, Ch. Böinghoff, G. Kersting, and V. A. Vatutin. Conditional limit theorems for intermediately subcritical branching processes in random environment *to appear in Ann. I.H. Poincaré (B)* (2012)

[3] Ch. Böinghoff and G. Kersting. Simulations and a conditional limit theorem for intermediately subcritical branching processes in random environment. *to appear in the Proceedings of the Steklov Institute* (2012)

OCS24 Random Graphs

Svd, Discrepancy, and Regular Structure of Contingency Tables

MARIANNA BOLLA*,†

*Institute of Mathematics, Budapest University of Technology and Economics
†email: marib@math.bme.hu 50:MariannaBolla.tex,session:OCS24

We will use the factors obtained by correspondence analysis to find biclustering of a contingency table such that the row–column cluster pairs are regular, i.e., they have small discrepancy. In our main theorem, the constant of the so-called volume-regularity is related to the SVD of the normalized contingency table. Our result is applicable to two-way cuts when both the rows and columns

are divided into the same number of clusters, thus extending partly the result of [2] estimating the discrepancy of a contingency table by the second largest singular value of the normalized table (one-cluster, rectangular case), and partly the result of [1] for estimating the constant of volume-regularity by the structural eigenvalues and the distances of the corresponding eigen-subspaces of the normalized modularity matrix of an edge-weighted graph (several clusters, symmetric case).

Acknowledgment. This research was partially supported by the Hungarian National Research Fund, grant No.: OTKA-KTIA 77778; further, by the TÁMOP-4.2.2.B-10/1-2010-0009 and the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 projects, latter one supported by the European Union, co-financed by the European Social Fund.

References

[1] Bolla, M. (2011) Spectra and structure of weighted graphs. *Electronic Notes in Discrete Mathematics*, **38**, 149-154.

[2] Butler, S. (2006) Using discrepancy to control singular values for nonnegative matrices. *Linear Algebra and Its Applications*, **419**, 486-493.

Estimating and Detecting Jumps in Functional Processes

IS9 Functional Time Ser.

DENIS BOSQ*,†

*Université Paris 6 - Pierre et Marie Curie, France

†email: denis.bosq@upmc.fr 51:DenisBosq.tex,session:IS9

The purpose of this talk is to estimate and detect constant or random jumps of functional cadlag ("continu à droite, limite à gauche") random processes.

Let $X = (X(t), 0 \leq t \leq 1)$ be a cadlag random process. X can be considered as a D-valued random variable, where $D = D[0,1]$ is equipped with the Skorokhod topology. One observes independent or dependent copies of X and wants to estimate the jump's intensity of X and, if data are collected in discrete time, to detect the position of jumps.

In the first part of the talk we suppose that there exits only one jump at $t_0 \in]0,1[$, where t_0 is constant. Then, under suitable conditions, it is possible to obtain limit theorems for the empirical estimators of $E(X(t_0) - X(t_0-))$ and $E\left|X(t_0) - X(t_0-)\right|$.

The second part is devoted to high frequency data. The situation is more intricate since one has to detect t_0 before estimating the jump's intensity. If the sample paths of X satisfy a Hölder condition on $[0, t_0[$ and on $[t_0, 1]$ one obtains a consistent detector of t_0 and similar asymptotic results as in the continuous case.

Next, we study the case where the jump's position T is random. Again we obtain limit theorems in continuous and discrete time. We also study the asymptotic behavior of the kernel density estimator of T.

If X has two random jumps S and T the problem is to distinguish between the "$S - jump$" and the "$T - jump$". We present a simple discrimination method that leads to consistent estimators of the jumps intensity as well in continuous time or in discrete time.

We apply the above results to the i.i.d. case and to linear D- valued processes, in particular the cadlag moving average process and the cadlag autoregressive process. Finally the case of an infinity of jumps is considered.

CS6B Funct. Est., Regression

Estimation of a Distribution from Data with Small Measurement Errors

ANN-KATHRIN BOTT*,‡, LUC DEVROYE†, MICHAEL KOHLER*

*Technical University Darmstadt, Germany,
†McGill University, Montreal, Canada
‡email: abott@mathematik.tu-darmstadt.de 52:Ann-KathrinBott.tex,session:CS6B

We study the problem of estimation of a distribution from data which contains small measurement errors. Here the only assumption on the measurement errors is that the average absolute measurement error converges to zero for sample size tending to infinity with probability one. In particular we do not assume that the measurement errors are independent with expectation zero. We assume that the distribution, which has to be estimated, has a density with respect to the Lebesgue-Borel measure.

We show that the empirical measure based on the data with measurement error leads to an uniform consistent estimate of the distribution function. Furthermore, we show that in general no estimate is consistent in the total variation sense for all distributions under the above assumptions. However, in case that the average measurement error converges to zero faster than a properly chosen sequence of bandwidths, the total variation error of the distribution estimate corresponding to a kernel density estimate converges to zero for all distributions. In case of a general additive error model we show that this result even holds if only the average measurement error converges to zero. The results are applied in the context of estimation of the density of residuals in a random design regression model where the residual error is not independent from the predictor.

CS6H Copula Estim.

Modelling Unbalanced Clustered Multivariate Survival Data Via Archimedean Copula Functions

ROEL BRAEKERS*,‡, LEEN PRENEN*, LUC DUCHATEAU†

*Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Diepenbeek, Belgium,
†Department of Comparative physiology and biometrics, Universiteit Gent, Gent, Belgium
‡email: roel.braekers@uhasselt.be 53:RoelBraekers.tex,session:CS6H

In an analysis of clustered multivariate survival data, two different types of models are commonly used which take the association between the different lifetimes into account: frailty models and copula models. Frailty models assume that conditional on a common unknown frailty term for each cluster, the hazard function of each individual within that cluster is independent. These unknown frailty terms with their imposed distribution are used to express the association between the different individuals in a cluster. Copula models on the other hand assume that the joint survival function of the individuals within a cluster is given by a copula function, evaluated in the marginal survival function of each individual. Hereby it is the copula function which describes the association between the lifetimes within a cluster.

A major disadvantage of the present copula models over the frailty models is that the size of the different clusters must be small and equal in order to set up manageable estimation procedures for the different parameters in this model. We describe in this talk a copula model for unbalanced clustered survival data. This is done by focusing on the class of Archimedean copula functions with completely monotone generators and exploiting the Laplace transform-expression of these generators to simplify the likelihood function. Hereby we note that the size of each cluster does not have to be equal anymore, and moderate to large cluster sizes are also allowed.

For this model, we develop one- and two-stage procedures to estimate the association parameter for the copula function. In the one-stage procedure, we consider a parametric model for the marginal survival functions while in the two-stage procedure we model the marginal survival functions by either a parametric or a non-parametric model.

As results, we show the consistency and asymptotic normality of the maximum likelihood estimators for the different parameters. We perform a simulation study to investigate the finite sample properties of this estimator and finally we illustrate this copula model on a real life data set in which we study the time until first insemination for cows which are clustered within different herds.

Acknowledgment. In this research support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

Bayesian Temporal Compositional Analysis in Water Quality Monitoring

CS13A Epidem. Models

MARK J. BREWER*,†

*Biomathematics and Statistics Scotland, Aberdeen, United Kingdom
†email: M.Brewer@bioss.ac.uk 54:MarkBrewer-.tex,session:CS13A

An important part of understanding the behaviour of river catchments is determining the geographical sources of water flow and how the variation in relative proportions contributed by different sources changes over time and in response to drivers. The method of end-member mixing analysis (EMMA) has been developed in the hydrological literature [3], often relying on solving mass-balance equations for which assessments of uncertainty are poorly defined. Recent work in the statistical literature has been concerned with Bayesian analysis of compositions [1], and we discuss extensions of that work which model the source distributions directly in the absence of suitable end-member samples.

Our motivating data set was collected for two streams at Loch Ard, central Scotland, over an 18 year period from 1988. Data were recorded approximately weekly until 1996 and fortnightly thereafter, in the form of measurements of alkalinity as tracer and of the rate of flow for both streams. Previous work [4] for this catchment has generated source distributions simply by using the corresponding alkalinity values for the highest and lowest flows, but there are problems with this method: the sample sizes can be very low and as a result, identification of end-member distributions is often unreliable. Our method using Weibull mixed models in combination with kernel density estimation for source identification has proven to be more robust [2].

We conclude there is evidence of a change in source distribution over time; that corresponding to low flow conditions exhibits a gradual increase in alkalinity for both of two streams studied, whereas for high flow conditions alkalinity appeared to be rising for only one stream.

Acknowledgment. This research was partially funded by the Scottish Government's Rural and Environment Science and Analytical Services Division.

References

[1] Brewer, M.J., Filipe, J.A.N., Elston, D.A., Dawson, L.A., Mayes, R.W., Soulsby, C., Dunn, S.M., 2005: A hierarchical model for compositional data analysis, *Journal of Agricultural, Biological and Environmental Statistics*, **10**, 19 - 34.

[2] Brewer, M.J., Tetzlaff, D., Malcolm, I.A., Soulsby, C., 2011: Source distribution modelling for end-member mixing in hydrology, *Environmetrics*, **22**, 921 - 932.

[3] Genereux, D., 1998: Quantifying uncertainty in tracer-based hydrograph separations, *Water Resources Research*, **34**, 915 - 919.

[4] Tetzlaff, D., Malcolm, I.A., Soulsby, C., 2007: Influence of forestry, environmental change and climatic variability on the hydrology, hydrochemistry and residence times of upland catchments, *Journal of Hydrology*, **346**, 93 - 111.

Joint Modeling of Longitudinal and Survival Data: An Application to CASCADE Dataset

CHIARA BROMBIN[*,†], CLELIA DI SERIO[*], PAOLA M. V. RANCOITA[*]

[*]University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Milan, Italy

[†]email: brombin.chiara@hsr.it 55:ChiaraBrombin_2.tex,session:

Disease process over time results from a combination of event history information and longitudinal process. Commonly, separate analyses of longitudinal and survival outcomes are performed. However, disregarding the dependence between these components may lead to misleading results. Separate analyses are difficult to interpret whenever one deals with an observational retrospective multicentre cohort studies where the longitudinally measured biomarkers are poorly monitored, while the survival component involves several sources of bias, multiple end-points, multiple timescales and is affected by informative censoring.

In particular we highlight these aspects within a HIV study, where patients who have been infected are observed until they develop AIDS or die, and condition of their immune system is monitored using markers such as the CD4 lymphocyte count or the estimated viral load.

We discuss how joint models (JM) represent an effective appropriate statistical framework to treat HIV data incorporating all information simultaneously and providing valid and efficient inferences in retrospective studies [1]. We present different approaches for modelling longitudinal and time-to-event data, arising from one of the largest AIDS collaborative cohort studies, the CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe). In particular, we evaluate CD4 lymphocyte progression over time (from the date of seroconversion) and the risk of death in a sample of 648 HIV infected patients enrolled in the Italian cohort.

Along with standard separate models for the outcomes of interest, we consider both parametric and fully Bayesian joint models [2] and joint latent class mixed model [3, 4]. Advantages and disadvantages of the different approaches are discussed. To compare the performances of these models, cross-validation procedures are finally applied.

References

[1] Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **4**, 465-480.

[2] Guo, X. and Carlin, B. (2004). Separate and joint modelling of longitudinal and time-to-event data using standard computer packages. *The American Statistician* **58**, 16-24.

[3] Lin, H., Turnbull, B., McCulloch, C. and Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data. *Journal of the American Statistical Association* **97**, 53-65.

[4] Proust-Lima, C., Joly, P. and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational Statistics & Data Analysis* **53**, 1142-1154.

A Nonparametric Permutation Approach for Assessing Longitudinal Changes in Facial Expression

CS33A Longitudin Data

CHIARA BROMBIN[*,§], LUIGI SALMASO[†], LARA FONTANELLA[‡], LUIGI IPPOLITI[‡]

[*]University Centre of Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Milan, Italy,
[†]Department of Management and Engineering, University of Padova, Italy,
[‡]Department of Economics, University of Chieti and Pescara, Italy
[§]email: brombin.chiara@hsr.it 56:ChiaraBrombin.tex,session:CS33A

Since non-verbal communication plays a crucial role in the overall communicative process, there is a growing interest among researchers and clinicians in analyzing expression of emotion. Several imaging and statistical methods have been developed to recognize and classify facial expression and to quantify emotional expressions.

In this work, we consider the FG-NET Database with Facial Expressions and Emotions from the Technical University Munich. This is an image database containing face images showing a number of subjects performing the six different basic emotions defined by Ekman and Friesen (1971) [1]. More in details, the database contains material gathered from 18 different individuals so far. Each individual performed all six desired actions three times. Additionally three sequences doing no expressions at all are recorded. All together this gives an amount of 399 sequences. Depending on the kind of emotion, a single recorded sequence can take up to several seconds. From each recorded sequence we have extracted 5 frames (times) summarising the dynamic of the expression and on each of the 18 configurations we have manually placed 34 landmarks. We focus only on disgusted, happy, sad and surprised expression.

At first we describe spatio-temporal changes in facial expression using biometric morphing techniques [2, 3]. Then we quantify and test changes of landmarks positions over time using nonparametric combination (NPC) methodology for repeated measurements [4].

The NPC methodology has been showed to provide flexible and robust solutions in complex-repeated measures problems under a set of less-stringent assumptions with respect to standard parametric methods.

References

[1] Ekman, P. and Friesen, W. V. (1971). Constants across culture in the face and emotion. *Journal of Personality and Social Psychology* **17**, 124-129.

[2] Pahuta, M. A., Rohlf, F. J. and Antonyshyn, O. M. (2009) Biometric morphing: A novel technique for the analysis of morphologic outcomes after facial therapy. *Annals of Plastic Surgery* **62**, 48-53.

[3] Brombin, C., Salmaso, L., Ferronato, G., Galzignato, P. (2011) Multi-aspect procedures for paired data with application to biometric morphing. *Communications in Statistics - Simulation and Computation* **40**, 1-12.

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

Linear Regression Analysis in Non-linear Populations

SIL Spec. Invited Lecture

LAWRENCE D. BROWN[*]

[*]Statistics Department, Wharton School, University of Pennsylvania, Philadelphia, US

57:Brown.tex,session:SIL

Linear regression analysis is often applied to populations that do not satisfy the conventional assumptions of linearity, homoscedasticity and normality of residuals. This talk surveys a comprehensive theory of linear regression that is free of any of these classical assumptions. Much of the

talk is "expository" in that it describes results already known. However, some new results will be included related to the structure of statistical errors in such situations. In particular, alternate forms of the familiar Huber-Eicker-White "sandwich" estimator are described.

OCS19 Multivar. funct. data

Estimation of a Linear Data-Driven Ordinary Differential Equations

NICOLAS J-B. BRUNEL[*,‡], QUENTIN CLAIRON[†]

[*]Laboratoire Statistique et Genome, Université d'Evry, France,
[†]Laboratoire Analyse et Probabilités, Université d'Evry, France,
[‡]ENSIIE, Evry, France
[§]email: nicolas.brunel@ensiie.fr 58:NicolasBrunel.tex,session:OCS19

When we observe several (random) curves, for instance several stochastic processes, that shares a common generating mechanism (for instance with different initial conditions), we want to identify a common Ordinary Differential Equation that drives the system for prediction or trend estimation. The interest in determining an approximate Ordinary Differential Equation (ODEs) that can sum up the data generating mechanism, is that the trend can be analyzed within the general framework of ODEs. Moreover, the variability inside the population of curves can be explained by random perturbations (or controls) with respect to the "consensus" ODE. Although we consider only additive perturbations, this ODE framework provides a nonlinear analysis of the variability among the population of curves.

At our level, we consider that the "consensus" ODE is given by a linear vector field, parametrized by a finite dimensional parameter that have to be estimated. We give then a method that can furnish an estimate of the unknown parameter, and also that can give an estimate of the random individual perturbations. Our approach is based on reformulation of the classical statistical least squares criterion as linear-quadratic optimal control, that enables to estimate directly and at the same time the common and the individual parts. Some extensions of the methodology to nonlinear models are also considered.

IS3 Branching Proc.

Societies and Survival in Resource Dependent Branching Processes

F. THOMAS BRUSS[*]

[*]Université Libre de Bruxelles
[†]email: tbruss@ulb.ac.be 59:Thomas_Bruss.tex,session:IS3

Resource Dependent Branching Processes (RDBP's) are models for populations in which individuals need resources and have to create new resources in order to be able to live. Different policies to distribute resources give rise to different societies. For given probability distributions of resource creation and consumption the society form may have a fundamental impact on the survival probability. Can one give reasonable bounds for the survival probability as a function of the chosen society form? Secondly, for which societies can one derive explicit survival criteria? And thirdly, do certain societies stand out by distinguished properties? Several parts of the first question can be answered using earlier work on BP's based on variations of the Borel-Cantelli Lemma (as e.g. B.(1980)), and with the notion of *average unit reproduction mean* (B.(1984)) some answers can be extended to bisexual Galton-Watson processes. The second question is more sophisticated. Using results by Coffman et al. (1987) and B. and Robertson (1991) we can show that this is possible for a reasonably large class of 'structured' societies. The answer to the third question is the Envelope Theorem for societies of B. & Duerinckx (2012)): Under natural hypotheses, there are two extreme societies which determine

the range of survival probabilities for all possible societies. Since the Envelope Theorem is paralleled by explicit criticality criteria for survival of these societies, the result is significant. (This talk is partially based on joint work with M. Duerinckx.)

References

[1] F. T. Bruss, *A Counterpart of the Borel-Cantelli Lemma*, J. Appl. Prob. **17**, 1094-1101, (1980).

[2] F. T. Bruss, *A note on extinction criteria for bisexual Galton-Watson processes*, J. Appl. Prob. **21**, 915-919, (1984).

[3] E. G. Coffman, L. Flatto and R. R. Weber, *Optimal selection of stochastic integrals under a sum constraint*, Adv. Appl. Prob. **19**, 454-473, (1987).

[4] F. T. Bruss and J. Robertson, *'Wald's Lemma' for sums of order statistics of i.i.d. random variables,* Adv. Appl. Prob. **23**, 612-623. (1991).

[5] F. T. Bruss and M. Duerinckx, *Resource dependent branching processes and the envelope of societies,* arXiv: 1212.0693, (2012).

Valid Post-Selection Inference

IS13 Model Selection

RICHARD BERK*, LAWRENCE BROWN*, ANDREAS BUJA*, KAI ZHANG†, LINDA ZHAO*

*The Wharton School, University of Pennsylvania, Philadelphia, USA,
†Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, USA

60:Buja.tex,session:IS13

It is common practice in statistical data analysis to perform data-driven model selection and derive statistical inference from the selected model. Such inference is generally invalid. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffé protection. Importantly it does *not* depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

Acknowledgment. Research supported in part by NSF Grant DMS-1007657.

Catalytic Branching Processes via Hitting Times with Taboo and Bellman-Harris Processes

CS24A Branching Proc.

EKATERINA VL. BULINSKAYA*,†

*Lomonosov Moscow State University, Russia
†email: bulinskaya@yandex.ru

61:Ekaterina_Bulinskaya.tex,session:CS24A

We propose and study the model of *generalized catalytic branching process* (GCBP). It describes a system of particles moving on a finite or countable set S and splitting only in the presence of catalysts. Namely, let at the moment $t = 0$ a particle start the movement viewed as an irreducible continuous-time Markov chain with 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 an exponentially distributed time with parameter 1. Afterwards the particle either splits or leaves w_k with respective probabilities α_k and $1 - \alpha_k$ $(0 < \alpha_k < 1)$.

If the particle splits, it dies producing a random non-negative integer number ξ_k of offsprings. It is assumed that all newly born particles evolve as independent copies of their parent.

When $S = \mathbb{Z}^d$, $d \in \mathbb{N}$, and the Markov chain is a symmetric, homogeneous and irreducible random walk, such model was introduced by S. Albeverio and L. Bogachev in 2000. Another particular case of GCBP was investigated by L. Döring and M. Roberts in 2012 for W consisting of a single catalyst.

For GCBP, as for other branching processes, the natural interesting problem is the analysis of behavior (as $t \to \infty$) of the total and local numbers of particles existing at time t. Here our approach is based on hitting times with taboo and a certain auxiliary indecomposable Bellman-Harris process with $N(N+1)$ types of particles. Note that the number of particles of type $k = 1, \dots, N$ in the auxiliary process at time t coincides with the number of particles located at w_k at time t in GCBP. The total particles number in these processes is the same up to neglecting the particles in GCBP which "go to infinity" and never hit W again. Thus, to study the asymptotic properties of GCBP it is enough to analyze the limit behavior of the specified Bellman-Harris process.

It is well-known that an indecomposable multi-type Bellman-Harris process is classified as supercritical, critical or subcritical if the Perron root of the mean matrix is > 1, $= 1$ or < 1, respectively. We call GCBP supercritical, critical and subcritical according to the classification of the corresponding Bellman-Harris process.

In this way we solve the following problem. Assume that we fix the Markov chain generator A in GCBP and vary "intensities" of catalysts, e.g., $m_k := \mathsf{E}\xi_k$, $k = 1, \dots, N$. What is the set $C \subset \mathbb{R}_+^N$ such that GSBP is critical iff $m = (m_1, \dots, m_N) \in C$? In particular, what is the proportion of "weak" and "powerful" catalysts (possessing small or large values of m_i, respectively). We obtain a complete description of C by means of equation involving the determinant of a specified matrix and indicate the smallest parallelepiped $[0, M_1] \times \ldots \times [0, M_N]$ containing C. One can control the choice of $m \in C$ as follows. Take $m_1^0 \in [0, M_1)$ and consider m in the section $C_{m_1^0} = C \cap \{m : m_1 = m_1^0\}$. Then $m_2 \in [0, M_2(m_1^0)]$. If we take $m_2^0 \in [0, M_2(m_1^0))$ then for m belonging to the section $C_{m_1^0, m_2^0}$ we can claim that $m_3 \in [0, M_3(m_1^0, m_2^0)]$. After the choice of $m_1^0, \dots, m_{N-1}^0$ the value $m_N = m_N^0$ such that $(m_1^0, \dots, m_{N-1}^0, m_N^0) \in C$ is determined uniquely. If for some step $k = 1, \dots, N-1$ we choose $m_k^0 = M_k(m_1^0, \dots, m_{k-1}^0)$ (set $M_1(\varnothing) = M_1$) then $m_i^0 = 0$ for all $i = k+1, \dots, N$. Moreover, if for some $k = 1, \dots, N-1$ we take $m_k > M_k(m_1^0, \dots, m_{k-1}^0)$ then GCBP is supercritical for any $m_{k+1}, \dots, m_N \in \mathbb{R}_+^{N-k}$. Note that if $m_N^0 > 0$ then the choice $(m_1^0, \dots, m_{N-1}^0, m_N)$ with $m_N > m_N^0$ or $m_N < m_N^0$ leads to supercritical or subcritical GCBP, respectively. We also provide the explicit formulae for M_i and $M_k(m_1^0, \dots, m_{k-1}^0)$, $i, k = 1, \dots, N$.

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

IS16 R. Fields, Geom.

Random Fields and Their Applications

ALEXANDER BULINSKI*,†

*Lomonosov Moscow State University, Russia

†email: bulinski@yandex.ru

62:AlexanderBulinski.tex,session:IS16

Random fields are often used as building blocks of various stochastic models. In this regard one can refer, e.g., to a quite recent book [Spodarev (ed.) (2013)]. We consider the problems related to the analysis of random fields trajectories and the asymptotic problems concerning the excursion sets when the windows of observations are growing and the level (or levels) determining these sets is fixed or tends to infinity. Such problems arise, e.g., in material sciences when one tries to investigate the local structure of matter. Here we prove and employ the central limit theorem for specified functionals in random fields. The dependence structure of random fields under consideration plays the

key role. The approach developed in [Bulinski, Shashkin (2007)] permits to study not only the Gaussian random fields. The approximations of initial random fields by auxiliary ones are also widely applied.

After that we will concentrate on the study of models having applications in genetics, see, e.g., [Bulinski et al. (2012)] and references therein. In this research domain we are interested in spatial models describing the relationship between the genotype and phenotype (the environmental explanatory variables are considered as well). Such problems are important for identification of the genetic factors which could increase the risk of complex disease. In this way we prove the results providing the base for dimensionality reduction of observations and discuss the problems of the model selection, optimal in a sense. Several approaches involving cross-validation, random graphs, trees and forests, logic regression and its modifications are in the scope of the talk. Simulation techniques is also tackled.

The final part of the talk is devoted to stochastic models using the spatial point processes. Here we treat the statistical problems related to the population dynamics. Beyond the survey of the state of art and the new established results some open problems are discussed as well.

Acknowledgment. This research was partially supported by the Russian Foundation for Basic Research, grant 13-01-00612.

References

[Spodarev (ed.) (2013)] Spodarev, E. (ed.), 2013: Stochastic Geometry, Spatial Statistics and Random Fields, LNM, **2068**, Springer, Heidelberg.

[Bulinski, Shashkin (2007)] Bulinski, A., Shashkin, A., 2007: Limit Theorems for Associated Random Fields and Related Systems, World Scientific, Singapore.

[Bulinski et al. (2012)] 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*, **2**, 73 - 87.

Ergodic Properties of Strong Solutions of Stochastic McKean-Vlasov Equations

CS23A Diffusions & Diff. Eq.

OLEG BUTKOVSKY*,†

*Moscow State University, Russia

†email: oleg.butkovskiy@gmail.com

63:OlegButkovsky.tex,session:CS23A

Consider the stochastic McKean-Vlasov (SMV) equation in $\mathbb{R}^d$

$$\begin{cases} X_t = X_0 + \int_0^t b(X_s, \mu_s)\, ds + W_t, \quad t \geq 0 \\ \mathrm{Law}(X_t) = \mu_t, \end{cases} \tag{1}$$

where W is a d-dimensional Brownian motion, the initial condition X_0 is a d-dimensional vector that is independent of W, and $b\colon \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$, where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$. These processes were introduced by H.P. McKean, and they naturally arise in the study of the limit behavior of a large number of weakly interacting Markov processes ("propagation of chaos").

If the drift b does not depend on the measure μ, then SMV equation (1) is a stochastic differential equation (SDE). Ergodic properties of solutions of SDEs were studied by Veretennikov, Klokov, Douc, Fort, Guillin and many others. It is known that the Veretennikov-Khasminskii condition on the drift combined with a certain non-degeneracy condition on the diffusion implies existence and uniqueness of the invariant measure and exponential ergodicity.

Using the ideas of Hairer and Mattingly, we extend this result to solutions of SMV equation.

Assume that the drift coefficient b can be decomposed into two terms

$$b(x,\mu) = b_\varepsilon(x,\mu) = b_1(x) + \varepsilon b_2(x,\mu), \quad x \in \mathbb{R}^d,\ \mu \in \mathcal{P}(\mathbb{R}^d), \tag{2}$$

where $\varepsilon > 0$. Assume also that the functions b_1 and b_2 are Lipschitz continuous, i.e.

$$|b_1(x) - b_1(y)| + |b_2(x,\mu) - b_2(y,\nu)| \le L(|x-y| + d_{TV}(\mu,\nu)),\ \ x,y \in \mathbb{R}^d,\ \mu,\nu \in \mathcal{P}(\mathbb{R}^d). \tag{3}$$

for some $L > 0$, where d_{TV} is the total variation distance between two measures. As was shown by Funaki (1984), under these conditions, equation (1) has a unique strong solution $(X_t^\varepsilon, \mu_t^\varepsilon)_{t\ge 0}$.

Theorem 2. *Assume that conditions (2) and (3) hold, and*

1) *The function b_1 satisfies the Veretennikov-Khasminskii condition, i.e., there exist $M > 0$, $r > 0$ such that*
$$\langle b_1(x), x\rangle \le -r|x|, \quad |x| \ge M.$$
Here $\langle\cdot,\cdot\rangle$ is the standard scalar product in $\mathbb{R}^d$.

2) *The function b_2 is uniformly bounded.*

Then there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in [0;\varepsilon_0]$ the strong solution of McKean-Vlasov equation (1) has a unique invariant measure π^ε. Furthermore, for any measure $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ such that $I(\mu_0) := \int_{\mathbb{R}^d} e^x\, \mu_0(dx) < \infty$, one has

$$d_{TV}(\mu_t^\varepsilon, \pi^\varepsilon) \le C(1 + I(\mu_0))e^{-\theta t}, \quad t \ge 0$$

for some positive $C = C(\varepsilon)$, $\theta = \theta(\varepsilon)$.

Acknowledgment. This research was partially supported by the RFBR grant 13-01-00612.

IS2 Bayesian Nonpar.

Nonparametric Bernstein–von Mises Theorems

ISMAËL CASTILLO*,‡, RICHARD NICKL†

*CNRS, LPMA Paris, †Statistical Laboratory, Cambridge

‡email: ismael.castillo@upmc.fr 64:Castillo.tex,session:IS2

We investigate Bernstein-von Mises Theorems in some non-parametric frameworks, such as the Gaussian white noise model. We show that under some mild conditions on the prior distribution on the unknown non-parametric quantity of interest, a suitably rescaled version of the posterior distribution converges weakly to a Gaussian limit, provided weak convergence is stated in a sufficiently large space. Particularly we investigate frequentist coverage properties of Bayesian credible sets. Applications include goodness-of fit tests, as well as general classes of linear and nonlinear functionals.

CS26B Life and Failure Time

On a Generalized Stochastic Failure Model under Random Shocks

JI HWAN CHA*,†

*Ewha Womans University, Seoul, Rep. of Korea

†email: jhcha@ewha.ac.kr 65:JiHwanCha.tex,session:CS26B

Many of the currently used failure models are developed under the premise that the operating environment is static. In these cases, the basic assumption is that the prevailing environmental conditions either do not change in time or, in case they do, have no effect on the deterioration and failure

process of the device. Therefore, in these cases, models not depending on the external environmental conditions are proposed and studied. However, devices often work in varying environments and so their performance is significantly affected by these varying environmental conditions. In this paper, we consider external (environmental) shocks as a cause for system's failure and deterioration. For instance, numerous electronic devices are frequently subject to random shocks caused by fluctuations of unstable electric power. In these cases, the changes in external conditions result either in immediate failure or deterioration of equipment. Shock models usually consider systems that are subject to shocks of random magnitudes at random times. Traditionally, one distinguishes between two major types: cumulative shock models (systems break down because of a cumulative effect) and extreme shock models (systems break down because of one single large shock). These shock models are widely used in many different areas such as reliability, structure and infrastructure engineering, insurance, credit risk, etc. Therefore, along with meaningful mathematical properties, they have significant practical importance and a wide range of applications. In this paper we discuss a reliability model that reflects the dynamic dependency of the system failure on the system stress induced by environmental shock process. Standard assumptions in shock models are that failures of items are related either to the cumulative effect of shocks (cumulative models) or that they are caused by shocks that exceed a certain critical level (extreme shocks models). In this paper, we present useful generalizations of this setting to the case when an item is deteriorating itself, e.g., when the boundary for the fatal shock magnitude is decreasing with time. Mathematically, this approach is based on the consideration of the shot noise process-type stochastic intensity as a model for shocks accumulation.

Model Selection for Relative Density Estimation

CS6C Func. Est., Smoothing

GAËLLE CHAGNY*,†

*Laboratoire MAP5, Université Paris Descartes, Paris, France
†email: gaelle.chagny@parisdescartes.fr 66:GaelleChagny.tex,session:CS6C

The comparison of two samples coming from different populations modeled by two real random variables X and X_0, with cumulative distribution functions (c.d.f.) F and F_0 respectively, is a main challenge in statistics. The associated methods are required to study the differences among groups in various fields (biology or social research for examples).

The aim of this work is to tackle this statistical problem by estimating the relative density of X with respect to X_0, which is a quite recent tool. Assume that f_0, the density of X_0, does not vanish on its support A_0, and denote by F_0^{-1} the inverse of F_0. The relative density is defined as the density of the variable $F_0(X)$ and can be expressed as

$$r(x) = \frac{f \circ F_0^{-1}(x)}{f_0 \circ F_0^{-1}(x)}, \quad x \in F_0(A),$$

where $\circ$ is the composition symbol, f is a density of X, and $A \subset \mathbb{R}$ its support.

Although nonparametric procedures are well-developed for two-samples problems, this function has only been little studied. The most classical methods to compare the c.d.f. F and F_0 are statistical tests such as Kolmogorov and Smirnov, Wilcoxon, or Mann and Whitney tests, which all check the null hypothesis of equal c.d.f.. Another usual tool is the Receiving Operating Characteristic (ROC) curve, which can be defined as the c.d.f. of the variable $1 - F_0(X)$ and is well-known in fields such as signal detection and diagnostic test for example. Kernel-based estimators as well as empirical estimators have been proposed to estimate the ROC curve and also a very similar function, the c.d.f. of $F_0(X)$. It is called the relative distribution function. The relative density r is its derivative. Kernel estimates were defined and studied for this function, from an asymptotic point of view, by Cwik and Mielniczuk (1993) [1] and Molanes-Lopez and Cao (2008) [2].

We propose to build and study a new adaptive estimator for the function r, from two independent samples of variables X and X_0 of size n and n_0 respectively. A collection of projection estimators on linear models is first defined by minimizing a contrast inspired by the classical density contrast, and the quadratic risk is studied: the upper-bound is non trivial, and requires non straightforward splittings. We obtain a bias-variance decomposition which permits to understand what we can expect at best from adaptive estimation: the model selection is then automatically performed in the spirit of the Goldenshluger-Lepski method in a data-driven way. Both nonasymptotic and asymptotic results are derived: an oracle-type inequality shows that adaptation has no cost, and rates of convergence are deduced, for functions r belonging to Besov balls. Such results are new for this estimation problem: especially, no assumption about a link between the sample sizes n and n_0 is required, and the regularity assumptions are not restrictive. Simulation experiments illustrate the method.

References

[1] Cwik, J. and Mielniczuk, J., 1993: Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Probab. Lett.* **16**, no. 5, 397-405.

[2] Molanes-López, E.M. and Cao, R. 2008: Plug-in bandwidth selector for the kernel relative density estimator. *Ann. Inst. Statist. Math.* **60**, no. 2, 273-300.

IS14 Causal Inference

Inference in Targeted Covariate-Adjusted Randomized Clinical Trials

ANTOINE CHAMBAZ[*,§], MARK J. VAN DER LAAN[†], WENJING ZHENG[†,‡]

[*]Université Paris Ouest Nanterre, France,
[†]UC Berkeley, USA,
[‡]Université Paris Descartes, France
[§]email: achambaz@u-paris10.fr

67:AntoineChambaz.tex,session:IS14

Randomized clinical trials (RCTs) are among the most reliable, though still imperfect, sources of causal information. In this talk, we will present the construction and asymptotic study of adaptive covariate-adjusted RCTs analyzed through the prism of the semiparametric methodology of targeted minimum loss estimation (TMLE).

We will show how to build, as the data accrue, a sampling design which targets a user-supplied optimal design. Interestingly, we do not require that the targeted design belong to a class of designs which depend on the baseline covariates only through a summary measure taking finitely many values. We will also show how to carry out a sound TMLE statistical inference based on such an adaptive sampling scheme. The procedure is robust, *i.e.*, consistent even if the working models, parametric or non-parametric, involved in its construction are misspecified. The resulting estimator enjoys an asymptotic linear expansion with known influence curve, from which we deduce its asymptotic normality and an estimator of its asymptotic variance.

We will present the results of a simulation study, comparing the performances of a variety of estimators based on various sampling schemes (adaptive or not).

Acknowledgment. This research was partially supported by a Chateaubriand Fellowship–Science, Technology, Engineering & Mathematics (STEM), and by the grant NIH R01 AI074345-06.

References

[Cartwright (2011)] Cartwright, N. 2011: The art of medicine. A philosopher's view of the long road from RCTs to effectiveness, *The Lancet*, **377**, April 23rd.

[Chambaz and van der Laan (2012)] Chambaz, A., van der Laan, M. J., 2012: Inference in targeted group-sequential covariate-adjusted randomized clinical trials, *Scand. J. Statist.*, to appear.

Application of Sequential Methods to Regression Models When The Number of Effective Variables Is Unknown

CS9A Model Sel, Lin Reg

ZHANFENG WANG*, YUAN-CHIN IVAN CHANG†

*University of Science and Technology of China, Hefei, China,
†Academia Sinica, Taipei, Taiwan
‡email: ycchang@sinica.edu.tw 68:YuanChinChang_-.tex,session:CS9A

As a result of novel data collection technologies, it is now common to encounter data where the number of the collected explanatory variables can be very large, while the number of variables that actually contribute to the model remains small. Thus, a method that can identify those variables with impact on the model without inferring other noneffective ones will make analysis much more efficient. Many methods are proposed to resolve the model selection problems under such circumstances, however, it is still unknown how large a sample size is sufficient to identify those "effective" variables. For a fixed large number of variables considered in the model, we apply sequential sampling method so that the effective variables can be identified sequentially and efficiently. When the sampling is stopped as soon as the "effective" variables are identified, the corresponding regression coefficients are estimated with satisfactory accuracy. This method is new to both sequential estimations and regression models, and can be easily extended to generalized linear models. We consider both fixed and adaptive designs in our study, and prove the asymptotic properties of estimates of the number of effective variables and their coefficients are established under sequential sampling scenario. The proposed sequential estimation procedure is shown to be asymptotically optimal. Simulation studies are conducted to illustrate the performance of the proposed estimation method. We then apply the proposed method to a diabetes data set, and the results will also be reported.

References

[Lai and Wei (1982)] Lai, T. and C. Wei, 1982: Least square estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist. 10*, 154–166.

[Siegmund (1985)] Siegmund, D., 1985: *Sequential Analysis: Tests and Confidence Interval*.

[Willems et al. (1997)] Willems, J., Saunders J., Hunt, D. and Schorling, J. 1997: Prevalence of coronary heart disease risk factors among rural blacks: a community based study. *Southern Medical Journal 90*, 814 – 820.

New York: Springer-Verlag.

[Woodroofe (1982)] Woodroofe, M., 1982: *Nonlinear renewal theory in sequential analysis*. CBMS-NSF regional conference series in applied mathematics.

Clustering in a Random Graph Model with Latent Space

CS28A Random Graphs

ANTOINE CHANNAROND*,†, JEAN-JACQUES DAUDIN*, STÉPHANE ROBIN*

*UMR518 AgroParisTech/INRA
†email: channarond@agroparistech.fr 69:AntoineChannarond.tex,session:CS28A

One way to add heterogeneity to networks consists in allocating positions Z to the nodes in an unobserved latent space. Here the edge between the nodes i and j has probability $f_n(d(Z_i, Z_j))$ where n is the graph size, d is a distance of the latent space and f_n is a decreasing probability function. Thus an edge is even more likely when the nodes are close in the latent space.

Hoff, Raftery and Handcock (2002) provide methods of unsupervised node classification and parameter estimation, the ties being drawn from a logistic regression on the node distance, and the density of the positions being distributed according to a Gaussian mixture. On the other hand in our work we are interested in the non-parametric framework, so that there is no predefined constraint on

the density shape or clustering structure. Here the clusters are defined as connected components of some level set of the density (Hartigan, 1975). The question is to test whether the density is clustered, and more generally, what can be inferred from the observed graph about the clustering structure of the position density in the latent space.

We propose a simple algorithm to estimate the number of such clusters, which generalizes the procedure of Biau, Cadre and Pelletier (2007). It extracts a graph induced by the observed one, which is proved to have as many connected components as the density has clusters, with probability tending to one when n goes to infinity. The algorithm is linear with respect to the number of edges and hence can process large graphs. A simulation study is carried out to illustrate the behavior of the algorithm as a function of n. The effect of some parameters, like intercluster distance or maximal internode connection distance is also illustrated. Furthermore theoretical arguments are provided to prove the consistency.

References

[1] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. ESAIM: P&S, 11:272–280, 2007.

[2] M.S. Handcock, A.E. Raftery, and J.M. Tantrum. Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(2):301–354, 2007.

[3] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. Journal of the American Statistical Association, 97(460):1090–1098, 2002.

IS17 Random Matrices

Large Rectangular Random Matrices with Additive Fixed Rank Deformation

FRANÇOIS CHAPON*,†

*Institut de Mathématiques de Toulouse, Université Paul Sabatier
†email: francois.chapon@math.univ-toulouse.fr 70:FrancoisChapon.tex,session:IS17

We will present the study in the large dimensional regime of the first order and fluctuations behavior of outliers among the singular values of random matrices of the form $XD + P$, where X is a renormalized rectangular Gaussian matrix, D a deterministic matrix, and P a fixed rank deformation matrix. This model finds applications in the fields of signal processing and radio telecommunications, where for instance P represents snapshots of a discrete time radio signal and XD is a temporally correlated and spatially independent noise. This is a joint work with R. Couillet, W. Hachem and X. Mestre.

OCS5 Anal Complex Data

Modelling Multiple Cut-Points for Subset Effect in Clinical Trials: A Bayesian Approach

BINGSHU E. CHEN*,†

*Queens University, Kingston, Canada
†email: bechen@ctg.queensu.ca 71:BingShu.tex,session:OCS5

The objective of this study is to identify the potential time-dependent risk window of miscarriage rate associated with the bivalent vaccine against human papillomavirus (HPV) type 16/18. A novel hierarchical Bayesian method is developed to make statistical inference simultaneously on the threshold and the treatment effect restricted on the sensitive subset defined by the risk window threshold. In the proposed method, the threshold parameter is treated as a random variable that takes values

with certain probability distribution. The observed data are used to estimate parameters in the prior distribution for the threshold, so that the posterior is less dependent on the prior assumption. We then conducted simulation studies to evaluate the performance of the proposed method and estimate the power in detecting the risk effect. Statistical power of the Bayesian approach and existing method (e.g. the permutation test) were compared. The proposed model was illustrated with application to the subpopulation of pregnant women from the Costa Rica Vaccine Trial (CVT).

On a Generalized Multiple-Urn Model

CS19E Lim. Thms.

MAY-RU CHEN*,†

*National Sun Yat-sen University, Taiwan, Republic of China
†email: mayru@faculty.nsysu.edu.tw 72:Chen_May-Ru_NewVersion.tex,session:CS19E

In [Chen et al. (1991)], the authors proposed a new two-urn model with red and white balls and showed that the fraction of red balls in each urn converges to the same limit. In this talk, we extend their works to a three-urn model, which is described as follows. Suppose there are three urns, urn A, urn B and urn C. At the beginning, each urn contains red and white balls. We first draw m_1 balls from urn A and note their colors, say i red and $m_1 - i$ white balls. Then return them into urn A and add bi red and $b(m_1 - i)$ white balls into urn B. Next, we draw m_2 balls from urn B and note their colors, say j red and $m_2 - j$ white balls. Then return them into urn B and add cj red and $c(m_2 - j)$ white balls into urn C. Finally, we draw m_3 balls from urn C and note their colors, say k red and $m_3 - k$ white balls. Then return them into urn C and add ak red and $a(m_3 - k)$ white balls into urn A. Repeat the above action n times and let X_n, Y_n and Z_n be the respective fraction of red balls in urns A, B and C. In this talk, we show some results about the limits of X_n, Y_n and Z_n. We also consider a multiple-urn model with red and white balls and show some results about the limits of the fraction of red balls in each urn.

References

[Chen et al. (1991)] Chen, M.-R. Hsiau, S.-R. and Yaun, T.-H. 2013: A New Two-Urn Model. Accepted by *J. Appl. Prob.*

A Bayesian Approach for Predicting Customers Patronage at a Drug Store

CS12A Hierarchica Bayesian

NAI-HUA CHEN*,‡, YI-TING HWANG†

*Department of Information Management, Chienkuo Technology University, †Department of Statistics, National Taipei University, Taipei, Taiwan
‡email: nhc@cc.ctu.edu.tw 73:NaiHuaChen.tex,session:CS12A

Analyzing customer purchase frequency is important for companies to know and to target the profitable segments and to transform the prospects to valuable ones. The traditional Gamma-Possion form of the negative binomial distribution (NBD) model, adopting conditional probability, proves itself a reasonable tool in predicting customer purchase. However one of the limitations of the NBD model is the purchase frequency does not estimate in each individual-level. We develop a hierarchical Bayesian model of the Poisson likelihood with a gamma distribution. The empirical data was simulated by the Markov Chain Monte Carlo process. Customer purchase frequency of cosmetics for a drug-chain shop is examined. Results show that proposed model can capture the purchase frequency successfully. The top three attributes that effects purchase frequency are package, promotion and brand image.

CS2A Stat. Genetics

An Extended Ancestral Mixture Model for Phylogenetic Inference under the HKY 85 DNA Substitution Model

SHU-CHUAN CHEN[*,‡], JAY TAYLOR[†], MINGZE LI[†]

[*]Idaho State University, Pocatello, ID, USA,
[†]Arizona State University, Tempe, AZ, USA
[‡]email: scchen@isu.edu 74:Shu-ChuanChen.tex,session:CS2A

Inferring phylogenies is important in many scientific areas such as biology, paleontology, biochemistry and bioinformatics. Many statistical methods have been proposed to reconstruct the phylogeny from the DNA sequences. Recently, Ancestral Mixture Models has been proposed by Chen and Lindsay (2006) to infer phylogeny from the binary DNA sequences. It assumes all DNA sequences are evolving under Jukes-Cantor one–parameter model that assumes the transitions and transversions are equal likely, and the frequencies of state A, G, C, T are equal.

In the talk, we will extend the ancestral mixture models to a more general model, HKY85 model, which meets the reality where transitions happens more often than transversions and the frequencies of these four states are often unequal. By varying the time parameter, one can create a hierarchical tree that estimates the population structure at each fixed backward point in time. Theoretical and computational properties and applications of the extended ancestral mixture model will be presented. Comparisons with other existing methods will also be discussed.

CS26A Extremes

A Confidence Region for the Extreme Values of Means of Exponential Populations

CHING-FENG HSU[†], SHUN-YI CHEN[*,§], SU-HAO LEE[‡]

[*]Tamkang University, Tamsui, Taiwan,
[†]Tajen University, Pingtung, Taiwan,
[‡]Lee-Ming Institute of Technology, Taisan, Taiwan
[§]email: sychen@mail.tku.edu.tw 75:ShunYiChen.tex,session:CS26A

We derive a single-stage sampling procedure to yield an optimal confidence interval for the largest (or smallest) location parameter of independent exponential populations in the cases of known and unknown common scale parameter, where "optimal" is defined based on minimal expected interval width and best allocation at a fixed confidence. This is equivalent to obtaining an optimal confidence interval for the largest (or smallest) mean of several exponential populations. The optimal interval will be asymmetric due to the shifting of the interval to the left of a point estimate of the largest mean, and the shifting of the interval to the right of a point estimate of the smallest mean. We also construct a confidence region for the largest and smallest (the extreme values) means of several independent exponential populations. A confidence region for both the largest and the smallest population means reveals the quality of the selected best and the worst one in a ranking and selection procedure. Using numerical methods we obtain the optimal choice of the interval for the largest (or smallest) location parameter when the scale parameter is known and when it is unknown. Tables of percentage points are computed for practitioners to use.

An Alternative Imputation Approach for Incomplete Longitudinal Ordinal Data

CS33A Longitudin Data

YI-JU CHEN*,†

*Tamkang University, Taipei, Taiwan
†email: ychen@stat.tku.edu.tw 76:YIJUCHEN.tex,session:CS33A

Missing data are common occurrences in longitudinal studies. The impact of missing data on the analysis of longitudinal data produces the biased parameter estimates and power reduction. Appropriately handling missing data plays a key role for the analysis of incomplete longitudinal data. The methods for dealing with missing data include complete-case analysis, weighting procedure, available-data analysis and imputation method. Among those methods, available-data analysis and imputation method are more efficient approaches for dealing with missing data. The available-data analysis covers part of the information from subjects who drop out, and provides the availability of estimates that takes incompleteness into account, while imputation methods fill in the missing values, construct a completed data set and then to apply standard methods for data analysis without further developing the new analytical approach.

There are three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). If the missingness does not depend on the values of the data, no matter the observed data or missing data, the missing data mechanism is called MCAR. If the missingness depends only on the observed responses of the data, but not on the missing responses, then it is called MAR. MCAR assumption is more stringent than MAR assumption. The missing data mechanism is called MNAR if the distribution of missing data depends on the missing values in the data. The missing data mechanisms, MCAR and MAR, are of primary interest.

An imputation strategy developed by Demirtas and Hedeker (2008) imputes incomplete longitudinal ordinal data, which converts discrete outcomes to continuous outcomes by generating normal values, employs the multiple imputation method based on normality, and reconverts to binary scale as well as ordinal one. In this work, an alternative imputation approach based on a random generation for ordinal responses is proposed for dealing with incomplete longitudinal ordinal data. Compared to the Demirtas and Hedeker method, the performance of the proposed method is evaluated in terms of standardized bias, root-mean-squared error and coverage percentage under MCAR and MAR by various configurations.

Acknowledgment. This research was partially supported by Taiwan National Science Council, Grant No.: NSC 101-2118-M-032-010.

References

[Demirtas and Hedeker (2008)] Demirtas, H., Hedeker, D., 2008: An imputation strategy for incomplete longitudinal ordianl data, *Statistics in Medicine*, **27**, 4086-4093.

Multivariate Generalized Gamma Mixed-Effects Model for Pharmaceutical Data and its Application to Bioequivalence Test

CS13B Envtl. & Biol. Stat.

YUH-ING CHEN*,†, CHI-SHEN HUANG*

*Institute of Statistics, National Central University, Jhongli 320, Taiwan
†email: ychen@stat.ncu.edu.tw 77:Yuh-Ing_Chen.tex,session:CS13B

In the pharmacokinetic (PK) study under a 2x2 crossover design that involves both the test and reference drugs, we propose a mixed-effects model for the drug concentration-time profiles obtained

from subjects who receive different drugs at different time periods. In the proposed model, the drug concentrations repeatedly measured from the same subject at different time points are distributed according to a multivariate generalized gamma distribution and the drug concentration-time profiles are described by a compartmental PK model with between-subject and within-subject variations. We then suggest a bioequivalence test based on the estimated bioavailability parameters in the proposed model. The results of a Monte Carlo study further show that the proposed model-based bioequivalence test is not only better on maintaining its level, but also more powerful for detecting the bioequivalence of the two drugs than the conventional bioequivalence test based on a non-compartmental analysis suggested by USA FDA or the one based on a well-known mixed-effects model with a normal error variable. The application of the proposed model and test is finally illustrated by using data sets in two PK studies.

CS34A Clinical Studies

Inferiority Index, Margin Function and Non-inferiority Trials with Binary Outcomes

GEORGE YH CHI[*,‡], GARY G KOCH[†]

[*]Janssen Research & Development, Raritan, USA
[†]University of North Carolina, Chapel Hill, USA
[‡]email: gchi@its.jnj.com 78:GeorgeChi.tex,session:CS34A

One of the challenges in the design of non-inferiority trials is in setting the non-inferiority margin. In particular, for non-inferiority trials with binary outcomes, various authors including Röhmel (2001), Garrett (2003), Munk, Skipka and Stratmann (2005), have proposed different functions for defining the margin. Their approaches can be described generally as follows. Let $X_T \sim$ Bernoulli(p_T) and $X_C \sim$ Bernoulli(p_C) denote the binary outcomes of interest for subjects under a treatment T and a control C with Bernoulli distributions, where p_T and p_C are their respective response rates. Let $h(p)$ denote a certain function of the response rate p. Then the margin $\delta_{RD}$for the relative difference measure (RD) is defined as $\delta_{RD} = h(p_C)$. However, their functions $h(p)$ do not appear to take into consideration any potential difference between the variance of p_T and the variance of p_C. It is important to note that the heterogeneity in variance is an intrinsic part of any difference between two Bernoulli distributions. This potential difference in variance is especially pronounced outside the region (0.20, 0.80) where the rate of change in variance is dramatic as the response rate approaches 0 or 1. Since effective anti-infective treatments usually have a response rate exceeding 0.80, it becomes important to take this difference into account in defining the non-inferiority margin.

Li and Chi (2011) introduced the concept of an inferiority index between two distributions and showed how it can be used to define margin function with a specified degree of tightness. In this presentation, we will show how to apply the theory of inferiority index developed for normal distributions to Bernoulli distributions and obtain margin functions for binary effect measures that naturally accommodate the heterogeneity in variance. In addition, it will be shown how equivalent margin functions are derived and equivalent non-inferiority hypotheses can be defined through these equivalent margin functions. Performance of the test statistics associated with these non-inferiority hypotheses are compared and also compared to the classical Wald tests. The results are illustrated with an application to the design of Hospital Acquired Bacterial Pneumonia or Ventilation Associated Bacterial Pneumonia trials.

Acknowledgment. Appreciation to Kim DeWoody and Janssen R&D for their continuing support.

References

[1] Röhmel, J. (2001): Statistical considerations of FDA and CPMP rules for the investigation of new antibacterial products. *Stat. in Med.*, 20, 2561-2571.

[2] Garrett, A.D. (2003): Therapeutic equivalence: fallacies and falsification, *Stat. in Med.*, 22, 741-762.

[3] Munk, A., Skipka, G and Stratmann, B. (2005): Testing general hypotheses under binomial sampling: the two sample case - asymptotic theory and exact procedures, *Comp. Stat. & Data Anal.*, 49, 723-739.

[4] Li, G. and Chi, G. (2011): Inferiority index and margin in non-inferiority trials, *Stat. in Biopharm. Res.*, 3, 288-301.

Asymptotically Unbiased Estimation of Motif Count in Biological Networks From Noisy Subnetwork Data

OCS7 Comp. Biology

KWOK PUI CHOI*,†

*Department of Statistics & Applied Probability, National University of Singapore, Singapore
†email: stackp@nus.edu.sg 79:Choi.tex,session:OCS7

Small over-represented subgraphs in biological networks are thought to represent essential functional units of biological processes. A natural question is to gauge whether a given subgraph occurs abundantly or rarely in a biological network, and so we need an accurate estimate of the number of its occurrences. However, current high-throughput biotechnology is only able to interrogate a portion of the entire biological network with non-negligible errors (i.e., the observed subnetwork contains spurious edges, and some edges are missing due to non-detection). We present a framework, which accounts for the missing and spurious edges, to estimate the number of occurrences of a given subgraph. The estimate, based on uniform sampling scheme, is shown to be "asymptotically" unbiased. Interestingly, (i) we do not need to make any further assumptions of the underlying random network model (such as scale-free, geometric or duplication) in order for the asymptotically unbiased property to hold; and (ii) the estimation of the number of occurrences of subgraph $\mathcal{M}$ is recursively dependent on the estimates of all subgraphs of $\mathcal{M}$.

Regularized LRT for Large Scale Covariance Matrices : One Sample Problem

POSTER Poster

CHI TIM NG*, JOHAN LIM*, YOUNG-GEUN CHOI*,†

*Department of Statistics, Seoul National University, Seoul, Korea
†email: eumjangi@gmail.com 80:YGChoi.tex,session:POSTER

The development of modern technology has made it easier to encounter high-dimensional data in many different disciplines. Examples are genomic data in biology, financial time series data in economics, and natural language processing data in machine learning. For these data, ordinary multivariate procedures, which assume that p (dimension of data) is fixed small, do not fit well, and so a significant amount of research are made to resolve difficulties from the dimension of the data.

This paper studies the inference of covariance matrix from high-dimensional data. To be specific, we propose a regularized likelihood ratio (LR) statistic to test high-dimensional covariance matrix in one sample. In estimation, the sample covariance matrix often becomes poorly conditioned and their eigenvalues spread out when p increase. For this reason, many estimators for large-scale covariance matrix are proposed in the literature. Among many, the linear shrinkage estimator, defined by

$$\widehat{\Sigma}_{\text{lin}} := \alpha\mathbf{S} + (1-\alpha)\mathrm{I}$$

is most popular and widely used for many multivariate procedures as an alternative to the sample covariance matrix $\mathbf{S}$. For example, it defines regularized Hotelling's T-statistic to test the equality of high-dimensional mean vectors (Chen et al. (2011))

In this paper, we propose a regularized likelihood ratio (rLRT) statistic to test high-dimensional covariance matrix in one sample. The statistic is defined by replacing the sample covariance in usual LR statistic with the linear shrinkage estimator :

$$\text{rLRT} := \text{Tr}(\widehat{\Sigma}_{\text{lin}}) - \log|\widehat{\Sigma}_{\text{lin}}| - p = \sum_{i=1}^{p} \big(\psi_i - \log\psi_i - 1\big),$$

where $\psi_i = \psi_\alpha(l_i) = \alpha l_i + (1-\alpha)$, $i = 1, 2, \ldots, p$ and l_i's are the smaple eigenvalues of $\mathbf{S}$. The rLRT is expressed as a functional of empirical spectral distribution of $\widehat{\Sigma}$, so the CLT for linear spectral statistics can be invoked (Bai and Silverstein, 2004). We analytically compute its asymptotic distribution of rLRT under three assumptions of true covariance matrix, (i) identity matrix (I_p), (ii) compound symmetry matrix ($(1-\rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}$), and (iii) a type of singular matrix (block diagonal matrix of I_{p-k} and $\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}}$). The asymptotic is for the case when both n and p increases, and p/n approaches to a constant γ in $(0, 1)$. We apply the asymptotic results to testing the covariance matrix in one sample. In particular, this study focuses on testing whether the covariance matrix is identity or not.

References

[Bai and Silverstein (2004)] Bai, Z.D., and Silverstein, J.W., 2004: CLT for linear spectral statistics of large-dimensional sample covariance matrices, *Ann. Prob.*, **32**, 553 - 605.

[Chen et al. (2011)] Chen, L.S., Paul, D., Prentice, R.L., and Wang., P., 2011: A regularized Hotelling's T^2 test for pathway analysis in proteomic studies, *J. Am. Stat. Assoc.*, **106(496)**, 1345 - 1360.

CS5C H-D Var. Selection

Mixture Model for Designs in High Dimensional Regression and the LASSO

STÉPHANE CHRÉTIEN*,†

*Laboratoire de Mathématiques de Besançon, Université de Franche Comté

†email: stephane.chretien@univ-fcomte.fr 81:Stephane_Chretien.tex,session:CS5C

The LASSO is a recent technique for variable selection in the regression model

$$y \quad = \quad X\beta + \epsilon, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ and ϵ is a centered gaussian i.i.d. noise vector $\mathcal{N}(0, \sigma^2 I)$. The LASSO has been proved to perform exact support recovery for regression vectors when the design matrix satisfies certain algebraic conditions and β is sufficiently sparse. Estimation of the vector $X\beta$ has also extensively been studied for the purpose of prediction under the same algebraic conditions on X and under sufficient sparsity of β. Among many other, the coherence is an index which can be used to study these nice properties of the LASSO. More precisely, a small coherence implies that most sparse vectors, with less nonzero components than the order $n/\log(p)$, can be recovered with high probability if its nonzero components are larger than the order $\sigma\sqrt{\log(p)}$. However, many matrices occuring in practice do not have a small coherence and thus, most results which have appeared in the litterature cannot be applied. The goal of this work is to study a model for which precise results can be obtained. In the proposed model, the columns of the design matrix are drawn from a Gaussian mixture model and the coherence condition is imposed on the much smaller matrix whose columns are the mixture's centers, instead of on X itself. Our main theorem states that $X\beta$ is as well estimated as in the case of small coherence up to a correction parametrized by the maximal variance in the mixture model.

Bayesian Health Monitoring of Farm Animals in Hi-Tech Farm Buildings

OCS11 ENBIS

SHIRLEY Y. COLEMAN[*,‡], MALCOLM FARROW[†]

[*]Industrial Statistics Research Unit, School of Maths and Stats, Newcastle University, Newcastle upon Tyne, UK,
[†]School of Maths and Stats, Newcastle University, Newcastle upon Tyne, UK
[‡]email: shirley.coleman@ncl.ac.uk 82:Coleman.tex,session:OCS11

Modern farm buildings can be equipped for remote data collection for various purposes including the monitoring of livestock. Water intake of farm animals is an important measure of their health. Changes in water consumption are an early indication of poor health and a monitoring system that reliably detects these changes is important. The paper describes work on the water intake of growing pigs with data consisting of intakes in 15-minute intervals over a 6 week period. The data are characterised by marked diurnal variation with frequent zeros in some periods. Bayesian inference was used to develop a model for the normal behaviour. The model was then used to develop an early warning system to detect deviations in the pattern of water intake. The aim is to improve the health and productivity of the pigs. The paper then reports on progress to apply these techniques in the wider context of farm building design, developing innovative algorithms to run in real time monitoring livestock on working farms.

LULU Smoothers on Online Data

CS4A Time Series II.

WILLIE CONRADIE[*,†]

[*]University of Stellenbosch, Stellenbosch, South Africa
[†]email: wjc@sun.ac.za 83:Willie_Conradie.tex,session:CS4A

The smoothing of online monitoring data is an important task, for example in the medical field, important life-related decisions are dependent on the outcome of such data. The basic goal is to extract the underlying, clinically relevant signal from the observed time series. Short-term fluctuations and outliers caused by, for example, the movement of the patient or measurement errors, should be removed, while sudden level shifts and monotonic trends should be preserved. Medical researchers should know which length of block pulse of blood pressure can be ignored, or is dangerous to the patient, and hence the smoothing algorithm can be set to correspond to these limits.

In this paper the compound LULU smoother of Conradie et.al (2009) is investigated to determine its ability to attenuate Gaussian noise, to resist outliers and to preserve abrupt shifts in case of a constant signal and in a trend period via simulations. This smoother belongs to the class of LULU smoothers that was introduced into the mathematical literature by Rohwer (1989). It is showed in the literature in a series of papers, that they have very attractive mathematical properties. These include idempotency-, co-indempotency-, stability-, trend preserving and variation decomposing properties which make them worthy competitors to the well-known nonlinear smoothers found in the literature.

The extent to which the compound LULU smoother reduces Gaussian and spiky impulsive noise and preserves level shifts in case of a constant signal and in a trend period, is determined in this paper. This investigation was motivated by the sound mathematical theory that underlies LULU smoothers as well as their good performance in the comparisons with a number of Tukey's median smoothers in their ability to recover a sinusoidal signal of different frequencies with contaminated normal noise and impulsive noise added.

This paper investigates the unexplored performance of LULU smoothers in handling spiky noise and level shifts under the circumstances mentioned above. Monte Carlo experiments for a comparison of the compound LULU smoother with the running median in basic data situations and in the presence of observational noise were carried out. It was found that the compound LULU smoothers perform extremely well in the cases when the level shift or outliers are in the opposite direction of the trend.

References

[Conradie et.al (2009)] Conradie, W.J., De Wet, T., and Jankowitz, M.D., 2009: Performance of nonlinear smoothers in Signal Recovery, *Applied Stochastic Models in Business and Industry* **25**, 425-444

[Rohwer (1989)] Rohwer, C. H., 1989: Idempotent One-Sided Approximation of Median Smoothers, *Journal of Approximation Theory* **58** (2), 151-163

CS3A Machine learning

Synthetic Data for Multi-label Classification

IVONA CONTARDO-BERNING[*,†], SAREL STEEL[*]

[*]Stellenbosch University, Stellenbosch, South Africa
[†]email: ivona@sun.ac.za 84:Contardo-Berning.tex,session:CS3A

Single label classification is concerned with learning from a set of instances where each instance is associated with a single label from a set of disjoint labels. This includes binary classification, where only two different labels are available, and multi-class classification for cases where more than two labels are available. Multi-label learning problems are concerned with learning from instances where each instance is associated with multiple labels. This is an important problem in applications such as textual classification, music categorization, protein function classification and the semantic classification of images.

Several procedures have been proposed for multi-label classification - see for example Madjarov et al. (2012) for an overview. The relative merits of these procedures are investigated in the literature in terms of their performance on benchmark data sets. This approach may however not give a true picture, since these benchmark data sets most probably do not provide a representative picture of all cases which may occur in practice. According to Luaces et al. (2012) the benchmark data sets used in many papers are fairly similar in their characteristics. It seems therefore that a case can be made for generating artificial multi-label data to use in simulation studies for comparing multi-label classification procedures.

There are only a few papers in the literature dealing with generating synthetic data for multi-label classification: see Luaces et al. (2012). In this paper a brief review of these contributions is presented, followed by a description of a new proposal. We propose generating the indicator matrix of an artificial data set by either using the method of overlapping sums (see Park et al., 1996), or by using appropriately thresholded values generated from a multivariate normal distribution. Both these methods make it possible to introduce correlation between the label indicator vectors. The feature values required to complete the synthetic data set are generated in such a way that the presence or absence of labels in a data case is reflected in the feature values. An attempt is also made to generate feature values in such a way that these values will be suitable to distinguish between relevant and irrelevant variables (with a view to variable selection).

We conclude the paper by applying several multi-label classification procedures to data generated using the new proposal.

References

[Luaces et al. (2012)] Luaces, O., Dìez, J., Del Coz, J.J., Barranquero, J., Bahamonde, A., 2012: Synthetic datasets for sound experimental evaluation of multilabel classifiers, *M-L Group*, Artificial Intelligence Center: Universidad de Oviedo and Gijón. 2012.

[Madjarov et al. (2012)] Madjarov, G.,Kocev, D., Gjorgjevikj, D., Dzđeroski, S., 2012: An extensive experimental comparison of methods for multi-label learning, *Pattern Recognition*, **45**, 3084 - 3104.

[Park et al. (1996)] Park, C.G., Park, T., Shin, D.W, 1996: A simple method for generating correlated binary variates, *The American Statistician*, **50**, 306 - 310.

Modified Progressive Censored Sampling

OCS27 Infer. Censored Sample

COŞKUN KUŞ*,†, YUNUS AKDOĞAN*, AHMET ÇALIK*, ILKAY ALTINDAĞ*, ISMAIL KINACI*

*Selcuk University, Konya, Turkey
†email: coskun@selcuk.edu.tr 85:Coskunkus.tex,session:OCS27

In this paper, modified progressive censored sampling plan is proposed. The maximum likelihood (ML) estimates of the parameters of Weibull distribution are discussed with EM algorithm under introduced sampling plan. Simulation study is implemented to compare the asymptotic variance-covariance matrix, mean square errors(MSEs) and biasses of the ML estimates of proposed scheme with those of progressive censoring scheme. Finally, a numerical example is also given to illustrate the methodology.

Acknowledgment. This research was partially supported by Selcuk University BAP Office.

Using Robust Principal Component Analysis to Define an Early Warning Index of Firms' Over-Indebtedness and Insolvency

CS25C Stoch. Finance II.

G. DAMIANA COSTANZO*,†, DAMIANO B. SILIPO*, MARIANNA SUCCURRO*

*University of Calabria, Rende (CS), Italy
†email: dm.costanzo@unical.it 86:G.DamianaCostanzo.tex,session:CS25C

The aim of the paper is to define an early warning index for firms' insolvency on the base of accounting data taken from the Amadeus database, *Bureau van Dijk*. A correct measure of firm's insolvency level is very important both for potential investors and for management, stockholders and actual or potential firm's competitors. An early warning index could signal a critical level of over-indebtedness behind which the financial status of the firm becomes pathological, therefore very difficult to rehabilitate. Given the international financial crisis and the actual recession of several economies, interesting applications of this analysis could be mentioned so as significant future research developments.

Preliminary steps in the analysis of the firms' over-indebtedness deal with:

1. the definition of a set of variables including several aspects of the indebtedness phenomenon (leverage, indebtedness capacity, form of the financial debt, net financial position, etc.);

2. the setting up of criteria which allow to establish when a firm may be considered over-indebted.

Following this approach, to evaluate the financial condition of the firm we built up a debt index which includes eleven financial ratios; while to measure firm's *debt sustainability*, we built up a second index including three ratios related to the capability of the firm to meet financial obligations with current income. On the base of accounting data taken from the Amadeus database (*Amadeus Correspondence table for Italian companies*), we estimated previous indices by considering a Robust Principal Component Analysis methodology. In fact, it is well known that classical Principal Component Analysis, was originally developed for multivariate normal data, its performance and applicability in real scenarios are limited by a lack of robustness to outlying observations; both the variance (which is being maximized) and the covariance matrix (which is being decomposed) are very sensitive to anomalous observations. Consequently, the first components are often attracted towards outlying points, and may not capture the variation of the regular observations. Therefore, data reduction based on Principal Component Analysis becomes unreliable if outliers are present in the data, and that was the case in our financial data.

We then compared our *early warning index OI*, to some other famous indices in literature as for example the well-known z-score. Since these indices were also defined by using standard statistical methods (i.e. Discriminant Analysis and Logit Models) interesting considerations emerged both from statistical and interpretative point of views.

CsörgőL Csörgő Mem. Lecture

In Memoriam Sándor Csörgő: an Appreciative Glimpse of his Manifold Contributions to Stochastics, a Tribute to my brother Sándor

MIKLÓS CSÖRGŐ[*,†]

[*]School of Mathematics and Statistics, Carleton University, Ottawa, Canada
[†]email: mcsorgo@math.carleton.ca 87:MiklosCsorgo.tex,session:CsorgoL

IS7 Forensic Stat.

Statistical Interpretation of Forensic Glass Evidence

JAMES M. CURRAN[*,†]

[*]Department of Statistics, University of Auckland, New Zealand
[†]email: j.curran@auckland.ac.nz 88:JamesMCurran.tex,session:IS7

The field of statistical evidence interpretation involves objectively evaluating the weight of the evidence for the court. Physcial evidence can take many forms: biological material, fingerprints, bullets, paint, fibres, glass, etc. In this talk I will discuss models for the evaluation of forensic glass evidence. Glass evidence is primarily quantified either by its refractive index, or by its elemental composition. Both of these methods of measurement have their own interesting characteristics which in turn require careful statistical consideration. Glass evidence evaluation is also quite unique in that it is one of the few evidence types that includes assessment of the transfer and persistence mechanisms.

OCS32 Valuation in Stoch. Fin.

How to Detect Asset Bubbles and Crises

ANNA CZENE[*,‡], MIKLÓS LUKÁCS[*,†]

[*]Eötvös Loránd University, Budapest, Hungary,
[†]Corvinus University of Budapest, Hungary
[‡]email: czene.anna@gmail.com 89:Czene_Lukacs.tex,session:OCS32

This talk discusses the modelling of appearance of bubbles in asset valuations. There are at least two major lines in the literature, one relates bubbles to asset price volatility, the other associates

bubbles to trends or drift coefficients. The first lines are best represented by the works of Jarrow and Protter, see [Protter 2012] and references therein, whereas for the other line Ledoit-Johansen-Sornette (LJS) [Ledoit et al. 1999] is a standard starting-point reference, with some more recent works also to mention [Sornette et al. 2012].

Our model is built up similarly to the LJS one, where the growing of a bubble is associated with the drift coefficient, while the price dynamics satisfy a simple stochastic differential equation with drift and jump, and the dynamics of the jumps is governed by the crash hazard rate.

In contrast to the LJS model, where drift and jump intensity are predefined functions of time, independent of the actual price process, in our model the drift and the jump probability (the probability of the burst of the bubble) depends on the price process itself through an exponentially weighted moving average trend. With this moving average we introduce a self exciting growth-factor putting more emphasis on the last moment growth and this way we are able to take into account the greed affecting the strategies of traders.

We calculate the expected value of returns by solving the Backward-Kolmogorov equations and fit the solution to the price series data. The fitted function is a concave curve where the linear term and the curvature parameters are quantified. This way we estimate (as in LJS) those parameters that carry information of the possibility of a crash. Strong empirical evidence exists that the estimated parameters change with an approaching crisis, and the direction of change is always the same. The method was tested on the house price bubble in 2007, further, on many stock price time series like e.g. Apple.

In a neutral market there exists an initial range of parameters and as the bubble grows and reaches an extreme volume before its burst, the parameters suddenly change, and this change can be detected efficiently. After the burst the parameters return quickly to the initial neutral range. The difference in the neutral and extreme position is correlated to the extent of the crash.

Acknowledgment. We are grateful to László Márkus for drawing the topic to our attention, and for the subsequent consultations. We are also very thankful to Zsolt Bihary for regular consultations on the model building.

References

[Protter 2012] Philip Protter. A mathematical theory of financial bubbles. 2012. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2115895

[Ledoit et al. 1999] O. Ledoit, A. Johansen, D. Sornette. Predicting financial crashes using discrete scale invariance. Journal of Risk, 1(4):5–32, 1999.

[Sornette et al. 2012] D. Sornette, W. Yan, R. Woodard. Diagnosis and prediction of market rebounds in financial markets. 2012. Physica A: Statistical Mechanics and its Applications Vol. 391, Issue 4, 2012, Pp.: 1361–1380.

A Restricted Mixture Model for Dietary Pattern Analysis in Small Samples

POSTER
Poster

A. RITA GAIO*,†, JOAQUIM PINTO DA COSTA*

*Departamento de Matemática, Faculdade de Ciências da Universidade do Porto,Portugal, CMUP-Centro de Matemática da Universidade do Porto, Portugal
†email: argaio@fc.up.pt 90:ARitaGaio.tex,session:POSTER

Most of the statistical methods that have been used for the construction of dietary patterns involve factor or principal component analysis, thus exploring the correlations among food groups, and/or combinatorial algorithms of cluster analysis, hence minimizing some predefined cluster criterion. Within these methods, and with the purpose of evaluation of diet-disease relationships, it has

been advocated that a two-step procedure computing covariate-adjusted food consumption residuals before the cluster construction should be followed. This is to ensure that the relationships are specific, and not a reflection of a more general association with total quantity of food consumed. Covariates should include total energy intake and may also consist of other descriptive variables such as age, body weight or body mass index.

Recent studies suggested the use of finite mixture models for the identification of eating habits. In these models, data are viewed as coming from a mixture of probability densities, each representing a different cluster, and, if necessary, both the mixing proportions and the probability densities can be conditioned on covariates of interest.

Mixture modelling brings several advantages over the previous approaches: it allows problems such as the choice of the number of clusters and of the clustering method to be recast as statistical model choice problems; it produces posterior cluster membership probabilities for each subject, given individual food intakes and any other relevant variables, therefore providing measures of uncertainty of the associated classification; it allows for food consumption covariates adjustment simultaneously with the fitting process, instead of the usual two-step procedure from factor analysis mentioned earlier; it allows for pattern prevalence to depend on a set of (concomitant) variables, which avoids biases in the usual three-step procedure given by: finite mixture modelling (with no covariates) $\rightarrow$ classification of individuals based on posterior probabilities $\rightarrow$ multinomial regression analysis relating classes to covariates [A. Rita Gaio et al. (2012)]; if different food group variances are permitted, it becomes scale invariant and thus makes variable standardization redundant; it allows for correlated measurement errors between some or all food groups, by suitable parameterizations of the covariance matrix.

Different parameterizations of covariance matrices may include the diagonal, the eigenvalue decomposition, the intraclass correlation or one-factor model, generalizations of this based on factor analysis and structural equations, autoregressive and other parameterizations common in time series, and models in which covariances are functions of distance in either Euclidean or a deformed space. These different approaches may be impractical in small to moderately sized samples due to the high number of estimated parameters that they require.

An issue that typically arises when analysing dietary data is food non-consumption. Large numbers of food non-consumers result in zero-inflated distributions and estimation difficulties when the food group variables are treated as continuous variables. Possible solutions include dichotomizing food groups whose non-consumption is higher than 50%, a two-part model combining an indicator of food non-consumption with a continuous measurement for consumers, the use of multinomial or censored normal distributions for food group variables, and, for compositional data, the assumption that the data arise from a Euclidean projection of a multivariate Gaussian random variable onto the unit simplex.

Still related to food group distributions and their different shapes is the choice of food group units (e.g. daily servings, daily grams, percent energy) and their treatment (e.g. log transformed, square-root transformed, adjusted for total energy intake, standardized), which vary across studies.

In this work, we present a restricted mixture model approach that greatly reduces the number of model parameters of the generic mixture models. The approach is illustrated with an analysis of Portuguese data from a food-frequency questionnaire, and with a simulation study. It is shown that the model has a good performance in the associated classification, particularly when applied to small samples.

References

[A. Rita Gaio et al. (2012)] A. Rita Gaio, Joaquim Pinto da Costa, Ana Cristina Santos, Elisabete Ramos, Carla Lopes, 2012: A restricted mixture model for dietary pattern analysis in small samples, *Statistics in Medicine*, **31 (19)**, 2137 - 2150.

Phase Synchronization and Cointegration: Bridging Two Theories

OCS30 Stoch. Neurosci.

RAINER DAHLHAUS*,†

*Heidelberg University, Heidelberg, Germany
†email: dahlhaus@statlab.uni-heidelberg.de 91:RainerDahlhaus_2.tex,session:OCS30

In neuroscience the understanding of phase synchronization is of high importance since phase synchronization is regarded as essential for functional coupling of different brain regions. In this talk we point out that, after some mathematical transformation, there exists a close connection between the theories of phase synchronization and cointegration which has been the leading theory in econometrics with powerful applications to macroeconomics during the last decades. As a consequence several techniques on statistical inference for cointegrated systems can immediately be applied for statistical inference on phase synchronization based on empirical data. This includes tests for phase synchronization, tests for unidirectional coupling and the identification of the equilibrium from data including phase shifts. We give an example where a chaotic Rössler-Lorenz system is identified with the methods from cointegration. Cointegration may also be used to investigate phase synchronization in complex networks.

Spectral Density Estimation and Spectrum Based Inference for Nonstationary Processes

OCS3 Spectral Analysis

RAINER DAHLHAUS*,†

*Heidelberg University, Heidelberg, Germany
†email: dahlhaus@statlab.uni-heidelberg.de 92:RainerDahlhaus.tex,session:OCS3

For nonstationary processes there usually exists no unique spectral representation. For locally stationary processes it can be shown that the time varying spectral density exists uniquely in an asymptotic sense. We discuss different estimates of the time varying spectrum and derive it's properties including uniform convergence.

In addition many statistics of interest can be written as a functional of time varying spectral estimates. A key role in the theoretical treatment of these statistics is played by the empirical spectral process. We present a functional central limit theorem for such processes indexed by function spaces, a Bernstein type exponential inequality and a maximal-inequality. Furthermore, we indicate how the above results on the empirical spectral process can be used for the theoretical treatment of inference problems for locally stationary processes.

OCS22 Numeric SPDE

Adaptive Wavelet Methods for the Numerical Treatment of SPDEs

PETRU CIOICA*, STEPHAN DAHLKE*,¶, NICO DÖHRING§, FELIX LINDNER†, THORSTEN RAASCH‡, KLAUS RITTER§, RENÉ SCHILLING†

*Philipps-Universität Marburg,
†TU Dresden,
‡Johannes Gutenberg Universität Mainz,
§TU Kaiserslautern
¶email: dahlke@mathematik.uni-marburg.de 93:StephanDahlke.tex,session:OCS22

The first part of the talk is concerned with the theoretical foundation of adaptive numerical schemes. It is well–known that the order of approximation that can be achieved by adaptive and other nonlinear methods is determined by the regularity of the exact solution in a specific scale of Besov spaces. In contrast, the approximation order of nonadaptive (uniform) methods is determined by the Sobolev smoothness. Therefore, to justify the use of adaptive schemes, sufficiently high Besov smoothness compared to the Sobolev regularity has to be established. We show that for linear stochastic evolution equations in Lipschitz domains the spatial Besov regularity of the solution is commonly much higher than its Sobolev smoothness, so that the use of adaptive schemes is completely justified. In the second part of the talk, we introduce new (spatial) noise models which are based on wavelet expansions. The approach provides an explicit control on the Besov smoothness of the realizations. We study different linear and nonlinear approximation schemes and discuss adaptive wavelet algorithms for stochastic elliptic equations based on these new random functions.

Acknowledgment. This work has been supported by Deutsche Forschungsgemeinschaft, grants DA 360/13-1,2, RI 599/4-1,2, SCHI 419/5-1,2.

IS10 High-Dim. Inference

Indirect Sparsity and Robust Estimation for Linear Models with Unknown Variance

ARNAK S. DALALYAN*,‡, CHEN YIN†

*ENSAE ParisTech, CREST, GENES,
†Imagine, LIGM, Ecole des Ponts ParisTech, France
‡email: arnak.dalalyan@ensae.fr 94:DALALYAN.tex,session:IS10

In this talk, I will present a novel approach to the problem of learning sparse representations in the context of indirect or fused sparsity and unknown noise level. We propose an algorithm, termed Scaled Fused Dantzig Selector (SFDS), that accomplishes the aforementioned learning task by means of a second-order cone program. A special emphasize is put on the particular instance of indirect sparsity corresponding to the learning in presence of outliers. I will present finite sample risk bounds and carry out an experimental evaluation on both synthetic and real data.

The paper can be found at http://hal.archives-ouvertes.fr/hal-00742601/.

Efficient Estimation of the Distribution of Time to Composite Endpoint When Some Endpoints Are Only Partially Observed

IS14 Causal Inference

RHIAN M. DANIEL*,‡, ANASTASIOS A. TSIATIS†

*Department of Medical Statistics and Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, UK, †Department of Statistics, North Carolina State University
‡email: Rhian.Daniel@LSHTM.ac.uk 95:Daniel_Rhian.tex,session:IS14

Two common features of clinical trials, and other longitudinal studies, are (1) a primary interest in composite endpoints, and (2) the problem of subjects withdrawing prematurely from the study. In some settings, withdrawal may only affect observation of some components of the composite endpoint, for example when another component is death, information on which may be available from a national registry. In this work, we use the theory of augmented inverse probability weighted estimating equations to show how such partial information on the composite endpoint for subjects who withdraw from the study can be incorporated in a principled way into the estimation of the distribution of time to composite endpoint, typically leading to increased efficiency without relying on additional assumptions above those that would be made by standard approaches.

On Generalized Multinomial Models and Joint Percentile Estimation

CS40A Logistic & Multinom. Distr.

ISHAPATHIK DAS*,†, SIULI MUKHOPADHYAY*

*Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India
†email: ishapathik@math.iitb.ac.in 96:IshapathikDas.tex,session:CS40A

This article proposes a family of link functions for the multinomial response model. The link family includes the multicategorical logistic link as one of its members. Conditions for the local orthogonality of the link and the regression parameters are given. It is shown that local orthogonality of the parameters in a neighbourhood makes the link family location and scale invariant. Confidence regions for jointly estimating the percentiles based on the parametric family of link functions are also determined. A numerical example based on a combination drug study is used to illustrate the proposed parametric link family and the confidence regions for joint percentile estimation.

Modeling and Analysis of High-Throughput Count Data in Genomics and Proteomics

CS2A Stat. Genetics

SUJAY DATTA*,†

*University of Akron, Akron, USA
†email: sd85@uakron.edu 97:SujayDatta.tex,session:CS2A

High-throughput genomic, epi-genomic and proteomic technologies are now in common use in molecular biology and the biomedical sciences. Some examples are microarrays, serial analysis of gene expression (SAGE), tandem mass spectrometry (MSMS), chromatin immuno-precipitation (ChIP) and next-generation sequencing. Some of these technologies produce large amounts of *count* data (e.g. SAGE, LC-MSMS, RNA-seq) that are noisy, over-dispersed and/or zero-inflated. Conventional statistical methodologies for discrete data do not work well in those cases. Modelling and analysis strategies are needed that take into account the over-dispersion and/ or zero-inflation. As with any high-throughput technology, the choice of an analysis methodology is critical to interpreting the data and understanding the underlying biological mechanism. Here we discuss the

modelling and analyses of three such datasets (a genomic, a proteomic and a next-generation sequencing dataset) and the associated statistical issues. These analyses are motivated by important scientific questions coming from real-life scientific investigations such as a study of the host genomics in bovine *Salmonella* infection, prediction of the abundances of undetected proteins in *Desulfovibrio vulgaris*, etc. Techniques such as parametric and semi-parametric Bayesian hierarchical modeling, zero-inflated Poisson regression and two-stage Poisson model are used to address those questions. Brief summaries of results and conclusions will be presented, time permitting.

CS34A Clinical Studies

Statistical Inference in Dependent Middle Censoring

NASSER DAVARZANI[*,‡], AHMAD PARSIAN[†], RALF PEETERS[*]

[*]Department of Knowledge Engineering, Maastricht University, Maastricht, Netherlands,
[†]School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, IRAN.
[‡]email: n.davarzani@maastrichtuniversity.nl 98:DAVARZANI_Nasser.tex,session:CS34A

Middle censoring refers to data that becomes unobservable if it falls within a random interval. Middle Censoring is a generalization of the existing left censoring, right censoring and double censoring schemes. It was introduced by Jammalamadaka and Mangalam (2003) and recently has been studied in a few papers in which the censoring mechanism is non-informative (independent). In any lifetime study, if the subject is temporarily withdrawn from the study we obtain this middle censoring situation. Practically, this censoring also occurs in clinical trials if the clinic where the observations are being taken is closed for a period (off-period), or when patients are temporarily withdrawn from clinic for some internal or external causes. Usually the association between the survival time (time-to-event) and the beginning of the censoring time is controversial. In some clinical trials, potentially harmful therapies (e.g. radiation or chemotherapy) may have side effects on patients, depending on the toleration of patients and the harmful degree of the therapy. Then patients may need to be withdrawn from the clinic for a period of recovery. In such cases the time of withdrawal depends on the risk of time-to-event (e.g. relapse), and if the event happens within the off-period (censoring interval), we deal with dependent middle censored data.

In this paper, we deal with dependent middle censoring with a censoring interval of fixed length where the lifetime and lower bound of censoring interval are variables with a Marshall-Olkin bivariate exponential distribution. The ability to estimate a survival parameter in the presence of censoring is an important issue which has been studied extensively in recent years. In this set up, we derive a characterization of maximum likelihood estimators for survival parameters, but not in closed form. To compute estimates of the unknown parameters, we use an iterative numerical method. We also propose Bayes estimators of the parameters and the survival function under gamma priors and the squared error loss function in a closed form. A Monte Carlo simulation is carried out to compare these estimators. Finally, using simulated data, we compare the real survival curve with both its ML and Bayes estimates.

References

[Jammalamadaka and Mangalam (2003)] Jammalamadaka, S. Rao and Mangalam, V., 2003: Non-parametric estimation for middle censored data. *Journal of Nonparametric Statistics*. **15**, 253 - 265.

On the Distribution of Infimum of Reflected Processes

CS24A Branching Proc.

KRZYSZTOF DĘBICKI[*,§], KAMIL KOSIŃSKI[†,‡], MICHEL MANDJES[†,‡]

[*]Mathematical Institute, University of Wroclaw, Wroclaw, Poland,
[†]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands,
[‡]EURANDOM,Eindhoven University of Technology
[§]email: debicki@math.uni.wroc.pl 99:KrzysztofDebicki.tex,session:CS24A

In the talk we will consider both Lévy and Gaussian driven queues (i.e. processes reflected at 0), and focus on the distribution of $M(t)$, that is, the minimal value attained in an interval of length t (where it is assumed that the queue is in stationarity at the beginning of the interval).

In the first part of the talk, that deals with Lévy queues, we will provide explicit characterization of this distribution, in terms of Laplace transforms, for spectrally one-sided Lévy processes (i.e., either only positive jumps or only negative jumps). Then, both for Gaussian and Lévy case, we will analyze the asymptotics of $P(M(T_u) > u)$ (for different classes of functions T_u and u large).

Poisson Random Sampling of Almost Periodic Processes and Circular Bootstrap Method

OCS26 Resampling Nonstat T.S.

DOMINIQUE DEHAY[*,‡], ANNA DUDEK[†]

[*]Institut de Recherche Mathématique de Rennes, UMR CNRS 6625, Université Rennes 2, Rennes, France,
[†]AGH University of Science and Technology, Krakow, Poland.
[‡]email: dominique.dehay@univ-rennes2.fr 100:DEHAY.tex,session:OCS26

Let $\{X(t), t \in \mathbb{R}\}$ be an almost periodically correlated process and $\{N(t), t \geq 0\}$ be a homogeneous Poisson process (also denoted by its time jumps $\{t_k : k \geq 1\}$). We assume that $\{X(t), t \in \mathbb{R}\}$ and $\{N(t), t \geq 0\}$ are independent. Moreover, the process $\{X_t : t \in \mathbb{R}\}$ is not observed continuously. Only the time series $\{Z_k\} = \{X_{t_k}\}$ is observed. In this work we focus on the estimation of the cyclic mean of $\{X(t), t \in \mathbb{R}\}$. The asymptotic normality of the estimator is shown. Additionally, the bootstrap method based on circular block bootstrap is proposed. The consistency of the bootstrap technique is proved and the bootstrap quantiles for the normalized errors of estimation are computed. A simulation data example is also provided.

Exact and Limit Laws for Precedence Tests

IS14 Causal Inference

PAUL DEHEUVELS[*,†]

[*]LSTA - University of Paris 6, 7 avenue du Château, F92340 Bourg-la-Reine, France
[†]email: paul.deheuvels@upmc.fr 101:PaulDeheuvels.tex,session:IS14

We consider two-sample precedence-type tests based on the number of observations of the first sample less than the k-th order statistic in the second sample. Such tests have been discussed at length (refer to the monograph of Balakrishnan and Ng (2006)) in the literature since their introduction by Mosteller (1948). Besides exact formulæ for the distribution function of these statistics (under the assumption of equal continuous distributions), we show that their limit law when the sample sizes increase to infinity, with their ratio tending to a limit, for a fixed k, is negative binomial. We extend these results to the case where k depends upon the sample sizes.

CS19C Lim. Thms. Sums of RVs

Variance of Partial Sums of Stationary Processes

GEORGE DELIGIANNIDIS*,†

*Department of Statistics, University of Oxford, Oxford UK

†email: deligian@stats.ox.ac.uk 102:Deligiannidis.tex,session:CS19C

We give necessary and sufficient conditions for the variance of the partial sums of stationary processes to be regularly varying in terms of the spectral measure associated with the shift operator.

Similar results will be given for functionals of stationary reversible Markov chains, using Tauberian theory. Finally for stationary Markov chains with normal transition operator, the spectral measure of the Markov transition operator is linked to that of the shift operator through the use of the Poisson kernel, and necessary and sufficient conditions are expressed in terms of harmonic measure.

This is joint work with S. Utev(University of Nottingham, UK) and M. Peligrad (University of Cincinnati, USA).

CS26A Extremes

Heavy-Tailed Random Fields and Tail Index Estimation

STÉPHAN CLÉMENÇON*, ANTOINE DEMATTEO*,†,‡

*Télécom ParisTech, Paris, France,

†Gaz Transport & Technigaz, Saint-Remy lès cheuvreuse, France

‡email: antoine.dematteo@telecom-paristech.fr 103:AntoineDematteo.tex,session:CS26A

The objective of this paper is to provide a sound theoretical framework for risk assessment, when dangerous events coincide with the occurrence of extremal values of a (continuous) random field with an intrinsic regularly varying behaviour.

The dataset is provided with the courtesy of GTT(Gaz Transport & Technigaz), a society designing the insulation membrane for the tanks of the ships conveying liquefied natural gas (LNG) at $-163°\mathrm{C}$. When the ships are set into motion they sometimes experience an hydrodynamic phenomena called *sloshing* that leads to the creation of waves of LNG that hit the tank walls and might damage the insulation membrane. GTT studies sloshing pressure loads by means of small scale tanks (1/40) instrumented with arrays of sensors.

The theory of regular variations provides a non asymptotic semi-parametric framework, making possible an appropriate description of heavy-tail phenomena. In risk assessment, this conservative approach avoids underestimating the probability of occurrence of extreme events and is the main mathematical tool to carry out worst-case risk analyses in various fields such as meteorology, finance or insurance. Yet, a general theory for spatial processes with intrinsic marginal regularly varying behaviour has not been developed so far, to the best of our knowledge. The present work extends the concept of (multivariate) regular variations to the spatial setup and generalises the heavy-tailed property for multivariate random vectors to random fields.

The key parameter of heavy-tailed modelling is the tail index. In the univariate setup, the celebrated Hill estimator is commonly used and has been extended to a wide variety of data, especially to mixing-sequences. A significant contribution of this work is to propose a new aggregated tail-index estimator for Heavy-tailed random fields as a combination of the marginal Hill estimators of the tail index. Formally, suppose the field is strictly stationary and is observed on a discrete grid, say at the locations $s_1, \ldots, s_d$. Denote by α the tail index of the margins of the field and by $H_k^{(i)}$ the hill estimator at the location s_i. In the case when the observations at two different locations are independent, the asymptotic distribution of any combination of the marginal Hill estimators is straightforward to derive. However, in the presence of dependences, which will generally be the

case, the derivation of the asymptotic distribution of the aggregated estimator is much harder. We show that $\sqrt{k}\left(H_k^{(1)} - 1/\alpha, \ldots, H_k^{(d)} - 1/\alpha\right)$ is asymptotically Gaussian and we compute the asymptotic covariance matrix. We use this result to derive the asymptotic properties of the aggregated estimator. This result enables to combine all the observations of the field at all the locations, even in the case of dependencies, and to obtain a much more accurate estimation of the tail index.

This work make possible an efficient modelling of continuous spatial processes with intrinsic marginal heavy-tailed behaviour together with a very accurate estimation of the tail index, taking into account the dependencies between the observations. It is successfully applied on GTT's dataset.

References

[Resnick (2007)] Resnick, S. , 2007: Heavy-tail phenomena: probabilistic and statistical modeling, *Springer, New York*.

Covariance Estimate for Indicator Functions of Associated Random Variables and Applications

CS20A R. Fields & Geom.

VADIM DEMICHEV*,†

*Lomonosov Moscow State University, Moscow, Russia
†email: vadim.demichev@gmail.com

104:DemichevVadim.tex,session:CS20A

We study families of positively or negatively associated random variables (see the definitions and a number of examples in [Bulinski, Shashkin (2007)]).

Theorem 1. *Let (X, Y) be a positively or negatively associated square integrable random vector with values in $\mathbb{R}^2$. Suppose that X and Y have bounded densities p_X and p_Y respectively. Then the following inequality holds*

$$\sup_{x,y\in\mathbb{R}} |cov(\mathbb{I}\{X > x\}, \mathbb{I}\{Y > y\})| \leqslant C_1 f^{inv}(\min\{\|p_X\|_\infty \|p_Y\|_\infty |cov(X, Y)|, 1\}). \quad (1)$$

Here C_1 is a positive constant and $f(0) = 0$, $f(x) = x^2/\log(e/x)$, $x \in (0, 1]$.

The optimality of this estimate is demonstrated by

Theorem 2. *For any $r \in (0, 1]$ one can find a positively associated square integrable random vector (X, Y) with values in $\mathbb{R}^2$, such that $cov(X, Y) \in (0, r]$, $\|p_X\|_\infty = \|p_Y\|_\infty = 1$ and*

$$\sup_{x,y\in\mathbb{R}} cov(\mathbb{I}\{X > x\}, \mathbb{I}\{Y > y\}) \geqslant C_2 f^{inv}(cov(X, Y))$$

where $C_2 > 0$ is a constant.

Consider now an application to the theory of excursion sets. Let $\xi = \{\xi_t,\ t \in \mathbb{R}^d\}$ be a measurable strictly stationary associated random field with continuous covariance function $r(t) = cov(\xi_0, \xi_t)$, $t \in \mathbb{R}^d$. For $u \in \mathbb{R}$ and $n = (n_1, \ldots, n_d) \in \mathbb{N}^d$ introduce

$$S_n(t) = \langle n\rangle^{-1/2} \int_{E_{n,t}} (\mathbb{I}\{\xi_x > u\} - \mathsf{P}(\xi_0 > u))dx, \quad t = (t_1, \ldots, t_d) \in [0, 1]^d,$$

here $\langle n\rangle = n_1 \cdot \ldots \cdot n_d$ and $E_{n,t} = [0, n_1 t_1] \times \cdots \times [0, n_d t_d]$. In other words $S_n(t)$ is the centered and normalized excursion set volume of the field ξ on $E_{n,t}$ at level u. Using (1) we establish (cf. [Bulinski et al. (2012)])

Theorem 3. *Let ξ_0 have a bounded density. Suppose also that $r(t) = O(\|t\|^{-\alpha})$, $\alpha > 2d$. Then for any*

sequence $n^{(k)} = (n_1^{(k)}, \dots, n_d^{(k)})$ *with* $n_i^{(k)} \to \infty$, $i = 1, \dots, d$, *the random fields* $S_{n^{(k)}}$ *converge weakly in the space* $C([0,1]^d)$ *as* $k \to \infty$ *to the Gaussian random field* $\sigma_u W$ *where*

$$\sigma_u^2 = \int_{\mathbb{R}^d} cov(\mathbb{I}\{\xi_0 > u\}, \mathbb{I}\{\xi_x > u\})dx$$

and W *is a* d*-dimensional Brownian motion.*

References

[Bulinski, Shashkin (2007)] Bulinski, A. V., Shashkin, A. P., 2007: *Limit Theorems for Associated Random Fields and Related Systems*, World Scientific, Singapore.

[Bulinski et al. (2012)] Bulinski, A., Spodarev, E., Timmermann, F., 2012: Central limit theorems for the excursion set volumes of weakly dependent random fields, *Bernoulli*, **18**:1, 100-118.

IS21 Spatial Point Proc.

Consistency of the Maximum Likelihood Estimator for General Gibbs Point Process

DEREUDRE DAVID[*,‡], LAVANCIER FRÉDÉRIC[†]

[*]Laboratoire Paul Painlevé, Université Lille1, France,
[†]Laboratoire Jean Leray, Université de Nantes, France
[‡]email: david.dereudre@univ-lille1.fr 105:DEREUDRE.tex,session:IS21

We present a general result about the consistency of the maximum likelihood estimator for Gibbs Point Processes. In our setting, consistency means the almost sure convergence of the estimator to the true parameter when the observation window goes to the full space.

During the talk, a particular interest will be given to the Strauss model. Let us recall that its conditional distribution, in any compact set $\Lambda \subset \mathbb{R}^d$ and given the configuration φ_{Λ^c}, is absolutely continuous with respect to the homogeneous Poisson Point Process in Λ with the density

$$f_\Lambda(\varphi) = C_\Lambda(\varphi_{\Lambda^c}).z^{N_\Lambda(\varphi)}\gamma^{S_\Lambda(\varphi)},$$

where $C_\Lambda(\varphi_{\Lambda^c})$ is a normalization constant, $N_\Lambda(\varphi)$ is the number of points of φ in Λ and

$$S_\Lambda(\varphi) = \sum_{\substack{\{x,y\}\in\varphi \\ \{x,y\}\cap\Lambda\neq\emptyset}} \mathbf{1}_{|x-y|\leq R}.$$

Our Result allows to prove the simultaneous consistency of the Maximum Likelihood Estimator of z, γ, R although the density is discontinuous with respect to R.

In a general setting, let us consider a Gibbs Point Process P in $\mathbb{R}^d$ for an interaction which depends on a parameter θ in a compact set Θ in $\mathbb{R}^p$. Let us consider a sequence of finite windows Λ_n in $\mathbb{R}^d$ which converges to the full space $\mathbb{R}^d$. Let Φ be a realisation of P in $\mathbb{R}^d$ and $\hat{\theta}_n(\Phi)$ be the Maximum Likelihood Estimator in Λ_n based on Φ_{Λ_n} (the restriction of Φ in Λ_n). Our main result claims that under suitable assumptions, for P-almost every realization Φ, the estimator $\hat{\theta}_n(\Phi)$ converges to θ.

The main technical assumption is a variational principle in the spirit of statistical physics. It means that the Gibbs Point Processes are exactly the minimizers of the free energy. This assumption is satisfied in several models [Georgii (1994), Dereudre and Georgii (2009)]. The other assumptions are standard and, in particular, the regularity of the interaction with respect to the parameters is not required.

The scheme of proof was first proposed in [Mase (2000)] for exponential models. It means that the interaction depends linearly on the parameters. The originality of our work is, therefore, to provide a general proof of the consistency of the MLE when the interaction depends arbitrary on the parameters.

References

[Mase (2000)] Mase, S., 2000: Consistency of MLE's of Gibbs Distribution, *preprint*, 1 - 11.

[Georgii (1994)] Georgii, H.-O. 1994: Large deviations and the equivalence of ensembles for Gibbsian particle systems with superstable interaction, *Probab. Theory Relat. Fields*, **99** 171-195.

[Dereudre and Georgii (2009)] Dereudre, D and Georgii, H.-O. 2009: Variational principle of Gibbs measures with Delaunay triangle interaction, *Electron. J. prob.*, **14** 2438-2462.

Symbolic Representation of Non-Central Wishart Random Matrices with Applications

CS22A R. Matrices

ELVIRA DI NARDO*,†

*University of Basilicata, Potenza, Italy

†email: elvira.dinardo@unibas.it 106:ElviraDiNardo.tex,session:CS22A

Let $\{X_{(1)}, \ldots, X_{(n)}\}$ be random row vectors independently drawn from a p-variate normal distribution with zero mean $\mathbf{0}$ and full rank covariance matrix Σ. Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n$ be row vectors of dimension p. The *non-central Wishart distribution* is the distribution of the square random matrix of order p

$$W_p(n, \Sigma, M) = \sum_{i=1}^{n} (X_{(i)} - \boldsymbol{m}_i)^T (X_{(i)} - \boldsymbol{m}_i), \qquad \text{with } M = \sum_{i=1}^{n} \boldsymbol{m}_i^T \boldsymbol{m}_i.$$

The matrix $\Omega = \Sigma^{-1} M$ is the non-centrality matrix and it is usually used instead of M to parametrize the Wishart distribution.

Wishart distributions have applications in several areas: a good review of fields, within these distributions are successfully employed, is given in [Withers et al. (2012)]. In parallel to its spread in the applications, also its mathematical properties including determinant, eigenvalues and other characterizations have received great attention in the literature. In particular the derivation of explicit expressions for moments and cumulants is still object of in-depth analysis and the computation of joint moments

$$E\left\{ \mathrm{Tr}\left[W(n) H_1\right]^{i_1} \cdots \mathrm{Tr}\left[W(n) H_m\right]^{i_m} \right\} \qquad \text{with} \quad H_1, \ldots, H_m \in \mathbb{C}^{p \times p} \tag{1}$$

is a very general task: for example, when $H_1, \ldots, H_m$ are sparse matrices, equation (1) returns joint moments of entries of $W_p(n, \Sigma, M)$.

The aim of this contribution is twofold: to give a closed form formula to compute joint moments (1) and to introduce a symbolic method, known in the literature as the classical umbral calculus, as a natural way to deal with moments of Wishart distributions. This method allows us to manage moment sequences without making hypothesis on the existence of an underlying distribution probability. Many results rely on the representation of the non-central Wishart random matrix as convolution of its central component and a matrix of formal variables (or symbols), whose entries are *uncorrelated* with those of the central component. In order to take into account the cyclic property of traces, the notion of necklace is fruitfully employed in their computation allowing to set up an efficient symbolic procedure. The algorithm in `Maple 12` is available on demand. Comparisons with the other techniques proposed for computing moments of non-central Wishart distributions are also given.

References

[Withers et al. (2012)] Withers, C.S., Nadarajah, S., 2012: Moments and cumulants for the complex Wishart. *J. Mult. Anal.*, **112**, 242 - 247.

CS5A
H-D Dim. Reduction

Simultaneous Statistical Inference in Dynamic Factor Models

THORSTEN DICKHAUS*,†

*Humboldt-University, Berlin, Germany

†email: dickhaus@math.hu-berlin.de 107:Dickhaus.tex,session:CS5A

Dynamic factor models (DFMs) are popular tools in econometrics to describe the dynamics of a multi-dimensional time series by a lower-dimensional set of (possibly latent) common factors.

Assume that a p-dimensional, covariance-stationary stochastic process $\mathbf{X} = (\mathbf{X}(t) : 1 \leq t \leq T)$ can be observed in discrete time. Then, a DFM takes the form

$$\mathbf{X}(t) = \sum_{s=-\infty}^{\infty} \Lambda(s)\, \mathbf{f}(t-s) + \varepsilon(t),\ 1 \leq t \leq T.$$

Thereby, $\mathbf{f}(t) = (f_1(t), \ldots, f_k(t))^\top$ with $k < p$ denotes a k-dimensional vector of common factors and $\varepsilon(t) = (\varepsilon_1(t), \ldots, \varepsilon_p(t))^\top$ denotes a p-dimensional vector of specific factors, to be regarded as error or remainder terms. The entry (i,j) of the matrix $\Lambda(s)$ quantitatively reflects the influence of the j-th common factor at lead or lag s, respectively, on the i-th component of $\mathbf{X}(t)$, where $1 \leq i \leq p$ and $1 \leq j \leq k$.

Based on the theory of multiple statistical hypothesis testing, we elaborate simultaneous statistical inference methods in such models. Specifically, we employ structural properties of multivariate chi-square distributions in order to construct critical regions for vectors of Wald statistics for testing linear hypotheses regarding parameters of the frequency-domain representation of the model. The autocovariance function of the observable process $\mathbf{X}$, $\Gamma_{\mathbf{X}}$ for short, and its spectral density matrix $S_{\mathbf{X}}$ (say), can be expressed as

$$\begin{aligned}
\Gamma_{\mathbf{X}}(u) &= \mathbb{E}[\mathbf{X}(t)\mathbf{X}(t+u)^\top] = \sum_{s=-\infty}^{\infty} \Lambda(s) \sum_{v=-\infty}^{\infty} \Gamma_{\mathbf{f}}(u+s-v)\Lambda(v)^\top + \Gamma_{\varepsilon}(u), \\
S_{\mathbf{X}}(\omega) &= (2\pi)^{-1} \sum_{u=-\infty}^{\infty} \Gamma_{\mathbf{X}}(u) \exp(-i\omega u) \\
&= \tilde{\Lambda}(\omega) S_{\mathbf{f}}(\omega) \tilde{\Lambda}(\omega)' + S_{\varepsilon}(\omega),\ -\pi \leq \omega \leq \pi,
\end{aligned}$$

where $\tilde{\Lambda}(\omega) = \sum_{s=-\infty}^{\infty} \Lambda(s) \exp(-i\omega s)$ and the prime stands for transposition and conjugation. A localization technique allows to apply likelihood-based inference methods to the model, assuming that the model is identified, model restrictions are testable, and the sample size T is large. In this, we make use of the asymptotic distribution of the vector of test statistics.

Examples of important multiple test problems arising in dynamic factor models demonstrate the relevance of the proposed methods for practical applications. In particular, we demonstrate how the problems "Which of the specific factors have a non-trivial autocorrelation structure?" and "Which of the common factors have a lagged influence on $\mathbf{X}$?" can be addressed by our methodology.

Performing Model Selection in Mixture Cure Models for the Analysis of Credit Risk Data

CS9B Model Sel, Info Crit

LORE DIRICK*,†, GERDA CLAESKENS*, BART BAESENS*

*ORSTAT and Leuven Statistics Research Center, KU Leuven, Leuven, Belgium
†email: Lore.Dirick@kuleuven.be 108:LoreDirick.tex,session:CS9B

In recent research, several survival analysis techniques (originally mainly used in medical science) are being used to analyze credit loan information. The advantage of this method compared to previously used methods of credit scoring (e.g. cluster analysis, logistic regression) is that, when using survival analysis, one is able to predict *when* creditors will default and not only *whether* they will default. In the credit risk context, the survival function $S(t) = P(T > t)$ can be interpreted as the probability that some customer will still be repaying his loan at timepoint t.

Despite the fact that there are certain analogies between standard survival and survival in a credit loan context, there are also differences that might make the standard survival analysis approach inappropriate for the analysis of credit data. The main problem is that, typically, a very high proportion of credit data is right-censored, not only because the customer default is not observed during the observation period, but simply because default does not take place in the entire loan lifetime. Because of this, Tong et al. (2012) use a mixture cure approach to analyze the credit risk of a specific customer. The idea behind such models is that the population of loan applicants comprises two subpopulations, one that contains applicants that are susceptible to default (hence will default eventually), and another one that contains applicants that are not susceptible, or immune, and will never default. The susceptibility of a certain loan applicant is modeled by the so-called *incidence* model part, using logistic regression. Survival times of the susceptible subpopulation part are consequently modeled by the *latency* model part using proportional hazards regression. Recently, an R-package was introduced by Cai et al. (2012) to estimate such semi-parametric mixture models.

Using a mixture cure model involves working with a separate parameter vector for each of the two model parts. These two covariate vectors may or may not contain the same elements, which suggests that performing model selection on these kind of models could be very useful. However, hardly any attempts to perform model selection have been made in previous research. The use of widespread model selection criteria such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) for those mixture cure models is not straightforward, because the exact likelihood can not be computed because there is a problem of missing information. We develop and present an AIC for mixture cure models, using the idea of Cavanaugh and Shumway (1998) of calculating complete data log likelihoods performing the (supplemented) EM-algorithm.

References

[Tong et al. (2012)] Tong, E. N. C., Mues, C., and Thomas, L. C., 2012: Mixture cure models in credit scoring: if and when borrowers default. *Eur. J. Oper. Res.*, **218**, 132–139.

[Cai et al. (2012)] Cai, C., Zou, Y., Peng, Y. and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Comput. Meth. Prog. Bio.*, **108**, 1255–1260.

[Cavanaugh and Shumway (1998)] Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Stat. Plan. Infer.*, **67**, 45–65.

OCS30 Stoch. Neurosci.

Estimation in the Partially Observed Stochastic Morris-Lecar Neuronal model with Particle Filter and Stochastic Approximation Methods

SUSANNE DITLEVSEN[*,‡], ADELINE SAMSON[†]

[*]Department of Mathematical Sciences, University of Copenhagen, Denmark,
[†]Université Paris Descartes, France
[‡]email: susanne@math.ku.dk 109:SusanneDitlevsen.tex,session:OCS30

Parameter estimation in multi-dimensional diffusion models with only one coordinate observed is highly relevant in many biological applications, but a statistically difficult problem. In neuroscience, the membrane potential evolution in single neurons can be measured at high frequency, but biophysical realistic models have to include the unobserved dynamics of ion channels. One such model is the stochastic Morris-Lecar model, defined by a non-linear two-dimensional stochastic differential equation. The coordinates are coupled, i.e. the unobserved coordinate is non-autonomous, the model exhibits oscillations to mimick the spiking behavior, which means it is not of gradient-type, and the measurement noise from intra-cellular recordings is typically negligible. Therefore the hidden Markov model framework is degenerate, and available methods break down. The main contributions of this paper are an approach to estimate in this ill-posed situation, and non-asymptotic convergence results for the method. Specifically, we propose a sequential Monte Carlo particle filter algorithm to impute the unobserved coordinate, and then estimate parameters maximizing a pseudo-likelihood through a stochastic version of the Expectation-Maximization algorithm. It turns out that even the rate scaling parameter governing the opening and closing of ion channels of the unobserved coordinate can be reasonably estimated. Performance on simulated data and on intracellular recordings of the membrane potential of a spinal motoneuron are very encouraging.

CS16A Empirical processes

Change Point Estimation in Regression Models with Random Design

MAIK DÖRING[*,†]

[*]Universität Hohenheim, Stuttgart, Germany
[†]email: maik.doering@uni-hohenheim.de 110:MaikDoering.tex,session:CS16A

In this talk we consider a simple regression model with a change point in the regression function. Let for $n \in \mathbb{N}$ the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. $\mathbb{R}^2$-valued random variables with the same distribution as (X, Y). We assume that the distribution of X is absolutely continuous with a density function d_X, which is uniformly bounded on the unit interval $[0, 1]$. Further we assume that the response variables Y_i are given by the following regression model with an unknown change point $\theta_0 \in [0, 1)$ and unknown exponent $q_0 \in (0, \infty)$

$$Y_i = f_{\theta_0, q_0}(X_i) + \epsilon_i, \quad 1 \leq i \leq n,\ n \in \mathbb{N}.$$

For $(\theta, q) \in [0, 1] \times [0, \infty)$ the regression function $f_{\theta,q} : \mathbb{R} \to \mathbb{R}$ is given by

$$f_{\theta,q}(x) := (x - \theta)^q 1_{(\theta,1]}(x),$$

where 1_A is the indicator function of a set A. Let $\epsilon_1, \ldots, \epsilon_n$ for $n \in \mathbb{N}$ be i.i.d. real valued integrable random variables with the same distribution as ϵ. We assume that $E(\epsilon|X) = 0$ a.s. and that $P\big(X \in (\theta_0, 1)\big) > 0$.

We investigate the consistency with increasing sample size n of the least square estimates $(\hat{\theta}_n, \hat{q}_n)$ of the change point θ_0 and the exponent q_0. It turns out that $\sqrt{n}(\hat{q}_n - q_0) = O_P(1)$ and that we have the asymptotic normality property of the estimator for the exponent.

Further, it turns out that the rates of convergence of the change point estimator $\hat{\theta}_n$ depend on the exponent q_0. We show that $r_n\big(\hat{\theta}_n - \theta_0\big) = O_P(1)$ as $n \to \infty$, where the sequence $(r_n)_{n\in\mathbb{N}}$ is defined by

$$r_n := \begin{cases} n^{1/(2q_0+1)} & 0 \le q_0 < 1/2 \\ \big(n\ln(n)\big)^{1/2} & q_0 = 1/2 \\ n^{1/2} & 1/2 < q_0. \end{cases}$$

For $0 < q_0 < \frac{1}{2}$ the change point estimator converges to a maximizer of a Gaussian process. Some simulations suggest that the limit distribution is not normal. At $q_0 = \frac{1}{2}$ itself the rate of convergence of the change point estimator is $\sqrt{n \cdot ln(n)}$. Interestingly, the limit distribution is also normal. But for $q_0 > \frac{1}{2}$ we have a constant rate of $\sqrt{n}$ and the asymptotic normality property of the change point estimator.

Further we have that the change point estimator and the estimator for the exponent are asymptotically independent for $0 < q_0 \le \frac{1}{2}$.

Ergodicity of Observation-Driven Time Series Models and Consistency of the Maximum Likelihood Estimator

OCS8 Time Series

RANDAL DOUC[*,§], PAUL DOUKHAN[†], ERIC MOULINES[‡]

[*]Department CITI, CNRS UMR 5157, Telecom Sudparis, Evry. France.,
[†]Department of mathematics, University of Cergy-Pontoise, France.,
[‡]Department LTCI, CNRS UMR 5141, Telecom Paristech, Paris. France.
[§]email: randal.douc@telecom-sudparis.eu 111:RandalDouc.tex,session:OCS8

This paper deals with a general class of observation-driven time series models with a special focus on time series of counts. We provide conditions under which there exist strict-sense stationary and ergodic versions of such processes. The consistency of the maximum likelihood estimators is then derived for well-specified and misspecified models.

Inequalities for f-Divergence Measure and Applications

CS9C Model Selection

SILVESTRU SEVER DRAGOMIR[*,†,‡]

[*]Victoria University, Melbourne, Australia,
[†]University of the Witwatersrand, Johannesburg, South Africa
[‡]email: sever.dragomir@vu.edu.au 112:sever.tex,session:CS9C

The concept of *f-divergence* was introduced in the literature by I. Csiszár in his seminal paper [Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 1967 299–318] as a comprehensive and natural generalization of many divergence measures in Probability Theory & Statistics including the Kullback-Leibler distance, Hellinger discrimination, Jeffrey's distance and other divergence measures in Information Theory. The above paper is highly cited in the literature as one can learn from the Google Scholar database where more than 1000 citations are listed.

The purpose of our presentation is to survey some recent results obtained by the author in providing various inequalities for f-divergences by employing techniques and results from Analytic Inequalities Theory. Amongst these, we mentions reverses of Jensen inequality, Grüss type and Ostrowski type inequalities, to mention only a few.

Applications for some divergence measures of interest are also provided.

The interested reader may find details in the "*Research Group in Mathematical Inequalities & Applications*" website located at: http://rgmia.org/monographs/csiszar.htm, where preprint version of numerous related papers are provided and are available freely online.

IS15 Percol., R. Graphs

Viral Processes by Random Walks on Random Regular Graphs

MOEZ DRAIEF

Imperial College London
email: m.draief@imperial.ac.uk 113:Draief.tex,session:IS15

We study the SIR epidemic model with infections carried by k particles making independent random walks on a random regular graph. We give a edge-weighted graph reduction of the dynamics of the process that allows us to apply standard results of Erdös-Rényi random graphs on the particle set. In particular, we show how the parameters of the model produce the following phase transitions: In the subcritical regime, $O(\ln k)$ particles are infected. In the supercritical regime, for a constant $C \in (0,1)$ determined by the parameters of the model, Ck get infected with probability C, and $O(\ln k)$ get infected with probability $(1-C)$. Finally, there is a regime in which all k particles are infected. Furthermore, the edge weights give information about when a particle becomes infected. We demonstrate how this can be exploited to determine the completion time of the process by applying a results on first-passage percolation.

OCS4 3D Images

3D Shape Analysis in Ambient or Quotient Spaces

IAN L. DRYDEN[*,‡], ALFRED KUME[†], HUILING LE[*], ANDREW T.A. WOOD[*]

[*]University of Nottingham, Nottingham, United Kingdom
[†]University of Kent, Canterbury, United Kingdom
[‡]email: ian.dryden@nottingham.ac.uk 114:IanLDryden.tex,session:OCS4

One of the most basic summaries of a population or sample of shapes is a mean shape. However, there are several different notions of mean shape, and it is important to distinguish between them. Appropriate methodology and properties of the methods can differ substantially depending on the particular type of mean shape that is of interest.

A key issue in dealing with shapes of objects is whether to choose a model in the ambient space of the objects, or in the quotient space of objects modulo registration transformations. In the former case the distributions can become complicated after integrating out the registration information, whereas in the latter case the geometry of the space can be complicated after optimizing over registrations. Although in general these approaches are different, in many applications there are often practical similarities in the resulting inference due to a Laplace approximation.

We will focus on landmark shape analysis, for example points obtained from different 2D images of a face and then reconstructed as 3D co-ordinates. In this case the ambient space mean could be the shape of the mean from a multivariate normal model, and the quotient space mean could be a Fréchet mean, using either intrinsic or extrinsic distances in an embedding space, for example using multidimensional scaling (MDS) on the landmark co-ordinates.

We explore many of these issues through a set of examples, including some 3D face data and 3D molecule data. We demonstrate that there can be substantial differences in artificial examples depending on the resistance properties of the estimators. However, in practical real datasets the means are usually extremely close.

Block Bootstrap in the Second Order Analysis for Signals

OCS26 Resampling Nonstat T.S.

ANNA DUDEK*,†

*AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland.
†email: aedudek@agh.edu.pl 115:AnnaDudek.tex,session:OCS26

The purpose of this talk is to present the new methods of detection the significant frequencies corresponding to the autocovariance function of PC and APC processes. The presented techniques are based on Moving Block Bootstrap and Generalized Seasonal Block Bootstrap. The comparison with existing methods will be provided. The simulation and real data examples will also be presented.

Estimating Biophysical Parameters of Computational Neural Models from Spike Trains Using a Point Process Particle Filter Algorithm

IS22 Stat. Neuronal Data

URI EDEN*,†

*Department of Mathematics and Statistics, Boston University, Boston, MA, USA
†email: tzvi@bu.edu 116:Eden.tex,session:IS22

Since the seminal work by Hodgkin and Huxley, conductance based dynamical systems models have been used to simulate physiological ionic currents and characterize their relation to neural spiking activity. An important class of neural modeling problems focuses on identifying currents and estimating conductance levels in dynamical models to achieve experimentally observed patterns of spiking activity. However, parameter estimation for these models is often performed in an ad-hoc manner by manually adjusting parameters to achieve spiking patterns that are qualitatively similar to observed spike trains. Here, we develop a framework to combine conductance based neural modeling with point process statistical theory to estimate model components directly from a set of observed spike times. We construct estimation algorithms using sequential Monte Carlo (particle filter) methods that combine future and past spiking information to update a collection of model realizations that are consistent with the observed spiking data.

We apply these methods to both simulated data and to in vitro recordings from 4 layer 5 intrinsically bursting (IB) cells. For the simulation analyses, the goal is to verify that we are able to recover the true parameter values generating the observed spiking activity. We examined multiple classes of models, including Fitzhugh-Nagumo, standard Hodgkin-Huxley, and a Hodgkin-Huxley type model with an additional slow current, and find that the estimation method returns a collection of models that are consistent with the spiking data, including the one used to generate the data. These methods suggest specific experimental interventions that can be used to identify the generating model.

For the in vitro data analyses, we applied the methods to estimate an unknown current driving the slow, bursting activity characteristic of layer 5 IB cells. Multiple features of this ionic current, such as its conductance, reversal potential, and temporal properties, are unspecified and must be estimated from the spike time data. The procedure converges on a slow, depolarization activated, hyperpolarizing current for all the cells even though some fire in bursting patterns while others fire tonically.

These results indicate that spike time data carries information about cellular biophysical processes. The estimation methods provide an important new link between dynamical systems modeling and statistical neural data analysis.

Acknowledgment. This research was partially supported by the National Science Foundation, grant No.: IIS-0643995.

IS7 Forensic Stat.

Forensic Fingerprints Unicity and Interpretational Models

NICOLE M. EGLI ANTHONIOZ*,†

*Institut de police scientifique, Université de Lausanne, Switzerland

†email: Nicole.EgliAnthonioz@unil.ch 117:Nicole_EgliAnthonioz.tex,session:IS7

Fingerprint comparison has a long history; for over 100 years, fingerprints have been used to uniquely identify individuals, as well as allowing to find the donor of a mark left at a crime scene. The evolution of the use of this type of impression has heavily relied on the premise of individuality. The fingerprint has become the ideal of the identifying feature, to a point where when DNA analysis was introduced, it was referred to as "DNA fingerprinting", in order to highlight the interest of the method. More recently however, the demonstration of fingerprint unicity has been questioned, and even more so the conclusion of identity of source between a mark and a print. A few models computing the rarity of a set of fingerprint features have been created (allowing to compute small probabilities of duplication). The unicity of a complete fingerprint, even if it could be proven, would not be sufficient to demonstrate that a partial, degraded mark left on a crime scene can be uniquely attributed to a source. The focus will therefore be on two models allowing the computation of a likelihood ratio related to the comparison of a mark to a print. Up to now, these models are based on one-dimensional scores or distance measures that describe the comparison between the mark and the print, for characteristics that are called minutiae. Two such models will be discussed, and a third one, built on a different type of characteristics (the pores) shown.

OCS3 Spectral Analysis

Trek Separation and Latent Variable Models for Multivariate Time Series

MICHAEL EICHLER*,†

*Department of Quantitative Economics, Maastricht University, The Netherlands

†email: m.eichler@maastrichtuniversity.nl 118:MichaelEichler.tex,session:OCS3

In systems that are affected by latent variables conditional independences are often insufficient for inference about the structure of the underlying system. One common example is a system in which four observed variables X_1, X_2, X_3, and X_4 are conditionally independent given a fifth unobserved variable Y. While there are no conditional independences among the observed variables, they must satisfy the so-called tetrad constraints (e.g. Spirtes *et al.*, 2001)

$$\begin{aligned}\rho_{X_1X_2}\,\rho_{X_3X_4} - \rho_{X_1X_4}\,\rho_{X_2X_3} &= 0,\\ \rho_{X_1X_3}\,\rho_{X_2X_4} - \rho_{X_1X_4}\,\rho_{X_2X_3} &= 0,\\ \rho_{X_1X_2}\,\rho_{X_3X_4} - \rho_{X_1X_3}\,\rho_{X_2X_4} &= 0.\end{aligned}$$

Recently, Sullivant *et al.* (2010) discussed such additional non-Markovian constraints and provided a characterisation in terms of low-rank conditions on submatrices of the covariance matrix. Graphically these general constraints can be identified by a new separation concept called trek separation.

In this talk, we discuss the extension of the results to the multivariate time series case. Because of the commonly present serial correlation, the results are not directly applicable. For instance, the above tetrad constraints do not hold if the variables $X_1, \ldots, X_4$ and Y (as time series) have non-zero auto-correlation. Graphically, this corresponds to that fact that any instances of the variables $X_1, \ldots, X_4$ cannot be separated by a single instance of Y. As an alternative, we consider mixed graphs in which each node corresponds to a complete time series. Such graphical descriptions for

time series have been considered for instance by Dahlhaus (2000) and Eichler (2007). We show that trek separation in such graphs corresponds to low-rank conditions on the spectral matrix of the process. In particular, we obtain a spectral version of the above tetrad constraints in terms of spectral coherences. We discuss tests for vanishing tetrad constraints in the frequency domain based on asymptotic results and on bootstrap techniques.

References

[1] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika* **51**, 157–172.

[2] Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics* **137**, 334–353.

[3] Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. 2nd edn, MIT Press, Cambridge, MA.

[4] Sullivant, S., Talaska, K., and Draism, J. (2010). Trek separation for gaussian graphical models. *Annals of Statistics* **38**, 1665–1685.

Some Asymptotics for Localized Principal Components and Curves

CS38A Appl. Multivariate Tech.

JOCHEN EINBECK[*,‡], MOHAMMAD ZAYED[†]

[*]Durham University, Department of Mathematical Sciences, Durham City DH1 3LE, UK,
[†]Applied Statistics & Insurance Department, Mansoura University, Egypt
[‡]email: jochen.einbeck@durham.ac.uk 119:JochenEinbeck.tex,session:CS38A

We are given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sampled from a multivariate random vector $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$, variance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, and density $f(\cdot)$. We are interested in *localized PCA* in the sense of *locally weighted PCA*, where the weighting enters through multivariate kernel functions, $w^{\boldsymbol{x}}(\boldsymbol{x}_i) = K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x}) = |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{x})\right)$ centered at $\boldsymbol{x}$, with a positive definite bandwidth matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$. Let $\boldsymbol{\gamma}$ a vector with $||\boldsymbol{\gamma}|| = 1$ and $\boldsymbol{x}_i^{\boldsymbol{g}} = (\boldsymbol{I} - \boldsymbol{\gamma}\boldsymbol{\gamma}^T)\boldsymbol{m} + \boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{x}_i$ be the projection of $\boldsymbol{x}_i$ onto the line through $\boldsymbol{m}$ with direction $\boldsymbol{\gamma}$. The *localized first principal component* at $\boldsymbol{x}$ is the minimizer of

$$Q(\boldsymbol{m}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})||\boldsymbol{x}_i - \boldsymbol{x}_i^g||^2 - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1)$$

with solution $\boldsymbol{\Sigma}^{\boldsymbol{x}}\boldsymbol{\gamma} = -\lambda\boldsymbol{\gamma}$, where $\boldsymbol{\Sigma}^{\boldsymbol{x}} = \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}^{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}^{\boldsymbol{x}})^T / \sum_{i=1}^{n} w^{\boldsymbol{x}}(\boldsymbol{x}_i)$ is the localized covariance matrix at $\boldsymbol{x}$, and $\boldsymbol{\mu}^{\boldsymbol{x}}$ is the local mean at $\boldsymbol{x}$. It is easy to see that $\boldsymbol{\gamma}$ needs to be the first eigenvector of $\boldsymbol{\Sigma}^x$, which we denote by $\boldsymbol{\gamma}^{\boldsymbol{x}}$ henceforth. By exploiting asymptotic properties of the mean shift, $\boldsymbol{\mu}^{\boldsymbol{x}} - \boldsymbol{x}$, we show that, asymptotically [i.e., for $n^{-1}|\boldsymbol{H}|^{-1/2} \longrightarrow 0$ and $\boldsymbol{H} \longrightarrow 0$ as $n \longrightarrow \infty$] this direction is approximated by the density gradient rotated by the bandwidth matrix,

$$\tilde{\boldsymbol{\gamma}}^{\boldsymbol{x}} = -\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}) / ||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x})||.$$

The nonparametric equivalent to a (globally fitted) principal component line is a principal curve, which can be descriptively defined as a 'smooth curve through the middle of the data cloud'. The "local principal curve" algorithm is explicitly based on iterating local PCA steps. At j-th iteration, this algorithm can be described by

$$\boldsymbol{x}_{(j+1)} = \boldsymbol{\mu}_{(j)} \pm t\boldsymbol{\gamma}_{(j)}$$

where t is a step size, the sign in '$\pm$' is given by $\text{sign}(\boldsymbol{\gamma}_{(j)} \circ \boldsymbol{\gamma}_{(j-1)})$, and the sequence of $\boldsymbol{\mu}_{(j)}$'s (that is, local means at $\boldsymbol{x}_{(j)}$) constitutes the actual principal curve. We use the previously obtained result to show that, asymptotically,

$$\boldsymbol{\mu}_{(j+1)} - \boldsymbol{\mu}_{(j)} = \left[\frac{\mu_2(K)}{f(\boldsymbol{x}_{(j)})} \pm \frac{t}{||\boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})||}\right] \boldsymbol{H}\boldsymbol{\nabla} f(\boldsymbol{x}_{(j)})$$

with a kernel moment $\mu_2(K)$. If the curve is moving downhill, then the mean shift step will pull the curve backwards – towards higher densities –, so this distance will be rather small. The curve will stop at some point close to the boundary of the support of f if the two contributions are exactly the same. This theoretical result is confirmed through simulation, and it is used to provide a boundary extension which allows the principal curve to proceed beyond its natural endpoint. This is of particular importance for data with time series character, such as, for instance, multivariate consumer price index data. For data of this type, the current time point, which is likely to be the point of interest, will in general be a boundary point.

OCS13 H-D Stat, R. Matrices

Random Matrices and High-Dimensional M-Estimation: Applications to Robust Regression, Penalized Robust Regression and GLMs

NOUREDDINE EL KAROUI*

*Department of Statistics, UC Berkeley
†email: nkaroui@stat.berkeley.edu 120:NoureddineElKaroui.tex,session:OCS13

I will discuss the behavior of widely used statistical methods in the high-dimensional setting where the number of observations, n, and the number of predictors, p, are both large. I will present limit theorems about the behavior of the corresponding estimators, their asymptotic risks etc... The results apply not only to robust regression estimators, but also Lasso-type estimators and many much more complicated problems.

Many surprising statistical phenomena occur: for instance, maximum likelihood methods are shown to be inefficient, and loss functions that should be used in regression are shown to depend on the ratio p/n. This means that dimensionality should be explicitly taken into account when performing simple tasks such as regression.

It also turns out that inference is possible in the setting I consider. We will also see that the geometry of the design matrix plays a key role in these problems.

Mathematically, the tools needed mainly come from random matrix theory, measure concentration and convex analysis.

The talk is based on several papers, including joint works with Derek Bean, Peter Bickel, Ching-whay Lim and Bin Yu.

OCS24 Random Graphs

Spectra and Multiple Strategic Interaction in Networks

MARIANNA BOLLA*, AHMED ELBANNA*,†, ILDIKÓ PRIKSZ*

*Budapest University of Technology and Economics
†email: ahmed@math.bme.hu 121:AhmedElbanna.tex,session:OCS24

Strategic interaction is a game in which groups of players are looking for a common goal such as profit, see [1]. The communication between the players is realized through a graph given by its adjacency matrix $\mathbf{G}$, and their strategies also depend on positive parameters δ and α, characterizing the payoffs with and without bilateral influences, respectively. In the case of strategic complements, and also of substitutes when δ is 'small', a unique inner equilibrium exists; it can be found by matrix inversion and related to the Katz–Bonacich centrality of the network. However, in the case of strategic substitutes, for larger δ's (exact limits can be given in terms of the eigenvalues of $\mathbf{G}$ and its complement), corner equilibria appear. To find them, quadratic optimization is used, and in [2] the authors define an algorithm which examines all subsets of vertices for possible corner solutions.

This is computationally not tractable if the number of vertices is 'large'. Instead, we may approximate corner equilibria by spectral clustering tools. More generally, we consider multiple strategies. The k-dimensional strategies $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^k$ of the agents are the row vectors of the $n \times k$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, and they can be thought of as intensities of agents buying k different stocks or farmers planting k different crops. The simultaneous maximization of the agents' payoff is equivalent to maximizing the potential function

$$P(\mathbf{X}) = \sum_{j=1}^{k} \mathbf{x}_j^T((\alpha-1)\mathbf{I} - \delta\mathbf{G})\mathbf{x}_j = \operatorname{Tr} \mathbf{X}^T((\alpha-1)\mathbf{I} - \delta\mathbf{G})\mathbf{X}$$
$$\text{subject to } \mathbf{X}^T\mathbf{X} = \mathbf{I}_k.$$

Its maximum is $\sum_{j=1}^{k}(\alpha - 1 - \delta\lambda_j)$, where $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\mathbf{G}$, and it is attained by the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ as columns of $\mathbf{X}$. Then this spectral relaxation is approximated by stepwise constant vectors of nonnegative coordinates. The subspace of these so-called partition-vectors is close to the subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_k$ provided there is a gap between λ_k and λ_{k+1}. In this case, the clusters of agents pursuing similar strategies are obtained by applying the k-means algorithm to the optimum strategy vectors, row vectors of the optimum $\mathbf{X}$.

Acknowledgment. This research was supported by the TÁMOP-4.2.2.B-10/1-2010-0009 project.

References

[1] Ballester, C., Calvo-Armengol A., Zenou, Y. (2006) Who's who in networks. Wanted: the key player. *Econometrica*, **74**, 1403-1417.

[2] Bramoullé, Y., Kranton, R. E. (2007) Public goods in networks. *Journal of Economic Theory*, **135**, 478-494.

Statistical Properties of the Site-Frequency Spectrum Associated with Lambda-Coalescents

IS28 Stoch. in Biol.

MATTHIAS BIRKNER*, JOCHEN BLATH†, BJARKI ELDON†

*Johannes-Gutenberg-Universität, Institute für Mathematik, 55099 Mainz, Germany
†TU Berlin, Institute für Mathematik, 10623 Berlin, Germany
‡email: eldon@math.tu-berlin.de 122:BjarkiEldon.tex,session:IS28

Lambda-coalescents arise naturally in population models with skewed offspring distributions, in which individuals can have very many offspring, or up to the order of the population size. Lambda-coalescents differ from the Kingman coalescent, the usual 'null' model in population genetics, by admitting multiple mergers of ancestral lineages. In multiple mergers, any number of active ancestral lineages can coalesce, as opposed to only two in the Kingman coalescent. The occurrence of multiple mergers changes the topology of the genealogy, and hence patterns of neutral genetic diversity, and therefore the site-frequency spectrum should be different from that predicted by the Kingman coalescent. One also expects different Lambda-coalescents to predict different site-frequencies (counts of mutations in different copy numbers).

Statistical properties of the site frequency spectrum obtained from Lambda-coalescents are our objects of study. In particular, we derive recursions for the expected value, variance, and covariance of the site frequency spectrum (SFS), extending earlier results obtained by Fu (1995) for the Kingman coalescent. The recursions can be extended in a straight-forward way to Xi-coalescents admitting *simultaneous* multiple mergers of ancestral lineages.

The recursions allow us to apply an approximate likelihood approach to distinguish between certain parametric subclasses of Lambda-coalescents using the full frequency spectrum, and to distinguish between the Kingman coalescent (the null) and Lambda coalescents.

Finally, we provide some empirical studies. In particular, we investigate by simulation how large sample size and mutation rates have to be so that asymptotic formulas derived by Berestycki, Berestycki and Limic (2012) begin to describe the observed spectra sufficiently well. In addition, we infer the coalescent model for Atlantic cod datasets, e.g. the full mitochondrial cytochrome b dataset of Árnason (2004).

OCS30 Stoch. Neurosci.

Neuropercolation and Related Models of Criticalities

PÉTER ÉRDI[*,†,‖], RÓBERT KOZMA[‡], MARKO PULJIC[§], JUDIT SZENTE[*,¶]

[*]Center for Complex Systems Studies Kalamazoo College, Kalamazoo, MI, USA,
[†]Institute for Particle and Nuclear Physics, Wigner Research Centre for Physics, Hungarian Academy of Sciences, Budapest, Hungary,
[‡]University of Memphis, Memphis, TN, USA,
[§]Tulane University, New Orleans, LA, USA,
[¶]Western Michigan University, Kalamazoo, MI, USA
[‖]email: perdi@kzoo.edu

123:Peter_Erdi.tex,session:OCS30

Power laws, black swans, dragon kings Power-law statistics became ubiquitous in analyzing extreme events in nature and society. The concepts of "Black swans" and "Dragon Kings" were suggested to characterize extreme and super-extreme events related to self-organized criticality and intermittent criticality, respectively ([Sornette D and Ouillon G, 2012]). Neuronal avalanches might belong to both categories. Here we contribute to the understanding of generating mechanisms of dragon kings in general and to large activation clusterings in neural systems, specifically, by using the conceptual and mathematical framework of neuropercolation ([Kozma 2007]).

Neuropercolation is a family of stochastic models based on the mathematical theory of probabilistic cellular automata on lattices and random graphs motivated by structural and dynamical properties of neural populations. The existence of phase transitions was demonstrated both in continuous and discrete state space models, e.g. in specific probabilistic cellular automata and percolation models. Neuropercolation extends the concept of phase transitions to large interactive populations of nerve cells.

Models of the archetypal percolation problem, the vertices (sites) are the points of a lattice with edges (bonds) joining neighboring sites. In site percolation, sites are open independently with probability p and one wishes to answer questions about the size of the connected components formed by these open sites. The evolution rules adopted here are:
At time $t-1$, v_i's state $s_{i,t-1} = 0$ or 1. v_i senses a state of its neighbor $v_{i_j} \in I_{i,t-1}$ with chance $0 \leq \omega_{i_j,i,t-1} \leq 1$ and $\omega_{i_j,i} \in \Omega_{i,t-1}$. $s_{i,t}$ is the most common $s_{i_j,t-1}$ in $I_{i,t-1}$. More precisely, evolution starts at time $t = 0$, when $s_{i,t} = 0$. Then, at $t = 1, 2, .., T-1$, v_i is influenced by each $s_{i_j,t-1}$ when a random variable $R_{i,t} < \omega_{i_j,i,t-1}$, else v_i is influenced by $1 - s_{i_j,t-1}$. $s_{i,t}$ is set to majority influences, if there is such, otherwise $s_{i,t}$ is randomly set to 0 or 1.

Simulation results showed there is transition from self-organized critical regime to intermittent criticality. Around the critical point reduction in excitation implies super-exponential increase in the size of the neural activation avalanches, so it leads to the appearance of dragon kings.

References

[Sornette D and Ouillon G, 2012] Sornette D and Ouillon G, 2012: Dragon-kings: Mechanisms, statistical methods and empirical evidence, *European Physical Journal Special Topics* **205**, 1-26.

[Kozma 2007] Kozma R: Neuropercolatio, *Scholarpedia*, 2(8):1360.

Selecting The Model Using Penalized Spline Regression with Bayesian Perspective by Real Data

CS8A Bayesian Semipar.

MAHMUT SAMI ERDOGAN*, OZLEM EGE ORUC*

*Dokuz Eylul University, Izmir, Turkey

124:MahmutSamiErdogan.tex,session:CS8A

The flexibility of semiparametric regression method has a great advantage in modeling. Penalized spline regression with Bayesian perspective for the problem of selecting the model and coefficients are considered. For the method, the knot sequence of spline functions coincides with the end points of the interval. A real data example of ratios of export to import in Turkey is modeled and discussed.

References

[1] Ciprian M. C., Ruppert, D., Wand M. P., 2005: Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software* 14,

Score Function of Distribution: A New Inference Function

NYA Not Yet Arranged

FABIÁN ZDENĚK*,†

*Institute of Computer Sciences, Czech Academy of Sciences, Prague
†email: zdenek@cs.cas.cz

125:FabianZdenek.tex,session:NYA

Denoting by $f(x;\theta)$ the density of a regular continuous distribution F_θ, $\theta \in \Theta, \Theta \subseteq \mathbb{R}^m, m \geq 1$, the score function $S_F(x;\theta)$ of F_θ is a support-dependent function of the ratio $f'(x;\theta)/f(x;\theta)$. The function reflects main features of F_θ (namely properties of its tails). In certain cases it equals to the Fisher score function for the most important component of the vector of parameters, in other cases it is a new function. In contrast to the vector Fisher score function, S_F is a simple scalar function. In this talk we define it, discuss its properties for different distributions and show examples of its use:

1. Instead of the mean and variance, a typical value of a continuous distribution can be characterized by the solution $x^* = x^*(\theta)$ of equation $S_F(x;\theta) = 0$, and variability by the score variance $\omega^2 = 1/ES_F^2(\theta)$, where $ES_F^2(\theta)$ is the Fisher information for x^*. These new characteristics exist even in cases of heavy-tailed distributions.

2. Let $\hat{\theta}$ be an estimate of θ based on sample $[X_1, ..., X_n]$ from $F \in \{F_\theta, \theta \in \Theta\}$. Sample versions $x^*(\hat{\theta})$ and $\omega^2(\hat{\theta})$ of x^* and ω^2 and their standard deviations represent new descriptions of data samples, making possible to compare results of estimation under assumption of various models with different parametrization by a simple way.

3. Score function of distribution of a heavy-tailed distribution is bounded, score moments $ES_F^k(\theta)$ finite and estimate of θ based on empirical moments of $S_F^k(x;\theta)$ robust. A use of transformed samples $[S_F(X_1;\hat{\theta}), ..., S_F(X_1;\hat{\theta})]$ yields robust estimates of correlation and linear regression coefficients.

Acknowledgment. This work is done with institutional support RVO:67985807

POSTER Poster

Assessing the Risk of Implementing Some Convolution Models for Background Correction of BeadArrays

ROHMATUL FAJRIYAH*,†

*Institute of Staistics, TU Graz, Austria
†email: fajriyah@student.tugraz.at 126:Fajriyah.tex,session:POSTER

The robust multi-array average (RMA), since its introduction by Irizarry et al. (2003, 2006), has gained popularity among bioinformaticians.The RMA has evolved from the exponential-normal convolution to gamma-normal convolution, from single to two channels and from Affymetrix to Illumina platform.

The Illumina design has provided two probe types: the regular and the control probes. This design enables one to study the distribution of both and to apply the convolution model to compute the true intensity estimator. The availability of a benchmarking data set from Illumina platform, the so called Illumina spike-in, helps researchers to evaluate their proposed method for Illumina BeadArrays.

In an earlier paper we studied the existing convolution models for background correction of Illumina BeadArray and proposed a new model. By using the benchmarking and public data sets, we did a comparative study based on the benchmarking criteria adapted from Affycomp (2004), Shamilov et al. (2006), Xie et al. (2009), Plancade et al. (2011, 2012) and Chen et al. (2011). In both data sets, the proposed model performed considerably better than the existing models.

In this paper, we discuss the risk of using a particular convolution model for the data at hand, measured by the ratio of mean absolute deviation of the model against the true model. This is then applied to the benchmarking and the public data set.

CS26B Life and Failure Time

Reliability Analysis of a Series System with Bivariate Weibull Components under Step Stress Accelerated Life Tests

TSAI-HUNG FAN*,†, SE-KAI JU*

*National Central University, Jhongli, Taiwan
†email: thfanncu@gmail.com 127:TsaiHungFAN.tex,session:CS26B

A series system fails if any of the components fails. These components are all from the same system, hence they may be correlated. In this paper, we consider the step stress accelerated life test of a two-component series system with a bivariate Marshall-Olkin Weibull lifetime distribution on the components. It is often to include masked data in which the component that causes failure of the system is not observed. We apply the maximum likelihood approach via EM algorithm and derive the observed Fisher information based on the missing information principle. Statistical inference on the model parameters and the mean lifetimes and the reliability functions of the system as well as components under usual operating conditions are derived. The proposed method is illustrated through a numerical example and a simulation study under various masking levels.

A New Measure of Uniformity — Mixture Discrepancy

CS32A Nonparame

KAI-TAI FANG[*,†,‡]

[*]BNU-HKBU United International College, Zhuhai, China,
[†]Chinese Academy, Beijing, China
[‡]email: ktfang@uic.edu.hk 128:Kai-Tai_Fang.tex,session:CS32A

There are various discrepancies that are measures of uniformity of a set of points on the unit hypercube. The discrepancies have played an important role in quasi-Monte Carlo Methods. Each discrepancy has its own characteristic and some weakness. In this talk we point out some unreasonable phenomena of the commonly used discrepancies in the literature such as the L_p-star discrepancy, the center L_2-discrepancy (CD) and the wrap-around L_2-discrepancy (WD). Then, a new discrepancy, called mixture discrepancy (MD), is proposed. The mixture discrepancy is more reasonable than CD and WD in a certain sense. Moreover, the relationships between MD and other design criteria such as balance pattern and generalized word-length pattern are also given.

Covariance Modelling in Longitudinal Data with Informative Observation

OCS1 Longitudin Models

DANIEL FAREWELL[*,†], CHAO HUANG[*]

[*]School of Medicine, Cardiff University, UK
[†]email: farewelld@cf.ac.uk 129:DanielFarewell.tex,session:OCS1

When using generalized estimating equations to model longitudinal data, both inconsistency (due to informative observation) and inefficiency (due to misspecified working covariances) are often of concern. We describe a class of generalized inverses of singular working correlation matrices that allows flexible modelling of covariance within a subject's responses while offering robustness to certain kinds of informative observation. We demonstrate how this class corresponds to dynamic models on the increments in the longitudinal responses, and illustrate its application to a randomized trial of quetiapine in the treatment of delirium.

Bayesian Nonparametric Estimation of Discovery Probabilities

IS2 Bayesian Nonpar.

STEFANO FAVARO[*,‡,§], ANTONIO LIJOI[†,‡], IGOR PRÜNSTER[*,‡]

[*]University of Torino, Italy,
[†]University of Pavia, Italy,
[‡]Collegio Carlo Alberto, Torino, Italy
[§]email: stefano.favaro@unito.it 130:StefanoFavaro.tex,session:IS2

Species sampling problems have a long history in ecological and biological studies and a number of statistical issues, including the evaluation of species richness, the design of sampling experiments and the estimation of rare species variety, are to be addressed. Such inferential problems have recently emerged also in genomic applications, however exhibiting some peculiar features that make them more challenging: specifically, one has to deal with very large genomic libraries containing a huge number of distinct genes and only a small portion of the library has been sequenced. These aspects motivate the Bayesian nonparametric approach we undertake, since it allows to achieve the degree of flexibility typically needed in this framework. Basing on an initial observed sample of size n, focus will be on prediction of a key aspect of the outcome from an additional sample of size m,

namely the so-called discovery probability. In particular, conditionally on the observed initial sample, we derive a novel estimator of the probability of detecting, at the $(n+m+1)$-th observation, species that have been observed with any given frequency in the enlarged sample of size $n+m$. The result we obtain allows us to quantify both the rate at which rare species are detected and the achieved sample coverage of abundant species, as m increases. Natural applications are represented by the estimation of the probability of discovering rare genes within genomic libraries and the results are illustrated by means of two Expressed Sequence Tags datasets.

Acknowledgment. This work was supported by the European Research Council (ERC) through StG ŇN-BNPÓ 306406.

OCS23 Strong Limit Thm.

Limit Theorems for the Generalized Allocation Scheme

ISTVÁN FAZEKAS*,‡, ALEXEY CHUPRUNOV†

*University of Debrecen, Hungary,
†Kazan State University, Russia
‡email: fazekas.istvan@inf.unideb.hu

131:IstvanFazekas.tex,session:OCS23

Let $\xi_1, \xi_2, \ldots, \xi_N$ be independent identically distributed non-negative integer valued non-degenerate random variables. In the generalized allocation scheme introduced by [Kolchin (1999)], random variables $\eta'_1, \ldots, \eta'_N$ are considered with joint distribution

$$P\{\eta'_1 = k_1, \ldots, \eta'_N = k_N\} = P\Big\{\xi_1 = k_1, \ldots, \xi_N = k_N \;\Big|\; \sum_{i=1}^N \xi_i = n\Big\}.$$

This scheme contains several interesting particular cases such as the usual allocation scheme and random forests. Inequalities and limit theorems are proved for Kolchin's generalized allocation scheme.

Moreover, random variables $\eta_1, \ldots, \eta_N$ with joint distribution

$$P\{\eta_1 = k_1, \ldots, \eta_N = k_N\} = P\Big\{\xi_1 = k_1, \ldots, \xi_N = k_N \;\Big|\; \sum_{i=1}^N \xi_i \le n\Big\}$$

are also studied. It can be considered as a general allocation scheme when we place at most n balls into N boxes. In this general allocation scheme the random variable $\mu_{nN} = \sum_{i=1}^N I_{\{\eta_i = r\}}$ is the number of boxes containing r balls.

We study laws of large numbers, i.e. the convergence of the average $\frac{1}{N}\mu_{nN}$, as $n, N \to \infty$. We prove local limit theorems, i.e. we study the asymptotic behaviour of $P\{\mu_{nN} = k\}$. We obtain weak limit theorems for the maximum, i.e. we shall consider the asymptotic behaviour of $P\{\max_{1\le i\le N} \eta_i \le r\}$.

Acknowledgment. Partially supported by the Hungarian Scientific Research Fund under Grant No. OTKA T079128/2009. Partially supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[Chuprunov and Fazekas (2010) a] Chuprunov, A., Fazekas, I., 2010: An exponential inequality and strong limit theorems for conditional expectations. *Period. Math. Hungar.* **61**, no. 1-2, 103 - 120.

[Chuprunov and Fazekas (2010) b] Chuprunov, A., Fazekas, I., 2010: An inequality for moments and its applications to the generalized allocation scheme. *Publ. Math. Debrecen* **76**, no. 3-4, 271 - 286.

[Kolchin (1999)] Kolchin, V. F., 1991: *Random Graphs*, Cambridge University Press, Cambridge.

Mutual Distance Correlation

CS6F Copulas

GÁBOR J. SZÉKELY[*,†], TAMÁS FEGYVERNEKI[‡,§]

[*]National Science Foundation, USA,
[†]Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences,
[‡]Eötvös Loránd University, Budapest, Hungary
[§]email: tfegyi@cs.elte.hu 132:TamasFegyverneki.tex,session:CS6F

Distance correlation is a multivariate dependence coefficient, analogous to the classical Pearson product-moment correlation. It was defined by Gábor Székely [Székely et al. (2007)]. Distance covariance and correlation are applicable to random vectors of arbitrary and not necessarily equal dimension, and characterize independence, so they equal zero if and only if the random variables are independent. Distance correlation can be generalized by using α-th powers of Euclidean distances, having the same benefits if $0 < \alpha < 2$. (Distance correlation with $\alpha = 2$ leads to the absolute value of the classical product-moment correlation and thus does not characterize independence.)

Brownian covariance is also a measure of dependence between random vectors. It is a special case of a covariance with respect to a stochastic process, using the Brownian motion. It is shown that distance covariance coincides with Brownian covariance, so they can be called Brownian distance covariance [Székely et al. (2009)]. Brownian covariance is also a natural extension for classical covariance, as we obtain Pearson's product-moment covariance by replacing Wiener process in the definition with identity. Replacing the Wiener process with Lévy fractional Wiener process with Hurst exponent $\frac{\alpha}{2}$ leads to α-distance covariance.

Unlike other dependence coefficients that characterize independence, distance covariance and distance correlation can easily be computed. The empirical computation formula applies to all sample sizes $n \geq 2$, is not constrained by the dimension and does not require parameter estimation or matrix inversion. The simplicity of the formula allows us to easily define the extension of empirical distance covariance for more than two random vectors. We extend the bivariate formula in [Székely et al. (2007)] for three random vectors and discuss the meaning of the resulting mutual distance correlation. A general formula is also claimed for arbitrary many random vectors.

References

[Székely et al. (2009)] Székely, G. J., Rizzo, M. L., 2009: Brownian distance covariance, *Annals of Applied Statistics*, **3**, 1236-1265

[Székely et al. (2007)] Székely, G. J., Rizzo, M. L., Bakirov, N. K., 2007: Measuring and testing dependence by correlation of distances, *Annals of Statistics*, **35**, 2769-2794

Estimation of a Density in a Simulation Model

CS6C Func. Est., Smoothing

TINA FELBER[*,†], MICHAEL KOHLER[*]

[*]Technical University of Darmstadt, Germany.
[†]email: tfelber@mathematik.tu-darmstadt.de 133:TinaFelber.tex,session:CS6C

The problem of estimation of a density in a simulation model is considered, where given a value of a random $\mathbb{R}^d$-valued input parameter X the value of a real-valued random variable $Y = m(X)$ is computed. Here $m : \mathbb{R}^d \to \mathbb{R}$ is a (measurable) function which computes the quality $m(X)$ of a technical system with input X. It is assumed that both X and Y have densities. Given a sample of (X, Y) the task is to estimate the density of Y. We compute estimates by estimating in a first step m and the density of X, and by using these estimates in a second step to compute an estimate of the

density of Y. Results concerning L_1-consistency and rate of convergence of the estimates are proven and the estimates are illustrated by applying them to simulated and real data.

Acknowledgment. The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Center 805.

CS6E Function Est.

Local Variance Estimation for Uncensored and Censored Observations

PAOLA GLORIA FERRARIO*,†

*Universität zu Lübeck, Germany

†email: paola.ferrario@imbs.uni-luebeck.de 134:PaolaFerrario.tex,session:CS6E

This talk is motivated by a problem frequently considered in medical science. Assume that a patient is suffering from a certain disease and the physician wants to predict how the disease proceeds. The aim is in particular to predict the survival time of the patient in dependence on different factors ("predictors").
Mathematically speaking, the task is to estimate the mean survival time Y given a realization x of the d-dimensional predictor vector X : $\mathbb{E}\{Y|X = x\} =: m(x)$. The quality of the prediction by m can be globally defined by the (minimum) mean squared error and locally by the local variance $\sigma^2(x) := E\{(Y - m(X))^2|X = x\}$. The aim of the talk is to give estimators of the local variance function and to show some asymptotic properties of them. In particular, we deal with local averaging (partitioning, nearest neighbor) and least squares estimation methods.
A feature that complicates the analysis is that the follow-up program of the patients may be incomplete. After a certain censoring time, there is no information any longer about the patient. We estimate the local variance also under censoring, using in addition the product-limit estimator.

References

[Ferrario (2013)] Ferrario, P.G., 2013: *Local variance estimation for uncensored and censored observations*, Springer, to appear.

[Ferrario and Walk (2012)] Ferrario, P.G., Walk, H., 2012: Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors, *J. Nonaparametr. Stat.*, **24**, 1019 - 1039.

[Györfi et al. (2002)] Györfi, L., Kohler, M., Krzyżak, A., Walk, H., 2002: *A distribution-free theory of nonparametric regression*, Springer.

[Kohler (2006)] Kohler, M., 2006: Nonparametric regression with additional measurement errors in the dependent variable, *J. Statist. Plann. Inference*, **136**, 3339 - 3361.

[Liitiäinen et al. (2008)] Liitiäinen, E., Corona, F., Lendasse, A., 2008: On nonparametric residual variance estimation, *Neural Process. Lett.*, **28**, 155 - 167.

CS40A Logistic & Multinom. Distr.

Estimating Overdispersion in Sparse Multinomial Data

DAVID FLETCHER*,†

*Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin, New Zealand

†email: dfletcher@maths.otago.ac.nz 135:Fletcher.tex,session:CS40A

Consider the situation where $(n_1, \ldots, n_k)$ is multinomial (N, π) and we have a model for π that involves p parameters, where $p \ll k$. When a substantial number of the elements of π are small, testing lack-of-fit of the model can be prone to problems. Suppose we wish to measure the amount of lack-of-fit, rather than test for it, and we assume that it can be modelled as overdispersion, i.e. $\text{var}(n_i) = \phi N_i \pi_i (1 - \pi_i)$ $(i = 1, \ldots, k)$, where ϕ is the overdispersion parameter. An estimate of ϕ

can be used in calculating standard errors, and in performing model selection and model averaging. We propose a new estimator of ϕ based on a modified version of Pearson's lack-of-fit statistic and use simulation to assess its performance in the context of fitting a mark-recapture model. Our results suggest that the new estimator is preferable to the current alternatives, in terms of both bias and variance.

Testing Linearity for Nonlinear Count Time Series Models

CS4B Time Series I.

KONSTANTINOS FOKIANOS*,†, VASILIKI CHRISTOU*

*University of Cyprus, Nicosia, Cyprus
†email: fokianos@ucy.ac.cy 136:Konstantinos_Fokianos.tex,session:CS4B

We consider testing linearity against two special classes of nonlinear alternatives for count time series data. The first class contains of models which do not face the problem of nonidentifiability, that is all the parameters of the model are identified under the null hypothesis. For this class of models and under the null hypothesis of linearity, the score test statistic possesses an asymptotic $\mathcal{X}^2$ distribution. The second class of nonlinear models consists of models in which a nonnegative nuisance parameter exists under the alternative hypothesis but not when linearity holds. In this particular case the testing problem is nonstandard and the classical asymptotic theory for the score test does not apply.
We focus on count time series autoregressive models based on either the Poisson or the negative binomial distribution. After parameterizing suitably the negative binomial distribution so that it has the same mean as the Poisson, we employ quasi likelihood inference to get the consistent estimators. Once the estimators are obtained, we calculate the score test statistic and based on parametric bootstrap procedure, we investigate the size and the power of the test by a simulation study.

A Bayesian Vector Multidimensional Scaling Procedure Incorporating Dimension Reparameterization with Variable Selection

CS22A R. Matrices

DUNCAN K. H. FONG*,†

*Penn State University, University Park, PA, USA
†email: i2v@psu.edu 137:DuncanFong.tex,session:CS22A

To analyze data collected from surveys where subjects are asked to rate multiple stimuli, multidimensional scaling (MDS) is commonly used to produce a joint space map of subjects and stimuli in a reduced dimensionality in order to gain insights into the inter-relationships between these row and column entities. In this paper, we propose a new Bayesian vector MDS procedure incorporating dimension reparameterization with variable selection options to determine the dimensionality and simultaneously identify the significant covariates that help interpret the derived dimensions in the joint space map. We discuss how we solve identifiability problems that are typically associated with the vector MDS model, and show that the variable selection results are not affected by the proposed identification procedure. We demonstrate through simulation as well as a real data example how our proposed model outperforms a benchmark model.

CS33A Longitudin Data

Joint Modelling of Longitudinal and Event History Data

ALI REZA FOTOUHI[*,†]

[*]University of the Fraser Valley, Abbotsford, Canada
[†]email: ali.fotouhi@ufv.ca 138:AliRezaFotouhi.tex,session:CS33A

Longitudinal count data commonly arise in clinical trials studies where the response variable is the number of multiple recurrences of the event of interest and observation times are variable among cases. Two processes therefore exist. The first process is for recurrent event, for example bladder tumour count, and the second process is for duration between recurrences, the duration between two tumour removals. Many studies have been done to analyze the recurrent event and the duration between recurrences separately. These studies have assumed that the recurrent event and the duration between recurrences are independent. A few studies have been referred to the case that these two processes may be dependent. The same problem arises when longitudinal binary data and the repeated durations are observed simultaneously. For example, the status of a resident, mover or stayer, and the duration that the resident stays in the current status may be correlated. Separate modelling of longitudinal and event history data may result biased estimations and therefore misleading interpretation. In this research, joint and separate modelling of longitudinal data and event history data are studied and compared. In joint modelling approach we model the possible correlation between the longitudinal data and the repeated durations through random effects. The proposed models are evaluated in a simulation study and are applied to bladder cancer and residential mobility data.

IS15 Percol., R. Graphs

Ultra-fast Rumor Spreading in Social Networks

NIKOLAOS FOUNTOULAKIS[*,§], KONSTANTINOS PANAGIOTOU[†,¶], THOMAS SAUERWALD[‡,‖]

[*]School of Mathematics, University of Birmingham,
[†]Mathematics Institute, LMU, Munich,Germany,
[‡]Max Planck Institute for Informatics, Saarbrücken, Germany
email: [§]n.fountoulakis@bham.ac.uk, [¶]kpanagio@math.lmu.de,
[‖]sauerwal@mpi-inf.mpg.de 139:NikolaosFountoulakis.tex,session:IS15

We analyze the popular push-pull protocol for spreading a rumor in networks. Initially, a single node knows a rumor. In each succeeding round, every node chooses a random neighbor, and the two nodes share the rumor if one of them is already aware of it. We present the first theoretical analysis of this protocol on random graphs that have a power law degree distribution with an arbitrary exponent $\beta > 2$.

Our main findings reveal a striking dichotomy in the performance of the protocol that depends on the exponent of the power law. More specifically, we show that if $2 < \beta < 3$, then the rumor spreads to almost all nodes in $O(\log \log n)$ rounds with high probability. On the other hand, if $\beta > 3$, then $\Omega(\log n)$ rounds are necessary.

We also investigate the asynchronous version of the push-pull protocol, where the nodes do not operate in rounds, but exchange information according to a Poisson process with rate 1. Surprisingly, we are able to show that, if $2 < \beta < 3$, the rumor spreads even in constant time, which is much smaller than the typical distance of two nodes.

Correlated-Errors-in-Variables Regressions in Method Comparison Studies

CS9D Testing Mod. Structure

BERNARD G. FRANCQ*,†, BERNADETTE GOVAERTS*

*Institute of Statistics, Biostatistics and Actuarial sciences, Louvain-la-Neuve, Belgium
†email: bernard.g.francq@uclouvain.be 140:BernardG.Francq.tex,session:CS9D

The correlated-errors-in-variables regressions are widely applied for instance in engineering or chemometrics. Indeed, the needs of the industries to quickly assess the quality of products leads to the development and improvement of new measurement methods sometimes faster, easier to handle, less expensive or more accurate than the reference method. These alternative methods should ideally lead to results comparable to those obtained by a standard method. Ideally, there is no bias between these two methods or they should be interchangeable.

An approach based on a regression analysis (a linear functional relationship) is widely applied to assess the equivalence between measurement methods. On the other hand, the well-known and widely used Bland and Altman plot focuses directly to the observed differences between two measurement methods and the equivalence is assessed by predictive or agreement intervals. The regression approach focuses on the parameter estimates and their confidence interval (CI). To test statistically the equivalence between two measurement methods, a given characteristic of a sample (or a patient) can be measured by the methods in the experimental domain of interest. The pairs of measures taken by the reference method and the alternative one can be modeled by a linear regression (a straight line). Then, the parameter estimates are very useful to test the equivalence. Indeed, an intercept significantly different from zero indicates a systematic analytical bias between the methods and a slope significantly different from one indicates a proportional bias. To achieve this correctly, it is essential to take into account the errors in both axes and the heteroscedasticity if necessary. Some regressions are proposed to tackle the heteroscedasticity and when the error variances are known, the obtained coverage probabilities are very close to each other. Unfortunately under unknown and heteroscedastic error variances, the coverage probabilities drop drastically when the variances are locally estimated. Actually, the uncertainties on the estimated variances are not taken into account. Then, predicted variances (computed with a regression analysis) can be plugged into the correlated-errors-in-variables regression instead of locally estimated variances. The obtained coverage probabilities are then close to the nominal level.

Ideally, the statistical hypotheses on the parameters of the regression have to be tested jointly. This joint-hypothesis can be assessed by an ellipse or equivalently by a confidence band for the regression. These confidence bands are given from $x = -\infty$ to $x = \infty$ but ideally, they should be restricted to a range covered by the measurement method (for instance, from 80 to 220 mmHg with systolic blood pressure data). These restricted confidence bands can be computed with the OLS and we will explain how to modify them in the context of correlated-errors-in-variables regressions to assess the equivalence of two measurement methods in a given range of measurements.

Maxiset Performance of Hyperbolic Wavelet Thresholding Estimators

CS6E Function Est.

FLORENT AUTIN*, GERDA CLAESKENS†, JEAN-MARC FREYERMUTH†,‡

*Université d'Aix-Marseille 1, L.A.T.P., Marseille, France
†KU Leuven, ORSTAT, Leuven, Belgium
‡email: Jean-Marc.Freyermuth@kuleuven.be 141:Jean-MarcFreyermuth.tex,session:CS6E

In this talk we are interested in nonparametric multivariate function estimation. In [1], we deter-

mine the maxisets of several estimators based on thresholding of the empirical hyperbolic wavelet coefficients. That is we determine the largest functional space over which the risk of these estimators converges at a chosen rate. It is known from the univariate setting that pooling information from geometric structures (horizontal/vertical blocks) in the coefficient domain allows to get 'large' maxisets (see e.g [2, 4, 3]). In the multidimensional setting, the situation is less straightforward. In a sense these estimators are much more exposed to the curse of dimensionality. However we identify cases where information pooling has a clear benefit. In particular, we identify some general structural constraints that can be related to compound models and to a 'minimal' level of anisotropy.

References

[1] Autin, F., Claeskens, G., Freyermuth, J-M. (2013). Hyperbolic wavelet thresholding rules: the curse of dimensionality through the maxiset approach. *to appear in Applied and Computational Harmonic Analysis*.

[2] Autin, F., Freyermuth, J-M., von Sachs, R. (2011). Ideal denoising within a family of Tree Structured Wavelets. *Electronic Journal of Statistics*, **5**, 829-855.

[3] Autin, F., Freyermuth, J-M., von Sachs, R. (2013). Block-threshold-adapted estimators via a maxiset approach. *to appear in the Scandinavian Journal of Statistics.*

[4] Autin, F., Freyermuth, J-M., von Sachs, R. (2012) Combining thresholding rules: a new way to improve the performance of wavelet estimators. *Journal of Nonparametric Statistics*, **24**, no. 4, 905 - 922.

IS24 Single Molecule Exp.

Stochastic Models and Inference for Molecular Motors across Scales

JOHN FRICKS*,†

*Pennsylvania State University, University Park, USA

†email: fricks@stat.psu.edu 142:JohnFricks.tex,session:IS24

Linear molecular motors, such as kinesin and dynein, carry cargo, such as vesicles and organelles, through a cell along a microtubule network. The heads of these motors step along a microtubule and are on the order of nanometers, while the cargo size and the distance traveled can be on the order of hundreds or thousands of nanometers. Stochastic models of motors and motor-cargo complexes that describe both the mechanics and chemical kinetics will be presented. In addition, applications of limit theorems will be used to bridge these spatial and temporal scales along with a description of how data can be incorporated into these models at the various scales. Particular attention will be given to how data at one scale can inform biological mechanisms at other scales.

Acknowledgment. This research was partially supported by the US National Science Foundation, grant No.: DMS-0714939.

CS4A Time Series II.

Robust Shift Detection in Time Series

HEROLD DEHLING*, ROLAND FRIED†,‡, MARTIN WENDLER*

*Ruhr-Universität Bochum, Bochum, Germany,

†TU Dortmund University, Dortmund, Germany

‡email: fried@statistik.tu-dortmund.de 143:RolandFried.tex,session:CS4A

We present a robust test for detection of a shift in a time series which is based on the two-sample Hodges-Lehmann estimator. New limit theory for a class of statistics based on two sample U-statistics and two-sample U-quantile processes allows us to derive the asymptotic distribution of our test statistic under the null hypothesis in case of short range dependent data. We study the finite sample properties of our test via simulations and compare the test with the classical CUSUM test.

Interacting Particles with Different Jump Rates, Warren's Process with Drifts, and the Perturbed GUE Minor Process

OCS16 Interacting Particles

PATRIK L. FERRARI*, RENÉ FRINGS*,†

*University of Bonn, Germany
†email: frings@uni-bonn.de

144:ReneFrings.tex,session:OCS16

Interacting particles with different jump rates. The Borodin-Ferrari model [Borodin-Ferrari, 2008] generalizes the classical TASEP (totally asymmetric simple exclusion process) and describes the time evolution of $\frac{1}{2}N(N+1)$ interacting particles on a two-dimensional lattice, the so called (discrete) Gelfand-Tsetlin cone

$$\{(x^1, x^2, \dots, x^N) \in \mathbb{Z}^1 \times \mathbb{Z}^2 \times \cdots \times \mathbb{Z}^N : x_k^{n+1} < x_k^n \le x_{k+1}^{n+1}\}.$$

We consider a version of this process where the jump rates are level-dependent: Denote by $x_k^n \in \mathbb{Z}$ the position of a particle labeled by (k,n), with $1 \le k \le n \le N$, and call n the "level" of the particle. Particle (k,n) performs a continuous time random walk with one-sided jumps (to the right) of rate v_n. The interaction between levels is the following: (i) if particle (k,n) tries to jump to x and $x_{k-1}^{n-1} = x$, then the jump is suppressed, and (ii) when particle (k,n) jumps from $x-1$ to x, then all particles labeled by $(k+\ell, n+\ell)$ for some $\ell \ge 1$, which were at $x-1$, are forced to jump to x too. These two conditions ensure that the particle system has the discrete Gelfand-Tsetlin pattern as its state space.

Warren's process with drifts. Next, we take a diffusion scaling with appropriately scaled jump rates in the interacting particle model,

$$t = \tau T, \quad x_k^n = \tau T - \sqrt{T}\, W_k^n, \quad v_n = 1 - \frac{\mu_n}{\sqrt{T}}.$$

In the $T \to \infty$ limit, the particle process $\{x_k^n(t) : 1 \le k \le n \le N, t \ge 0\}$ will converge to Warren's process [Warren, 2007] $\{W_k^n(\tau) : 1 \le k \le n \le N, \tau \ge 0\}$ with drift vector $(\mu_1, \dots, \mu_N)$. The latter is constructed as follows: Consider a Brownian motion W_1^1 with drift μ_1 starting from the origin and then take two independent Brownian motions W_1^2 and W_2^2 having the same drifts μ_2 conditioned to start at the origin and, whenever they touch W_1^1, they are reflected off W_1^1. Similarly for $n \ge 2$, W_k^n is a Brownian motion with drift μ_n conditioned to start at 0 and being reflected off W_k^{n-1} (for $k \le n-1$) and W_{k-1}^{n-1} (for $k \ge 2$). By construction, this process lives in the non-discrete Gelfand-Tsetlin cone, a subset of $\mathbb{R}^1 \times \cdots \times \mathbb{R}^N$.

Perturbed GUE minor process. It turns out that Warren's process also appears in random matrix theory as a variant of the GUE minor process [Johansson-Nordenstam, 2006]. Consider an $N \times N$ Hermitian matrix H with eigenvalues $\lambda_1^N \le \cdots \le \lambda_N^N$ and denote by $\lambda_1^n \le \cdots \le \lambda_n^n$ the ordered eigenvalues of the submatrix that is obtained by keeping the first n rows and columns of H. In particular, $H^N = H$ and $H^1 = H_{11}$. The collection of all these eigenvalues $(\lambda^1, \dots, \lambda^N)$ then also forms a (non-discrete) Gelfand-Tsetlin pattern. Next, we let evolve these matrices in time and perturb them: We take $\{H(t) : t \ge 0\}$ to be a GUE matrix diffusion perturbed by a deterministic drift matrix $M = \mathrm{diag}(\mu_1, \dots, \mu_N)$, i. e., we consider $H(t) + t\,M$ with H evolving as the standard GUE Dyson's Brownian motion starting from the origin. This resulting eigenvalue process $\{\lambda_k^n(t) : 1 \le k \le n \le N, t \ge 0\}$ is the same as Warren's process $\{W_k^n(t) : 1 \le k \le n \le N, t \ge 0\}$ described above with identical drift vector $(\mu_1, \dots, \mu_N)$.

Acknowledgment. This research was supported by the German Research Foundation via the SFB611-A12 project.

IS18 Rough Paths

Applications of Rough Paths to Stochastic Control and Filtering

PETER FRIZ*,†,‡

*Department of Mathematics, TU Berlin,
†WIAS Berlin
‡email: friz@math.tu-berlin.de, friz@wias-berlin.de 145:FrizPeter.tex,session:IS18

We start with a discussion of optimally controlled rough differential equations. The value function is seen to satisfy a HJB type equation with "rough" time depedence. Deterministic problems of this type arise in the duality theory for controlled diffusion processes and typically involve anticipating stochastic analysis. We propose a formulation based on rough paths and then obtain a generalization of Roger's duality formula [Rogers, 2007] from discrete to continuous time. We also make the link to old work of [Davis–Burstein, 1987].

In the second part of the talk we discuss robust filtering. In the late seventies, Clark [The design of robust approximations to the stochastic differential equations of nonlinear filtering, Communication systems and random process theory, 1978] pointed out that it would be natural for π_t, the solution of the stochastic filtering problem, to depend continuously on the observed data $Y = \{Y_s, s \in [0,t]\}$. Indeed, if the signal and the observation noise are independent one can show that, for any suitably chosen test function f, there exists a continuous map θ_t^f, defined on the space of continuous paths $C([0,t],\mathbb{R}^d)$ endowed with the uniform convergence topology such that $\pi_t(f) = \theta_t^f(Y)$, almost surely. As shown by Davis [Pathwise nonlinear filtering, Stochastic systems: the mathematics of filtering and identification and applications, 1981] this type of *robust* representation is also possible when the signal and the observation noise are correlated, provided the observation process is scalar. For a general correlated noise and multidimensional observations such a representation does not exist. By using the theory of rough paths we provide a solution to this deficiency: The observation process Y is "lifted" to the process $\mathbf{Y}$ that consists of Y and its corresponding Lévy area process and we show that there exists a continuous map θ_t^f, defined on a suitably chosen space of Hölder continuous paths such that $\pi_t(f) = \theta_t^f(\mathbf{Y})$, almost surely.

References

[1] D. Crisan, J. Diehl, P. Friz, H. Oberhauser: Robust Filtering, Correlated Noise and Multidimensional Observation, arXiv: 1201.1858 and to appear in Annals of Applied Probability

[2] J. Diehl, P. Friz, P. Gassiat; Stochastic control with rough paths; arXiv: 1303.7160

[3] P. Friz, N. Victoir: Multidimensional stochastic processes as rough paths. Theory and applications. Cambridge University Press, Cambridge, 2010.

[4] Lyons, Terry: Differential equations driven by rough signals. Revista Matematica Iberoamericana 14.2 (1998): 215-310.

IS18 Rough Paths

Density of Solutions to Gaussian Rough Differential Equations

SAMY TINDEL*,†

*Université de Lorraine, Nancy, France
†email: samy.tindel@univ-lorraine.fr 146:SamyTindel.tex,session:IS18

Let $B = (B^1, \dots, B^d)$ be a d dimensional centered Gaussian process, whose components B^j are supposed to be i.i.d. One of the main example of this kind of process is provided by fractional Brownian motion with Hurst parameter $H > 1/4$. Recall that it means that the components B^j are i.i.d and satisfy the relation $\mathbb{E}[(B_t^j - B_s^j)^2] = (t-s)^{2H}$, so that H roughly represents the Hölder

continuity exponent of B. Nevertheless, we aim at covering general cases of Gaussian processes with some smoothness and nondegeneracy conditions on their covariance function R.

We are concerned here with the following class of equations driven by B:

$$X_t^x = x + \int_0^t V_0(X_s^x)ds + \sum_{i=1}^{d} \int_0^t V_i(X_s^x)dB_s^i, \tag{1}$$

where x is a generic initial condition and $\{V_i;\, 0 \leq i \leq d\}$ is a collection of smooth and bounded vector fields of $\mathbb{R}^m$. The unique solution to equation (1) is understood thanks to the rough paths theory.

Once equations like (1) are solved, it is natural to wonder how the density p_t of the random variable X_t behaves for an arbitrary strictly positive t. The first aim of this talk will be to present a smoothness result for p_t under very general conditions for the covariance function R of B, and assuming the classical Hörmander conditions on the vector fields $V_0, \ldots, V_d$. This result is contained in a joint work with T. Cass, M. Hairer and C. Litterer, and covers the fBm case for $1/4 < H < 1/2$.

We will then move to upper and lower Gaussian bounds for our density p_t, for which we work under standard elliptic assumptions:

$$V(z)\, V^*(z) \geq \epsilon\, \mathrm{id}_n, \quad \text{for all} \quad z \in \mathbb{R}^n, \tag{2}$$

where V stands for the matrix $(V^1, \ldots, V^d)$. Under this kind of assumptions, we are able to prove the following kind of theorem (taken from from joint works with M. Besalú, A. Kohatsu-Higa, F. Baudoin, E. Nualart and C. Ouyang):

Theorem 3. *Let B be a d-dimensional fBm, X the solution to (1) and V a smooth and bounded coefficient satisfying relation (2). Then if $H > 1/2$ and $t \in (0, 1]$ the density $p_t(z)$ of y_t satisfies*

$$\frac{c_1}{t^{nH}} \exp\left(-\frac{c_2\,|z-a|^2}{t^{2H}}\right) \leq p_t(z) \leq \frac{c_3}{t^{nH}} \exp\left(-\frac{c_4\,|z-a|^2}{t^{2H}}\right),$$

for strictly positive constants $c_1, \ldots, c_4$ depending on n, d, V, H.

Notice that up to constants, our bounds seem to be optimal in the sense that they mimic the Gaussian behavior of the underlying fBm itself. We shall also discuss generalizations to the case $H < 1/2$, for which methods of proof all rely on Gaussian analysis combined with rough paths techniques.

Modelling Multivariate Financial Returns Using Changepoint-Induced Multiscale Bases

IS6 Financial Time Ser.

PIOTR FRYZLEWICZ*,†, ANNA LOUISE SCHRÖDER*

*London School of Economics, UK
†email: p.fryzlewicz@lse.ac.uk

147:PIOTR_FRYZLEWICZ.tex,session:IS6

We motivate and describe a new model for multivariate financial returns, based on the the (adaptive) Unbalanced Haar basis. The use of this basis leads to a sparsely parameterised, interpretable and flexible model with naturally induced nonstationarity in the mean. The model provides a hierarchical description of the main features of the data, and offers the concept of a recoverable Unbalanced Haar 'spectrum'. A defining feature of this framework may be the use of an adaptive basis in its construction, which differs from the usual Fourier or fixed wavelet bases.

We discuss model-based forecasting of returns, as well as volatility matrix estimation, also in high dimensions.

CS25A Stoch. Finance I.

Strong Consistency of Maximum Likelihood Estimators of AR Parameter for a HJM Type Interest Rate Model

ERIKA FÜLÖP*,‡, GYULA PAP†

*University of Debrecen, Debrecen, Hungary,
†University of Szeged, Szeged, Hungary
‡email: fulop.erika@inf.unideb.hu 148:ErikaFulop.tex,session:CS25A

We consider a discrete time Heath-Jarrow-Morton (HJM) type forward interest rate model, where the interest rate curves are driven by a geometric spatial autoregression field. Such models were proposed by Gáll, Pap and Zuijlen [1]. Let $\{\eta_{i,j} : i,j \in \mathbb{Z}_+\}$ be i.i.d. standard normal random variables on a probability space $(\Omega, \mathcal{F}, P)$, where $\mathbb{Z}_+$ denotes the set of nonnegative integers. Let $\varrho \in \mathbb{R}$ be the autoregression coefficient. Let $f_{k,\ell}$ denote the forward interest rate at time k with time to maturity date ℓ $(k, \ell \in \mathbb{Z}_+)$. We assume that the initial values $f_{0,\ell}$ are known at time 0. If this market is arbitrage free then the forward rates are given by the following equations (see [1]):

$$f_{k,\ell} = f_{k-1,\ell+1} + \varrho(f_{k,\ell-1} - f_{k-1,\ell}) + \eta_{k,\ell} + \frac{1}{2}\sum_{i=0}^{2\ell} \varrho^i, \qquad (k, \ell \in \mathbb{N}).$$

The process $\{f^{(\varrho)}_{k,\ell} : k, \ell \geq 0\}$ is a spatial autoregressive process. It is called *stable*, *unstable*, or *explosive*, if $|\varrho| < 1$, $|\varrho| = 1$, or $|\varrho| > 1$, respectively.

Our aim is to test the autoregression parameter ϱ. When the sample is $(f^{(\varrho)}_{k,\ell})_{1\leq k\leq K_n, 0\leq \ell\leq L_n}$, where $K_n = nK + o(n)$ and $L_n = nL + o(n)$ with some $K, L > 0$, we proved strong consistency of maximum likelihood estimators in the stable and unstable cases [4] by the help of checking general conditions given in [3]. We also consider the sequence of statistical experiments in the paper [2]. Now we look another realistic model when the time to maturity date $L_n := L$ $(n \in \mathbb{N})$ is constant i.e. our sample proportional to n instead of n^2. The difficulty is that the underlying sample consists of nonindependent random variables. Moreover, no explicit formula is available for the maximum likelihood estimators of ϱ.

References

[1] Gáll, J. and Pap, Gy. and Zuijlen, M.v., 2006: Forward interest rate curves in discrete time settings driven by random fields. *Comput. Math. Appl.*, **51** no. 3–4, 387–396.

[2] Fülöp, E. and Pap, G., 2007: Asymptotically optimal tests for a discrete time random field HJM type interest rate model, *Acta Sci. Math.* , **73(3-4)**, 637–661.

[3] Fülöp, E. and Pap, G., 2008: Note on strong consistency of maximum likelihood estimators for dependent observations, *Proc. of the 7th International Conference on Applied Informatics (Eger, 2007),* Volume 1, pp. 223–228.

[4] Fülöp, E. and Pap, G., 2009: Strong consistency of maximum likelihood estimators for a discrete time random field HJM type interest rate model, *Lithuanian Math.J.*, **49**(1):5–25.

CS9D Testing Mod. Structure

On Estimates of R-values in Multiple Comparison Problems

ANDREAS FUTSCHIK*,‡, WEN-TAO HUANG†

*University of Vienna, Vienna, Austria,
†Tamkang University, Taipei, Taiwan
‡email: andreas.futschik@univie.ac.at 149:AndreasFutschik.tex,session:CS9D

In applications involving large numbers of hypotheses, it is well known that subset selection rules

can be quite conservative in the sense that the actual probability of correct selection is considerably above the target value, if the parameter configuration is not least favorable. We therefore propose an approach to better adapt to the actual parameter configuration. In order to ensure a wide applicability, we phrase our approach in terms of R-values. These quantities are related to p-values and have been introduced by Hsu (1984). Related approaches have been recently considered in multiple hypothesis testing where approximate procedures have been proposed that adapt to the actual number of true null hypotheses in situations where the number of parameters is large.

Both theoretical and simulation results indicate a desirable performance.

Resampling Methods for Weakly Dependent and Periodically Correlated Sequences

OCS26 Resampling Nonstat T.S.

ELZBIETA GAJECKA-MIREK*,†

*State Higher Vocational School, Nowy Sacz, Poland
†email: egajecka@gmail.com 150:ElzbietaGajeckaMirek.tex,session:OCS26

Many authors have used the mixing properties as a type of dependence in time series. Unfortunately some time series do not satisfy any mixing condition. In 1999 P. Doukhan introduced a new type of dependence in time series - weak dependence, which is weaker than the strong mixing property. This gives tools for the analysis of statistical procedures with very general data generating processes. One of such statistical procedures is resampling.

Resampling can be used if statistical inference for dependent data based on asymptotic distributions fails or there are problems with sample size. To apply subsampling it is sufficient to know if there exists a non-degenerated asymptotic distribution of the statistic.

For independent data and stationary time series resampling procedures are well investigated. Our research is focused on non-stationary periodically correlated time series.

In the poster the generalization of the model introduced by Politis and McElroy in 2007 is considered for periodically correlated processes with a known period.

The model investigated in the poster is: $X_t = \sigma_t G_t + \eta_t$, where σ_t and G_t are independent, σ_t, is i.i.d. mean μ different from zero and has the marginal distribution of an $\alpha-$stable random variable. Moreover G_t is periodically correlated time series with known period T and can be written as $G_t = f_t N_t$ for a long memory, stationary mean zero Gaussian process N_t, and f_t- bounded and scalar periodic sequence $f_t = f_{t+T}$, $\eta_t(= \eta(t))$ is periodic with the same, known period T as f_t.

In such model, the joint asymptotic behavior of the sample mean and the sample variance is investigated. The weak dependence property gives the tools to improve the subsampling consistency of self-normalized statistics.

Additionally the simulations and real data example will be given.

Multiple Break Detection in the Correlation Structure of Random Variables

OCS8 Time Series

PEDRO GALEANO*,‡, DOMINIK WIED†

*Universidad Carlos III de Madrid, Spain,
†TU Dortmund, Germany
‡email: pedro.galeano@uc3m.es 151:Pedro_Galeano.tex,session:OCS8

Correlations between random variables play an important role in applications, e.g., in financial analysis. More precisely, accurate estimates of the correlation between financial returns are crucial in

portfolio management. In particular, in periods of financial crisis, extreme movements in asset prices are found to be more highly correlated than small movements. It is precisely under these conditions that investors are extremely concerned about changes on correlations. Wied, Krämer and Dehling (2012) propose a CUSUM type procedure along the lines of Ploberger, Krämer and Kontrus (1989) to formally test if correlations between random variables remain constant over time. However, with this approach the practitioner is only able to see if there is a change or not; he cannot determine where a possible change occurs or how many changes there are. The present paper fills this gap by proposing an algorithm based on the correlation constancy test to estimate both the number and the timing of possible change points. For this purpose, we propose a binary segmentation procedure to detect the number and position of multiple change points in the correlation structure of random variables. The segmentation algorithm proceeds as follows: First, we determine the "dominating" change point and decide if this point is statistically significant. Then, we split the series in two parts and again test for possible change points in each part of the series. The procedure stops if we do not find any new change point any more. In this paper, we will analytically show that the algorithm asymptotically gives the correct number of change points and that - finitely many - change points are consistently estimated. Furthermore, we show that the algorithm gives reasonable results in finite samples through a large simulation study. Also, we apply the procedure to look for changes in the correlation structure of the log-return series of the Standard & Poors 500 Index and the IBM stock.

Acknowledgment. Financial support by MCI grants MTM2008-03010 and ECO2012-38442 and Deutsche Forschungsgemeinschaft (SFB 823, project A1) is gratefully acknowledged.

References

[Ploberger et al (1989)] Ploberger, W., Krämer, W. and Kontrus, K. 1989: A new test for structural stability in the linear regression model, *Journal of Econometrics*, **40**, 307 - 318.

[Wied et al (2012)] Wied, D. and Krämer, W. and Dehling, H. 2012: Testing for a change in correlation at an unknown point in time using an extended functional delta method, *Econometric Theory*, **28**, 570 - 589.

CS24A Branching Proc.

On Cycle Representations of Discrete Time Birth and Death Processes

CHRYSOULA GANATSIOU*

*University of Thessaly, Department of Civil Engineering,Volos, Greece

152:ChrysoulaGanatsiou.tex,session:CS24A

Following the context of cycle-circuit representation theory of Markov processes, regarding the representations of the finite-dimensional distributions of stochastic processes of Markovian type with discrete or continuous parameter having an invariant measure as linear or convex decompositons in terms of the cycle-circuit passage functions, the present research arises as an attempt to investigate the type problem (transience/recurrence) for the discrete time birth and death processes. In particular, the paper provides suitable criterions for positive/null recurrence and transience for the corresponding ad joint Markov chains describing uniquely the discrete time birth and death process by directed circuits and weights.

References

[1] Ganatsiou, Ch. , 2011: On cycle representations of random walks in fixed, random environments, *58th World Statistics Congress of the International Statistical Institute, Dublin, 21-26/8/2011.*

[2] Kalpazidou,S., 1995: Cycle representations of Markov Processes, *Springer-Verlag, New York.*

[3] MacQueen,J., 1981:Circuit processes,*Ann. Probab.*, **9**, 604-610.

[4] Minping,Qian, Min, Qian, 1982: Circulation of recurrent Markov chain, *Z. Wahrsch. Verw. Gebiete*,**59**,203-210.

MMCTest - A Safe Algorithm for Implementing Multiple Monte Carlo Tests

NYA
Not Yet Arranged

AXEL GANDY[*,†], GEORG HAHN[*]

[*]Imperial College London
[†]email: a.gandy@imperial.ac.uk 153:AxelGandy.tex,session:NYA

This talk discusses testing multiple hypotheses using tests that can only be evaluated by simulation such as permutation tests or bootstrap tests. In particular, it presents MMCTest, a sequential algorithm which modifies standard procedures which work with exact p-values, such as the Benjamini & Hochberg False Discover Rate (FDR) procedure. The algorithm gives the same classification as the original procedure with the exact p-values, up to an error probability that can be chosen by the user. The method also extends to controlling the Familywise Error Rate using the Bonferroni correction. At any stage, MMCTest can be interrupted and returns sets of hypotheses which can already be classiffied with satisfactory precision as being rejected or non-rejected and a set of hypotheses whose decision is still pending. A simulation study on actual biological data, given by a microarray dataset of gene expressions, shows that for a realistic precision, MMCTest draws level with the performance of current methods which unlike MMCTest do not give a guarantee on its classification being correct. An ad-hoc variant of MMCTest which forces a complete classification outperforms established methods.

Resampling Methods for Spatial Data with Applications

POSTER
Poster

PILAR GARCIA-SOIDAN[*,‡], RAQUEL MENEZES[†], OSCAR RUBINOS-LOPEZ[*]

[*]University of Vigo, Spain,
[†]University of Minho, Portugal,
[‡]email: pgarcia@uvigo.es 154:PilarGarciaSoidan.tex,session:POSTER

Inference in the spatial setting entails dealing with dependent data, which demands requiring hypotheses from the random process and making use of specific strategies to achieve reliable results. In this respect, different techniques have been designed for spatial data to address the characterization of the dependence structure or prediction at unsampled locations. The use of these approaches can require extra work to check the accuracy of the estimators employed or to derive their sampling distribution. Some of the aforementioned problems have been solved through a resampling method referred to as Bootstrap. The original Bootstrap procedures were designed for independent data by assuming knowledge of the distribution model or without the latter constraint, which are respectively called parametric or nonparametric Bootstrap methods. The former ones can be more easily adapted to the spatial setting, whereas the latter methods must be specifically redesigned for spatial data so as to guarantee consistency of the results. In this work, we describe nonparametric Bootstrap approaches to generate replicates from the available spatial data, observed at a set of locations, by first constructing an estimator of the joint distribution in a nonparametric way and then randomly drawing samples from it. Consistency follows for the suggested procedures, provided that the random process is strictly stationary or when this condition is relaxed by admitting a deterministic trend.

Applications of the resampling approaches can be found, for instance, to characterize the dependence structure of the random process. For the latter aim, the variogram or the covariance functions are typically approximated, depending on whether the spatial process is assumed to be intrinsic or second-order stationary, respectively. In a first step, the nonparametric methods may be applied for approximation of the latter functions, although they cannot be used directly for prediction by using

the kriging equations. We can cope with this problem by selecting an appropriate parametric model and then deriving optimal estimates of the parameters. However, the model misspecification is one of its main drawbacks. To overcome this issue, we will derive nonparametric tests for modelling spatial dependence of an intrinsic random process, which will be used to check the hypothesis of isotropy and to test the goodness of fit of a parametric model selected. The critical points of the referred tests need to be estimated, as they depend on unknown terms, and the normal approximation is not recommended for the latter aim, because of its slow rate of convergence. Thus, an alternative strategy for approximation of the critical points will be proposed in this work, based on the Bootstrap techniques.

Acknowledgment. The first and third authors acknowledge financial support from the projects TEC2011-28683-C02-02 and CONSOLIDER-INGENIO CSD2008-00068 of the Spanish Ministry of Science and Innovation. The second author's work has been supported by the project PTDC/MAT/112338/2009 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education.

References

[Efron (1979)] Efron, B., 1979: Bootstrap methods: another look at the Jackknife, *Ann Statist*, **7**, 1 - 26.

[Hall (1985)] Hall, P., 1985: Resampling a coverage pattern, *Stoch. Proces. Applic.*, **20**, 231 - 246.

CS6F Copulas

Nonparametric Independence Test Based on Copula Density

GERY GEENENS*,†

*University of New South Wales, Sydney, Australia

†email: geenens@unsw.edu.au 155:GeryGeenens.tex,session:CS6F

The concept of independence is central in statistics, and being able to test the assumption of independence between two random variables X and Y is evidently very important. In this work, such an independence test is proposed. It is able to detect any departure from the null hypothesis of independence (omnibus test) between two continuous random variables X and Y, without relying on any particular parametric assumptions on the underlying distributions (nonparametric test). The test is based on copula modelling, which has lately become a very popular tool for analysing the dependence structure of a random vector. Specifically, the test statistic is a Cramér-von Mises-type discrepancy measure between a (boundary-bias-corrected) kernel estimate of the copula density of (X, Y) and the independence copula density. Basing an independence test on the copula density has numerous advantages that will be discussed. In particular, this makes the test very powerful at detecting subtle departures from the null hypothesis of independence in any direction. This is explained through theoretical considerations and illustrated by substantial simulation studies.

CS1A Shape & Image

New Methods for Horizon Line Detection in Infrared and Visible Sea Images

ILAN LIFSHITZ*, EVGENY GERSHIKOV*,†,‡

*Ort Braude Academic College of Engineering, Karmiel, Israel,

†Technion - Israel Institute of Technology, Haifa, Israel

‡email: eugeny11@braude.ac.il 156:EvgenyGershikov.tex,session:CS1A

In this work we propose methods for horizon line detection in marine images captured by either infrared or visible light cameras. A common method for horizon line detection is based on edge detection [Canny (1986)] followed by the Hough transform [Duda and Hart (1972)]. This method

suffers from serious drawbacks when the horizon is not a clear enough straight line or there are other straight lines present in the image. We improve the algorithm performance by proposing a pre-processing stage eliminating some of the false detections.

We also propose a new method for horizon line detection. The new algorithm is based on the idea of segmenting the image into two regions (sky and sea) by comparing their regional probability distribution functions (PDFs). These PDFs can be approximated by the region histograms, calculated using the available image gray level statistics. The line maximizing a statistical distance between these PDFs among a set of candidate lines is chosen. The candidate lines examined can be lines at all possible positions and orientations or just a subset of this, e.g., in a given region of interest in the image. For marine images the horizon orientation is usually close to horizontal, thus only small line angles (measured relative to the horizontal axis) can be checked. Still this method has high computational complexity that can be reduced without loss of performance quality by reducing the size of the examined candidate line set while keeping the important candidates.

Thus, we proceed to introduce two new methods combining the basic edge detection and Hough transform method to detect several candidate lines in the image and a statistical criterion to find the optimal line among the candidates. The chosen criterion can be based on regional covariances [Ettinger et al. (2002)] or the distance between PDFs as described above. We show that choosing the second option provides the best performance among the methods tested and yields a low complexity fast algorithm.

We compare all methods quantitatively by their accuracy and relative speed as well as visually for several example images. Our conclusion is that the algorithms introduced in this work, based on statistical methods, can be beneficial for automatic processing of marine images for the purposes of tracking, navigation, target recognition or other applications.

Acknowledgment. We would like to thank the administration of Ort Braude Academic College and the Department of Electrical Engineering for providing the opportunity and financial means to conduct this research.

References

[Canny (1986)] Canny, J., 1986: A computational approach to edge detection, *IEEE Transactions on PAMI*, **8(6)**, 679-697.

[Duda and Hart (1972)] Duda, R. O., Hart, P. E., 1972: Use of the Hough transformation to detect lines and curves in pictures", *Comm. ACM*, **15(1)**, 11-15.

[Ettinger et al. (2002)] Ettinger, S., Nechyba, M., Ifju, P. , Waszak, M., 2002: Vision-guided flight stability and control for micro air vehicles, *IEEE Conf. on Intelligent Robots and Sys.*, 2134-2140.

Copula-based Semiparametric Quantile Regression

CS6F Copulas

ANOUAR EL GHOUCH*,†, HOHSUK NOH*, INGRID VAN KEILEGOM*

*Université catholique de Louvain, Louvain-La-Neuve, Belgium.

†email: anouar.elghouch@uclouvain.be 157:AnouarElGhouch.tex,session:CS6F

We propose a new approach to quantile regression modeling based on the copula function that defines the dependence structure between the variables of interest. The key idea for this approach is to rewrite the characterization of a regression quantile in terms of a copula and marginal distributions. After the copula and the marginal distributions are estimated, the new estimator is obtained as the weighted quantile of the response variable in the model. Along with the large and growing literature of copula, this approach provides a rich and flexible class of quantile regression estimators. We establish the asymptotic theory of our estimator when the copula is estimated by maximizing the pseudo log-likelihood and the margins are estimated nonparametrically including the case where

the copula family is misspecified. We also present finite sample performance of the estimator and illustrate its usefulness with real data.

The method is shown to be flexible, easy to implement and less influenced by the curse of dimensionality than competitors. This advantage, however, is attained at the price of a model risk in the copula modeling. In this work, we empirically checked that such a model risk is small when the copula family and the copula parameters are selected by the data but some theoretical analysis about the additional risk stemming from the model selection step is needed.

Acknowledgment. The authors acknowledge financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

OCS25 Long-mem. Time Ser.

Studentizing Weighted Sums of Linear Processes

VIOLETTA DALLA*, LIUDAS GIRAITIS†, HIRA L. KOUL‡

*University of Athens, Greece,
†Queen Mary, University of London, UK,
‡Michigan State University, East Lansing, USA

158:LiudasGiraitis.tex,session:OCS25

This paper presents a general method for studentizing weighted sums of a linear process where weights are arrays of known real numbers and innovations form a martingale difference sequence. This paper centers on estimation of the standard error, to make the normal approximation operational. Suggested studentization is easy to apply and robust against unknown type of dependence (short range and long range) in the observations. It does not require the estimation of the parameters controlling the dependence structure. A finite sample Monte-Carlo simulation study shows the applicability of the proposed methodology for moderate samples. Assumptions for studentization are satisfied by kernel Nadaraya-Watson type weights used for inference in non-parametric regression settings.

IS20 Space-Time Stat.

Positive Definite Functions on Spheres

TILMANN GNEITING*,†

*Heidelberg University, Germany
†email: t.gneiting@uni-heidelberg.de

159:Gneiting.tex,session:IS20

Positive definite functions on spheres play important roles in spatial statistics, where they occur as the correlation functions of random fields and star-shaped random particles. We review Schoenberg's classical characterization of the isotropic positive definite functions on spheres in terms of Gegenbauer expansions, and apply them to dimension walks, where monotonicity properties of the Gegenbauer coefficients guarantee positive definiteness in higher dimensions. Subject to a natural support condition, isotropic positive definite functions on the Euclidean space $\mathbb{R}^3$, such as Askey's and Wendland's functions, allow for the direct substitution of the Euclidean distance by the great circle distance on a one-, two- or three-dimensional sphere, as opposed to the traditional approach, where the distances are transformed into each other. Completely monotone functions are positive definite on spheres of any dimension and provide rich parametric classes of such functions, including members of the powered exponential, Matérn and generalized Cauchy families. The sine power family permits a continuous parameterization of the roughness of the sample paths of Gaussian processes on spheres. A set of sixteen research problems provides challenges for future work in mathematical analysis, probability theory, and spatial statistics.

A New Measure Of Association For Doubly Ordered Cross Tables

CS5B H-D Distribution

ATİLLA GÖKTAŞ*,‡, ÖZNUR İŞÇI*, PINAR GÖKTAŞ†

*Muğla Sıtkı Koçman University, Faculty of Sciences, Department of Statistics, Muğla/TURKEY,
†Muğla Sıtkı Koçman University, Faculty of Economic and Administrative Sciences, Department of Economy , Muğla/TURKEY
‡email: gatilla@mu.edu.tr 160:AtillaGoktas.tex,session:CS5B

Ordinal measures of association have been widely used in social sciences and some specific area of applied sciences for decades. Morover in most of the popular statistical packages do calculate those measures and present the statistical tests as well. Those measures have never been compared either analytically or emprically, until the recent studies performed by Göktaş and .Işçi in 2011. In their studies they compared the most commonly used measures of association for square tables and extend their studies in the next paper for rectangular doubly ordered cross tables checking both normality and the nominal level with emprical results. Both of their works have proven that especially for small sample sizes the Kurskall Gamma coefficient performed better than the other well-known ordinal measures of association. However the Kruskall Gamma is underestimating the actual degree of association for large tables whereas it is overestimating the true degree of association for small sample sizes.

The aim of this study is to present a new measure of association that is free of table dimension and sample size which is in fact robust. The new measure of association has also proven that both the underestimation and overestimation has been considerably adjusted.

Modeling Inflation Uncertainty: Case Of Turkey

CS25C Stoch. Finance II.

PINAR GÖKTAŞ*,†, CEM DIŞBUDAK*

*Muğla Sıtkı Koçman University, Faculty of Economic and Administrative Sciences, Department of Economy , Muğla/TURKEY
†email: pinargoktas@mu.edu.tr 161:Pinar_GOKTAS.tex,session:CS25C

The aim of this study is to build a generalized autoregressive conditional heteroskedasticity (GARCH) model for determining inflation variability and test the association between inflation variability and level of inflation rate. To realize this we first use the original monthly consumer price index series between 1994:01 and 2012:12 and convert it using the year over year transformation to generate the raw monthly inflation series between 1995:01 and 2012:12. In most studies there is always a positive correlation between level of inflation and inflation volatility no matter how devloped the country is. The test results of the model we obtained for case of Turkey indicate that there is a relationship between inflation and inflation variability and it is expressed in Granger Causality in both ways. In conlusion high inflation in Turkey leading to high volatility in inflation or vice versa does not contradict any study accross either developed or developing countries.

CS1A Shape & Image

Computing Intrinsic Mean Shape on Similarity Shape Spaces using a Highly Resistant Algorithm

MOUSA GOLALIZADEH[*,†], HAMIDREZA FOTOUHI[*]

[*]Tarbiat Modares University, Tehran, Iran
[†]email: golalizadeh@modares.ac.ir 162:MOUSA_GOLALIZADEH.tex,session:CS1A

Gradient Descent Algorithm (GDA) plays a key role in some of the statistical problems. Particularly, among many algorithms, it is a simple tool to derive an optimal quantity in dealing with an optimization problem in the linear space. However, most of the optimization procedures in the statistical shape analysis, which deals with the geometrical aspects of the objects invariant to the location, scale and rotation effects, are problems lending themselves to the non-linear space. Hence, one should pay a great attention to implement such optimization algorithms in this new fields of statistics. Particularly, while deriving the intrinsic mean shape, the geometry might be lost if the step size, threshold value and other elements of the GDA are not properly tuned. To both preserve the geometry and accelerate the convergence rate, we propose a dynamic step size and a new criterion to obtain the intrinsic mean on similarity shape space. We call it a Robust Gradient Descent Algorithm (RGDA).

A common Riemannian metric in the statistical shape analysis is the Procrustes distances and the mean shape is defined as the minimizer of the squared Procrustes distances. However, to derive a measure of variability among shapes and particularly to perform multivariate statistical analysis, the usual standard statistical tools, available on Euclidean space, cannot be directly utilized. So, among many methods to obtain the shape variance, the non-Euclidean shape space is approximated by a linearized space at the vicinity of the mean shape and then the Principal Component Analysis (PCA) is invoked there. Recently, a new tool which is called Principal Geodesic Analysis (PGA) was proposed to, directly, evaluate the variability on the curved manifolds including the shape space. The core basis of this method, which works well in some particular spaces, is mainly based upon GDA. Among many performance parameters of the PGA, the step size and the threshold value are key components on both accelerating the algorithm and guaranteeing the convergence to the optimum object.

In our talk, we will demonstrate how, at each stage of the GDA, an non-tuned step size both increases the time of convergence and fails to preserve geometrical structures of the objects before reaching the intrinsic mean shape. This is due to the fact that the optimum choice of step size parameter not only accelerate the convergence rate but also maintain the geometrical structure of the shape under study. It further helps the user to control the results in each step of the algorithm before reaching the optimum object which is the intrinsic mean shape in our case. Later, by introducing a new criterion for checking geometrical structure of objects, we propose a more sensible algorithm which works well in the shape analysis context. The performance of our proposed method is compared to the usual GDA in estimating the intrinsic mean shape of a real data set and also a simulation study.

IS5 Envtl. Epidem. Stat.

Prior Specification for Spatial Ecological Regression Models

ENRICO FABRIZI[*], FEDELE GRECO[†,‡], CARLO TRIVISANO[†]

[*]Catholic University of the Sacred Heart, Piacenza, Italy,
[†]University of Bologna, Bologna, Italy
[‡]email: fedele.greco@unibo.it 163:FedeleGreco.tex,session:IS5

We consider the problem of specifying priors for the variance components in the Bayesian analy-

sis of the Besag-York-Mollié (BYM) model, a model that is popular among epidemiologists for disease mapping. The model encompasses two sets of random effects: one spatially structured to model spatial autocorrelation and the other spatially unstructured to describe residual heterogeneity. In this model, prior specification for variance components is an important problem because these priors maintain their influence on the posterior distributions of relative risks when mapping rare diseases. This problem has only been partially addressed in a variety of papers on this topic, sometimes with controversial results. The choice of priors has a non-negligible influence on the posterior distribution of relevant parameters given the structure of the BYM model. Moreover, non-informative reference choices may lead to serious computational problems. In this talk we critically review some contributions on this topic widely used by practitioners. Moreover, we propose using Generalised Inverse Gaussian priors, a broad class of distributions that includes many distributions commonly used as priors in this context, such as inverse gammas. We discuss the prior parameter choice with the aim of balancing the prior weight of the two sets of random effects on total variation and controlling the amount of shrinkage. Even if the theoretical development of this strategy may appear quite complicated to practitioners, the final output turns out to be very intuitive. In fact, prior elicitation can be based on the specification of prior balance between structured and unstructured heterogeneity and on a prior guess on the variability of the relative risks or on the percentiles of their distribution in the study area. The suggested prior specification strategy is compared to popular alternatives using a simulation exercise and applications to real data sets in the case of a pure random effect model. Both the simulation study and the analysis of real case studies highlight that our prior specification strategy seems to produce better results in terms of fit and similar (or lower) amount of shrinkage when compared with results obtained under more popular priors. Some preliminary results in the context of spatial ecological regression are discussed.

Nonparametric Copula Estimation for Censored Data

CS6H Copula Estim.

SVETLANA GRIBKOVA[*,†], OLIVIER LOPEZ[*,‡]

[*]Université Pierre et Marie Curie, Paris, France
email: [†]svetlana.gribkova@etu.upmc.fr, [‡]olivier.lopez0@upmc.fr
164:GribkovaSvetlana.tex,session:CS6H

A copula is a cumulative distribution function (cdf) $\mathbb{C}(x_1, \ldots, x_d)$ on $[0,1]^d$ with uniform marginals, which joins the distribution function F of a random vector $(T_1, \ldots, T_d)$ to its marginal distributions $F_1(x_1), \ldots, F_d(x_d)$ by the relation $F(x_1, \ldots, x_d) = \mathbb{C}(F_1(x_1), \ldots, F_d(x_d))$. By Sklar's theorem, if the marginal df's are continuous, then the copula is unique. Copulas permit a separate modeling of the marginal distributions and the dependence structure.

The classical nonparametric copula estimator, based on fully observed data set is the empirical copula, proposed by Deheuvels. It is obtained by replacing the unknown distributions in the definition of copula by their empirical counterparts (empirical distribution functions). In the present work we are presenting an extension of this estimator to the framework of bivariate right censored data for some class of statistical models.

We are concerned with the estimation of the copula, linking two random variables T_1 et T_2, which are not fully observed. Under right random censoring instead of observing the variable T_1 (resp. T_2) one observes the variable $Y_1 = \min(T_1, C_1)$ (resp. Y_2) and $\delta_1 = \mathbb{I}_{T_1 \leq C_1}$ (resp. δ_2), where (C_1, C_2) is a couple of random variables supposed to be independent of the vector (T_1, T_2). Thus, the observed sample is of the form $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1 \leq i \leq n}$. As one does not observe the i.i.d. realizations of the vector (T_1, T_2), the empirical df and the empirical copula function are not available. Moreover, in general case of multivariate censored observations, one fails to construct an estimator of the df,

being consistent for all censoring schemes and defining at the same time a proper distribution function. However, for a large class of particular bivariate models that is possible and such estimator takes a general form $\hat{F}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \hat{W}_{in}\, \mathbb{I}_{Y_{1i}\le x, Y_{2i}\le y}$, where the random weights $\hat{W}_{in}$ depend on particular model assumptions. In that case, let us define our copula estimator by

$$\hat{\mathbb{C}}_n(u,v) = \hat{F}_n(\hat{F}_{1n}^{-1}(u), \hat{F}_{2n}^{-1}(v)),$$

and the corresponding empirical process by $\alpha_n(u,v) = \sqrt{n}(\hat{\mathbb{C}}_n(u,v) - \mathbb{C}(u,v))$. In the uncensored case, $\alpha_n(u,v)$ was studied by [1]. We show that, under some assumptions, $\alpha_n(u,v)$ converges weakly to some tight gaussian process in the space $l^\infty([0,1]^2)$ of uniformly bounded functions on $[0,1]^2$.

The defined copula estimator is a discrete function and can not be used directly for copula density estimation. Thus we construct its smoothed version, generalizing the approach of [2] to the case of censored data. Copula density estimator is then obtained from the smooth kernel estimator by derivation.

Asymptotic properties of the proposed estimators are investigated. Our results are illustrated by a simulation study and real data examples.

References

[1] Fermanian, J-D., Radulović, D. and Wegkamp, M., 2004: Weak convergence of empirical copula processes, *Bernoulli*, **10**, 847–860.

[2] Omelka, M., Gijbels, I., and Veraverbeke, N., 2009: Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing, *The Annals of Statistics*, **37**, 3023 – 3058.

OCS31 Stoch. Molecular Biol.

Novel Approximation Methods for Stochastic Biochemical Kinetics

RAMON GRIMA*,†

*University of Edinburgh, Edinburgh

†email: ramon.grima@ed.ac.uk

165:RamonGrima.tex,session:OCS31

It is well known that the kinetics of biochemical pathways is stochastic, a property stemming from the low molecule numbers of many chemical species. The traditional rate equations of physical chemistry cannot account for stochasticity and hence in recent years the stochastic simulation algorithm has increased in popularity and use. A major disadvantage of the latter is its computational inefficiency for moderately large sized pathways. In this talk I will present the Effective Mesoscopic Rate Equation (EMRE) and Inverse Omega Square (IOS) formalisms which provide accurate approximations to the mean concentrations and to the size of the fluctuations for a wide range of molecule numbers and which can be computed in a fraction of the time taken by the stochastic simulation algorithm. These linear equations provide a more accurate approximation than possible with the conventional linear-noise approximation and their application to various catalytic and gene regulatory systems has elucidated new kinetic laws and novel phenomena such as concentration inversions as well as explaining various features of experimental single cell oscillatory data. Finally I'll show case iNA (intrinsic noise analyzer), our new open-source software with a user-friendly graphical interface, which utilizes the aforementioned approximation methods to calculate the intrinsic noise statistics of a user-specified biochemical network.

Conformality, Criticality, and Universality in Two-Dimensional Stochastic Processes

Forum
Forum
Lecture

GEOFFREY GRIMMETT

University of Cambridge
email: grg@statslab.cam.ac.uk 166:GeoffreyGrimmett.tex,session:Forum

Two-dimensional physical systems have been studied intensively by physicists over the last fifty years. Despite many stimulating 'exact calculations' and the development of conformal field theory, *rigorous* mathematical progress has been slow. This serious lacuna was partly rectified around the year 2000 through the discovery by Oded Schramm of the *Schramm–Loewner evolution* (SLE), and the proof of conformality for percolation by Stanislav Smirnov (and others). These (and later) results by mathematicians have introduced novel methods and insights into this important field of science.

The target of these two Forum Lectures is to introduce (to a general statistical audience) the theory of critical stochastic processes in two dimensions. The three principal objects of study are the *percolation model*, the *Ising/Potts models* for the ferromagnet, and *self-avoiding walks*. The three properties under study are *conformality*, *criticality*, and *universality*. Emphasis will be placed on the many unsolved problems.

Percolation is a fundamental model for a disordered d-dimensional medium. The case $d = 2$ is special in that open clusters are blocked by *one-dimensional barriers*, and this permits certain exact calculations (the case $d = 3$ is, in a serious sense, still 'wide open'). We shall discuss the connection between critical percolation and conformal functions on the complex plane. Whereas a near-complete solution is now known for one very special percolation model, the corresponding picture for general models remains only a conjecture.

In contrast, it has been proved (with Ioan Manolescu) that a large and natural collection of *bond percolation models on so-called isoradial graphs* are critical, and form a single universality class. This analysis extends the Harris–Kesten theorem for the square lattice to a large family of (a)periodic graphs including Penrose tilings.

The *Ising model* is the 'standard' model for the ferromagnet. It is especially harmonious when the underlying graph is isoradial, for which case Dmitry Chelkak and Stanislav Smirnov have proved criticality and universality.

The third object of study in these two lectures is the *self-avoiding walk*. This was introduced before 1953 by Paul Flory as a model for polymerization in physical chemistry. There are a number of fascinating combinatorial and probabilistic questions concerning their number and their typical shape. We shall discuss recent combinatorial results and conjectures (with Zhongyang Li), and also the notorious conjecture that the asymptotic shape of a random n-step self-avoiding walk on the square lattice converges, when rescaled, to the random curve $\mathrm{SLE}_{8/3}$.

Paradifferential Calculus and Controlled Distributions

IS27
SPDE

MASSIMILIANO GUBINELLI[*,‡], PETER IMKELLER[†], NICOLAS PERKOWSKI[†]

[*]CEREMADE UMR 7534, Université Paris–Dauphine,
[†]Institut für Mathematik, Humboldt–Universität zu Berlin
[‡]email: gubinelli@ceremade.dauphine.fr 167:MassimilianoGubinelli.tex,session:IS27

We combine notions from paradifferential calculus and rough path theory to provide a framework to study non-linear operations on multi-dimensional distributions arising from stochastic PDEs or generally from singular PDEs.

Paradifferential calculus has been introduced in the '80 by Bony and others to study singularities of non-linear PDEs. Rough path analysis has been introduced by Lyons' in the late '90 to provide a robust analytical framework to study of differential equations driven by singular signals. Controlled distributions seems to be a natural language in which to reunite these two different area of analysis to tackle problems in which both the analytical and the stochastic aspects play important roles.

The resulting theory is a natural multidimensional generalisation of the controlled approach to the solutions of rough differential equations which abstracts on the particular structure of the differential operators and on the space dimensionality. Paraproducts and in general paradifferential operators provide a natural language in which to describe *controlled* distributions, that is distributions which locally "looks like" given reference distributions, usually quite singular. Non-linear operations on controlled distributions are synthetized using informations on the non-linear structure of the underlying reference distributions. These informations are usually the results of statistical properties of these objects like independence and scaling. We present several examples of applications of this approach: a multi-dimensional stochastic Burger-like equation, a two-dimensional parabolic Anderson model and a reformulation of Hairer's analysis of the Kardar-Parisi-Zhang equation.

IS9 Functional Time Ser.

Trends in Stratospheric Ozone Profiles Using Functional Mixed Models

AH YEON PARK*, SERGE GUILLAS*, IRINA PETROPAVLOVSKIKH†,‡

*University College London, UK,
†University of Colorado, Boulder, Colorado, USA,
‡NOAA/ESRL, Boulder, Colorado, USA

168:SergeGuillas.tex,session:IS9

This paper is devoted to the modeling of altitude-dependent patterns of ozone variations over time. Umkher ozone profiles (quarter of Umkehr layer) from 1978 to 2011 are investigated at two locations: Boulder (USA) and Arosa (Switzerland). The study consists of two statistical stages. First we approximate ozone profiles employing an appropriate basis. To capture primary modes of ozone variations without losing essential information, a functional principal component analysis is performed as it penalizes roughness of the function and smooths excessive variations in the shape of the ozone profiles. As a result, data driven basis functions are obtained. Secondly we estimate the effects of covariates - month, year (trend), quasi biennial oscillation (QBO), solar cycle, arctic oscillation (AO) and the El Niño/Southern Oscillation (ENSO) cycle - on the principal component scores of ozone profiles over time using Generalized Additive Models (GAMs). The effects are smooth functions of the covariates, and are represented by knot-based regression cubic splines. Finally we employ Generalized Additive Mixed Effects Models (GAMMs) incorporating a more complex error structure that reflects the observed seasonality in the data. The analysis provides more accurate estimates of influences and trends, together with enhanced uncertainty quantification. We are able to capture fine variations in the time evolution of the profiles such as the semi-annual oscillation. We conclude by showing the trends by altitude over Boulder. The strongly declining trends over 2003-2011 for altitudes of 32-64 hPa show that stratospheric ozone is not yet fully recovering.

Acknowledgment. Ah Yeon Park was supported by a doctoral fellowship from the Kwanjeong Educational Foundation. S. Guillas was partially supported by a Leverhulme Trust research fellowship on "stratospheric ozone and climate change" (RF/9/RFG/2010/0427). We also thank the "NOAA Atmospheric Composition and Climate program" as the source of funding for this paper.

A Moment Matching Market Implied Calibration for Option Pricing Models

CS25A Stoch. Finance I.

FLORENCE GUILLAUME[*,†], WIM SCHOUTENS[*]

[*]KU Leuven, Leuven, Belgium
[†]email: florence.guillaume@wis.kuleuven.be 169:FlorenceGuillaume.tex,session:CS25A

The calibration of a model on a given market situation is a critical and important part of any derivative pricing and risk management exercise. Traditionally, one solves the so-called inverse problem, which consists of finding the parameter set that is compatible with the observed market price of a set of liquidly traded derivatives. Typically, a perfect match is not plausible and one looks for an "optimal match". More precisely, one minimizes the distance between the model and the market prices of benchmark instruments by using a search algorithm. Most commonly, practitioners are minimizing the root mean square error between model and market vanilla prices or between model and market implied volatilities. Although the root mean square objective function is the current industry practice, there exist other alternatives just as suitable. The optimal parameter set typically strongly depends on the choice of the objective function, leading to significantly different prices for the more exotic and structured products. Besides the calibration risk issue, it is well known and documented that several additional problems can arise with the standard calibration methodology. Firstly, one faces the problem of selecting an appropriate starting value for the search algorithm used. Indeed, the objective function to be minimized is typically a non-convex function of the model parameter set and can thus have several local minima, making the solution of the standard calibration problem dependent on the initial parameter set, which is taken as starting value of the optimization algorithm and on the sophistication of the numerical search performed. Further, one has the typical related problem of finding a local minimum instead of the global minimum. Also, a calibration exercise can be quite time consuming, especially if the number of parameters to be calibrated is becoming large. Hence we provide a new calibration formalism which consists of matching the moments of the asset log-return process with those inferred from liquid market data. In particular, we derive a model independent formula for the moments of the asset log-return distribution function by expanding power returns as a weighted sum of vanilla option payoffs. The new calibration methodology rests on closed-form formulae only: it is shown that, for a model with N parameters, the moment matching calibration problem reduces to a system of N algebraic equations which give directly the optimal parameter set in terms of the market implied standardized moments of order 2 to order N and avoids thus the delicate choice of a particular objective function. The new calibration formalism provides thus an appealing alternative to the standard calibration problem since it does not depend on starting value for the model parameters, it is almost instantaneously delivering the matching parameters and it avoids local minima problems. For the numerical study, we work out different popular exponential Lévy models and illustrate how the new methodology outperforms the current market standard ones in terms of both the computation time and the quality of the fit.

Acknowledgment. Florence Guillaume is a postdoctoral fellow of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.).

POSTER Poster

Prediction Intervals for Linear Regression on Log-Normal Exposure Data

SARA GUSTAVSSON[*,†], EVA M. ANDERSSON[*]

[*]Department of Occupational and Environmental Medicine, Sahlgrenska University Hospital and Academy at Gothenburg University, Sweden
[†]email: sara.gustavsson@amm.gu.se 170:SaraGustavsson.tex,session:POSTER

We propose a maximum likelihood based method for estimating a linear regression for lognormal data. Much data in environmental science are approximatly log-normally distributed and in many situations it is the absolute effect of a predictor that is of interest (i.e. a linear model). We consider the following model for the response Y:

$$\ln(Y) = \ln\left(\mu_{Y|X} - \sigma_Z^2\right) + \epsilon,$$

where ϵ is iid $N(0; \sigma_Z^2)$. The expected value is a linear function of the predictors $X_1, X_2, \ldots, X_p$:

$$\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

We use a method in which the likelihood function from the log-normal distribution is used to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ_Z. These estimates are then used to construct the prediction intervals.

Prediction intervals can be used for establishing reference ranges for subpopulations, e.g. establishing a continuous age-dependent reference range. For the model above, Taylor expansion is used to estimate a $(1-\alpha)\cdot 100\%$ prediction interval as:

$$\exp\left(\mu_{Y|X} - \frac{S_Z^2}{2} \pm t_{\alpha/2,(n-p-1)} \cdot \sqrt{se\left(\hat{\mu}_{Y|X}\right) \cdot \frac{1}{\hat{\mu}_{Y|X}^2} + \frac{S_Z^4}{2(n-p-1)} + S_Z^2}\right).$$

This prediction interval was evaluated in a simulation study (with $\beta = [1.564, 0.122, 0.075]$, $\sigma_Z = 0.4$, $n = 600$ and $3 \cdot 10^6$ iterations), which showed the 95% prediction interval to have a coverage (proportion of observations covered by the corresponding interval) at about 0.95 (mean 0.951, sd= 0.006). The study also showed the interval to be symmetric with approximately 2.5% of the observation lower than the lower limit and 2.5% higher than the upper limit.

The proposed interval was applied to a linear model for CRP (inflammation marker) with insulin resistance (HOMA-IR) and waist circumference as predictors (data on 598 Swedish women, 64 years old). The proportion of observations covered by the prediction interval was 0.94 with 3% of the observations lower than the lower limit of the interval and 3% higher than the upper limit.

Our proposed method gave satisfactory results, both in the application and in the simulation study.

StPburgS St. Petersburg Mem. Sess.

Revisiting the St. Petersburg Paradox

ALLAN GUT[*,†]

[*]Uppsala University, Uppsala, Sweden
[†]email: allan.gut@math.uu.se 171:AllanGut.tex,session:StPburgS

It all began about 10 years ago a with some interesting observations that inspired a generalization of Feller's weak law for i.i.d. random variables to sums obeying more general normalizing sequences,

and an application to the St. Petersburg game. This was followed by an extension to unfair coins and later to the situation in which I toss a biased coin for which $P(\text{head}) = p$, $0 < p < 1$, repeatedly until head appears. If this happens at trial number k you receive sr^{k-1} Euro, where s, $r > 0$, which induces the random variable

$$P(X = sr^{k-1}) = pq^{k-1}, \quad k = 1, 2, \dots.$$

The particular case when $s = 1/p$ and $r = 1/q$ has been studied before. The special case $p = q = 1/2$ reduces the game to the classical one (of course).

We prove a weak law, and extend Martin-Löf's theorem concerning convergence in distribution along the geometric subsequence 2^n to an infinitely divisible, semistable law.

Two modifications are also considered. First, a trunctated game and the problem "How many games until game over?", and, then, a game in which the player can borrow money without limit for the stakes, but has to pay interest on the capital. Both models have earlier been studied by Martin-Löf for the classical game.

Most proofs run along the main lines of those for the classical case.

The more recent parts of this work is joint with Anders Martin-Löf.

Pointing in New Directions

IS20 Space-Time Stat.

PETER GUTTORP[*,‡,§], AILA SÄRKKÄ[†], THORDIS L. THORARINSDOTTIR[*]

[*]Norwegian Computing Center, Oslo, Norway,
[†]Chalmers University of Technology, Gothenburg, Sweden,
[‡]University of Washington, Seattle, Washington, U.S.A.
[§]email: peter.guttorp@nr.no 172:PeterGuttorp.tex,session:IS20

Connecting old-fashioned point process models with new-fangled methodology, we use a neurology data set on nerve endings as a laboratory for model assessment in general clustered point processes. Trying to determine what kind of distinctions a given set of data sets allows is important in choosing model components. We also look at a spatiotemporal data set, where stationarity is far from reasonable.

Simulation of the Multivariate Generalized Pareto Distribution of Logistic Type

CS26A Extremes

JÁNOS GYARMATI-SZABÓ[*,†], LEONID V. BOGACHEV[*]

[*]University of Leeds, Leeds, UK
[†]email: j.gyarmati.szabo@gmail.com 173:JanosGyarmati-Szabo.tex,session:CS26A

In recent years there has been growing evidence that the peaks-over-threshold approach to multivariate extreme value (MEV) modelling using the multivariate generalized Pareto distributions (MGPD) has certain advantages as compared to the classical block-maxima method based on MEV distributions (see, e.g., [2] and further references therein), thus providing an improved accuracy of estimation and inference. In the present work, a new direct method is developed for simulation of the MGPD of logistic type (both symmetric and asymmetric), and its performance is assessed numerically. Our approach is valid for arbitrary dimension and, in contrast with previous works (see, e.g., [1]), in the entire domain of the distribution support.

Assuming the unit Frechét marginals of the background (p-variate) MEV distribution (after a suitable coordinate transformation $\tilde{\mathbf{x}} = \big(\mathbf{1} + \boldsymbol{\xi}(\mathbf{x} - \boldsymbol{\mu})/\boldsymbol{\sigma}\big)^{\mathbf{1}/\boldsymbol{\xi}}$, with componentwise operations on vectors), the symmetric logistic MGPD function is given by (cf. [2])

$$H(\mathbf{x}) = \frac{m_*(\tilde{\mathbf{x}} \wedge \tilde{\mathbf{x}}^0) - m_*(\tilde{\mathbf{x}})}{m_*(\tilde{\mathbf{x}}^0)}, \qquad \mathbf{x} \in \mathbb{R}^p, \tag{1}$$

where $\wedge$ denotes the minimum operator, $\tilde{\mathbf{x}}^0 = \tilde{\mathbf{x}}|_{\mathbf{x}=0}$ and $m_*(\tilde{\mathbf{x}}) = \big(\tilde{x}_1^{-1/\alpha} + \cdots + \tilde{x}_p^{-1/\alpha}\big)^\alpha$ $(0 < \alpha \le 1)$. In the case $\alpha = 1$, corresponding to independent components $\tilde{X}_1, \ldots, \tilde{X}_p$ of the (transformed) sample vector, the logistic MGPD (1) is reduced to a convex combination of quasi-univariate distributions (cf. [2]). If $\alpha < 1$ then we show that the distribution function $H(\mathbf{x})$ is absolutely continuous, even though the definition (1) involves a cut-off. Furthermore, conditioned on the "radial" part $Z = Y_1 + \cdots + Y_p$, $Y_i = \tilde{X}_i^{-1/\alpha}$ (which has an explicit polynomial density $f_Z(z)$), the "angular" vector $(Y_1, \ldots, Y_p)$ is uniformly distributed on the simplex $\Xi_{p-1}(z) \subset \mathbb{R}^p$ with lateral side z or, when $z > z_0 = (\tilde{x}_1^0)^{-1/\alpha} + \cdots + (\tilde{x}_p^0)^{-1/\alpha}$, on the perforated simplex $\Xi_{p-1}(z) \setminus ((\tilde{\mathbf{x}}^0)^{-1/\alpha} + \Xi_{p-1}(z - z_0))$.

Based on these results, one can first simulate the radial part $Z = z$ using either the quantile inversion (which is possible for $z \le z_0$) or a suitable version of the conventional acceptance-rejection method. Then, given the value $Z = z$, the simulation of the angular part $(Y_1, \ldots, Y_p)$ can be carried out via the (multivariate) acceptance-rejection method, which however may be rather inefficient for $\alpha \approx 1$. As an alternative, we propose a dynamic Monte Carlo procedure based on the truncated Dirichlet distribution. Finally, simulation of the asymmetric MGPD can be handled via suitable conditioning leading to symmetric MGPDs on lower-dimensional subspaces.

Acknowledgment. J. Gyarmati-Szabó was supported by an EPSRC Doctoral Training Grant and a Mathematics Postgraduate Research Scholarship at the University of Leeds. L. V. Bogachev was partially supported by a Leverhulme Research Fellowship.

References

[1] Michel, R., 2007: Simulation of certain multivariate generalized Pareto distributions, *Extremes*, **10**, 83–107.

[2] Rootzén, H. and Tajvidi, N., 2006: Multivariate generalized Pareto distributions, *Bernoulli*, **12**, 917–930.

IS18 Rough Paths

Extracting Information from the Signature of Paths

LAJOS GERGELY GYURKÓ[*,†,§], TERRY LYONS[*,†], MARK KONTKOWSKI[‡], JONATHAN FIELD[‡]

[*]Oxford-Man Institue of Quantitative Finance, Oxford, UK,
[†]Mathematical Institute, University of Oxford, Oxford, UK,
[‡]Man Group plc., London, UK
[§]email: gyurko@maths.ox.ac.uk

174:Gyurko.tex,session:IS18

The paper aims to explore how multiple features observed in real market data can be characterised quantitatively via a property of sample paths known as the *signature* of the path.

At an abstract mathematical level, the notion of a signature as an informative transform of a multidimensional time series was established by Ben Hambly and Terry Lyons [1], meanwhile Ni Hao et al [2] have introduced the possibility of its use to understand financial data and pointed to the power this approach has for machine learning and prediction.

We evaluate and refine these theoretical suggestions against some real world data and practical examples. Moreover, the paper identifies a low dimensional part of the signature that preserves essential information for understanding and measuring market impact and other properties of trade execution algorithms.

The first part of the paper presents a few motivating examples which introduce the signature of time series and demonstrate what information it preserves with special attention on properties which are not captured by the traditional statistical indicators.

In the second part, we work with the signature of level one limit order book data to caracterize 'moods' of the market. Moreover, we use particular instances of various trade execution algorithms – provided by Man Group plc. – in order to characterise their market impact as well as the market environments that are favourable and/or unfavourable for each of these algorithms. We mainly focus on trade execution on futures markets.

References

[1] Ben Hambly, Terry Lyons, *"Uniqueness for the signature of a path of bounded variation and the reduced path group"*, Annals of Mathematics, Pages 109–167 from Volume 171 (2010), Issue 1

[2] Ni Hao, Daniel Levin and Terry Lyons – private communication, 2013

Demimartingale Inequalities and Related Asymptotic Results

OCS23 Strong Limit Thm.

TASOS C. CHRISTOFIDES*, MILTO HADJIKYRIAKOU†,‡

*University of Cyprus, Nicosia, Cyprus,
†Cyprus University of Technology, Limassol, Cyprus
‡email: miltwh@gmail.com 175:MiltoHadjikyriakou.tex,session:OCS23

In recent years concepts of dependence have been the focus of substantial research activity. A variety of these concepts can be found in the literature including positive and negative dependence, as well as extensions and generalizations. In particular, Newman and Wright (1982) introduced the concept of a demimartingale and a demisubmartingale as a generalization of martingales and submartingales. The definition serves the purpose of studying the behavior of the partial sum of mean zero associated random variables. The class of N-demimartingales studied later by Christofides (2003), generalizes the concept of negative association and includes as special case the class of martingales equipped with the natural choice of σ-algebras. The aim of our work is to present maximal inequalities for these new classes of random variables. These inequalities are instrumental in obtaining asymptotic results. The asymptotic results derived for demimartingales can be applied to the case of partial sums of positively associated random variables while the corresponding results concerning N-demimartingales can be applied to partial sums of negatively associated random variables and other statistical functions involving negatively associated random variables. Furthermore, we deal with the classes of $\mathcal{F}$-demimartingales and conditional N-demimartingales and for these new classes of random objects we obtain several maximal inequalities and asymptotic results. For a more detailed study of these large classes of random variables one can study the recent monograph of Prakasa Rao (2011).

References

[Christofides (2003)] Christofides, T.C., 2003: Maximal inequalities for N-demimartingales, *Arch. Inequal. Appl.*, **50**, 397 - 408.

[Newman and Wright (1982)] Newman, C.M., Wright, A.L., 1982: Associated random variables and martingale inequalities, *Z. Wahrsch. Verw. Geb.*, **59**, 361 - 371.

[Prakasa Rao (2011)] Prakasa Rao, B.L.S., 2011: Associated sequences, Demimartingales and Nonparametric Inference. Springer, Basel.

CS14A Stat. Neuronal Data

Non-parametric modeling of multivariate neuron spike times

NIELS RICHARD HANSEN*,†

*Department of Mathematical Sciences, University of Copenhagen, Denmark

†email: Niels.R.Hansen@math.ku.dk 176:NielsHansen.tex,session:CS14A

The observed spike activity for a collection of neurons can be modeled as a multivariate point process. Each coordinate is a process of spike times for a single neuron, whose intensity depends on the spike history of the entire process. We propose non-parametric statistical methods for estimation of a class of filter-based models, where the intensity is given in terms of filters of the spike times for all the neuron spike time processes.

To be specific we consider models where the intensity for the k'th neuron is given as

$$\lambda_k(t) = \phi\left(\beta_0^k + \sum_m \int_0^{t-} h^{km}(t-s)\mathrm{d}N_s^m\right)$$

with N^m denoting the counting process for the m'th neuron. The functions h^{km} are estimated non-parametrically using combinations of penalized maximum-likelihood estimation and basis expansions.

We consider, in particular, an estimation algorithm for h^{km} using gradients of the log-likelihood in a reproducing kernel Hilbert space. Combining this algorithm with the penalization

$$\lambda \sum_{km} ||h^{km}||$$

we achieve the selection property, that is, some h^{km} can be estimated to be identically 0. The local independence graph, which can be related to the connectivity of the network of neurons, is then directly given. Furthermore, we consider applications of the sparse group lasso algorithm using basis expansions of h^{km} in terms of basis functions with local support. This makes it possible to estimate the support of h^{km}, and in this way we learn about the delay and memory in neural networks.

The generel methods for multivariate point process modeling are implemented in the R package `ppstat` available from CRAN.

CS32A Nonparame

IMSPE Nearly-Optimal Experimental Designs for Gaussian Processes via Spectral Decomposition

OFIR HARARI*,†, DAVID M. STEINBERG*

*Tel Aviv University, Israel

†email: ofirhara@post.tau.ac.il 177:OfirHarari.tex,session:CS32A

Gaussian Processes are used to model data in numerous scientific and engineering fields and applications, including geostatistics, machine learning, electronics and many others. In particular, it has gained popularity in recent years in the field of computer experiments ([Sacks et al. (1989)]), where it is often used as a surrogate model for an otherwise rather complex mathematical model, coded into computer programs whose run time is typically long. As the computer runtime increases, strategic choice of the set of inputs becomes a necessity. Suitable sampling methods include model-independent designs like Latin hypercube sampling of different types or criterion-based designs which relate directly to the underlying model, such as the maximum entropy design.

Minimum integrated mean squared prediction error designs in the setting of computer experiments were first introduced by [Sacks et al. (1992)], and to date numerical integration has been the method of choice for deriving these designs. In this talk, we study the nature of the criterion, using Mercer's Theorem and the Karhunen-Loève decomposition of random processes to derive an alternative, approximate criterion to the integrated mean squared prediction error.

We proceed from theory to practice by working an example with an anisotropic, two-dimensional process, and generate several minimum integrated mean squared prediction error designs. We also provide an asymptotic bound for the relative approximation error, which is based on the energy conserved when truncating the Karhunen-Loève expansion. In addition, we reveal the inevitable relation to Bayesian linear regression, and then proceed to take advantage of the structured nature of our criterion to suggest a sequential sampling scheme in batches (often referred to as "adaptive design"), and finally we extend our framework to the slightly more complicated case of an unknown model intercept, where we investigate the optimality (or lack thereof) of the designs derived earlier.

References

[Sacks et al. (1989)] Sacks, Jerome, Welch, William J., Mitchell, Toby J. and Wynn, Henry P.: Design and Analysis of Computer Experiments, *Statistical Science*, **4**, 409 - 423.

[Sacks et al. (1992)] Sacks, Jerome, Schiller, Susannah B. and Welch, William J.: Designs for Computer Experiments, *Technometrics*, **31**, 41 - 47.

Bootstrap Confidence Regions of Functional Regression Coefficient Estimators

OCS26 Resampling Nonstat T.S.

MADAN G. KUNDU*, JAROSLAW HAREZLAK†,*,§, JACEK LEŚKOW‡

*Indiana University School of Medicine, Indianapolis, Indiana, USA,
†Indiana University Fairbanks School of Public Health, Indianapolis, Indiana, USA,
‡Politechnika Krakowska, Kraków, Poland
§email: harezlak@iupui.edu

178:Harezlak.tex,session:OCS26

Regression modeling in the functional data analysis (FDA) area plays a prominent role. Often we are interested in the estimation of the functional regression coefficients when either the outcome or the predictors are continuous functions. A number of methods have been developed over the past 20 years to address this topic. However, the inference in the FDA settings has been much less developed.

We study the estimation of the functional regression coefficient $\beta(t)$ in the model $E[y_i] = \alpha + \int X_i(t)\beta(t)dt$, where $X_i(t)$'s are the functional predictors and y_i's are the scalar outcomes. The estimation of the functional regression coefficient $\beta(t)$ is usually ill-posed, since the number of sampling points on the curve $X_i(t)$ is often larger than the number of observations y_i. In such cases, a regularization (penalization) methods are often used to obtain the regression function estimates. We use a recently developed approach of [Randolph et al. (2012)] to obtain the $\widehat{\beta(t)}$. In this presentation, we concentrate on the confidence band construction for the functional regression coefficient estimator $\widehat{\beta(t)}$. Most published methods provide a confidence band that is based on the pointwise confidence intervals at each location t with appropriate adjustment for the confidence interval multiplicity. In contrast, our approach gives the confidence limits on the tuning parameter λ controling the model's complexity level. These unidimensional λ confidence limits can be directly translated into the equivalent number of parameters quantifying the smoothness or wiggliness of $\widehat{\beta(t)}$.

We compare model-based confidence band based on the normality assumption for the random errors with several bootstrap-based confidence bands via a Monte Carlo simulation study. In particular, using the equivalence between the functional regression models and linear mixed models,

we consider four resampling proposals: (1) fully-parametric bootstrap, (2) model-based bootstrap, (3) mixed bootstrap, and (4) residual bootstrap. In each of the above mentioned approaches, the resampling is performed with fewer parametric assumptions on the distribution of the random effects and the measurement errors. Major advantage of our method stems from the fact that we directly quantify the variability of the estimated curve in terms of its complexity.

Finally, we apply the developed method to study the association of magnetic resonance spectroscopy metabolite spectra and neurocognitive impairment of the HIV infected patients. The data used in the application arises from an HIV neuroimaging consortium longitudinal study.

Acknowledgment. This research was partially supported by the grant U01MH083545 from the USA National Institutes of Health (NIH).

References

[Randolph et al. (2012)] Randolph, T.W., Harezlak, J., Feng, Z. Structured penalties for functional linear models – partially empirical eigenvectors for regression. *Electronic Journal of Statistics*, **6**, 323–353.

CS8B Bayesian Nonpar.

Reversible Jump Markov Chain Monte Carlo (RJMCMC) vs. Bayes Factor for Model Selection

KELSEY VITENSE*, ANEESH HARIHARAN*,†

*Quantitative Ecology and Resource Management, University of Washington, Seattle, WA, USA

†email: vitense@uw.edu, aneesh@amath.washington.edu

179:Aneesh_Hariharan.tex,session:CS8B

The question of model selection has been addressed extensively in statistical literature. Many approaches have been proposed over the years for dealing with this key issue. In this work, we attempt to compare two Bayesian methods of model selection, Bayes factors and Reversible Jump Markov Chain Monte Carlo (RJMCMC), for the logistic and generalized logistic models, and we address which model selection method is "better".

RJMCMC is a technique for situations where models may have different numbers of parameters. RJMCMC randomly walks around the parameter space of possible model structures, updating model parameters at each step. The algorithm produces a list of the model structures visited at each step and the corresponding set of parameters. RJMCMC is a quite promising approach for model selection in which the entire parameter space is traversed to yield optimal estimates for parameter values.

In Bayes factors, model selection is based on the ratio of the marginal likelihoods of the two candidate models. No optimization is necessary with Bayes factors, and an explicit measure of model complexity is not needed. However, these benefits are subject to the assumption that one can numerically integrate out the parameters, which is more often than not impossible. For this reason, we use Sampling Importance Resampling (SIR) to obtain samples from the posteriors. Using these samples, we compute the marginal likelihoods for the respective models using the Harmonic Mean Estimator.

The two models considered for this work of model selection are the logistic model,

$$y(t) = \frac{y_0 k}{y_0 + (k - y_0)e^{-rt}},$$

and the generalized logistic model,

$$y(t) = \frac{k}{\left[1 + (-1 + (\frac{k}{y_0})^v)e^{-rt}\right]^{1/v}}.$$

Four datasets were simulated for each model from a normal distribution with mean y(t) for t=1,2...100 and standard deviations of .01, .05, .10, and .20. For the logistic model, parameters were set to $y_0 = .001$, $k = 1$, and $r = .15$. For the generalized logistic model, parameters were set to $y_0 = .001$, $k = 1$, $r = .25$, and $v = 2$. Improper, non-informative priors were used for the parameter values.

Both RJMCMC and Bayes factors are successful methods for choosing the correct model given a range of noise in the data. The methods produce similar parameter estimates for cleaner data, but v and r estimates substantially differ for data with more variation. The cause for these differences warrants further exploration. Regardless, SIR makes up for its long run time (∼60x that of RJMCMC) with the relative simplicity of its implementation, while RJMCMC makes up for the complexity of its algorithm with its speed.

Acknowledgment. We wish to acknowledge Andre Punt, Professor and Director, School of Aquatic and Fishery Sciences, University of Washington, Seattle for posing this interesting problem.

Drift Estimation in Sparse Sequential Dynamic Imaging with Application to Nanoscale Fluorescence Microscopy

IS24 Single Molecule Exp.

ALEXANDER HARTMANN[*,§], STEPHAN HUCKEMANN[*], JÖRN DANNEMANN[*], ALEXANDER EGNER[†], CLAUDIA GEISLER[†], AXEL MUNK[*,‡]

[*]Institut für Mathematische Stochastik, Georg-August-Universität Göttingen, Germany,
[†]Laser Laboratory, Göttingen, Germany,
[‡]Max-Planck-Institut für biophysikalische Chemie, Göttingen, Germany
[§]email: alexander.hartmann@mathematik.uni-goettingen.de

180:AlexanderHartmann.tex,session:IS24

A major difficulty in time resolved microscopy such as stochastic marker switching nanoscale fluorescence microscopy is caused by the drift of the object of interest or its bottom layer during the observation process due to external sources, e.g. mechanical effects. Therefore, crucial for a sound registration of the sequence of images is the estimation of this drift. To this end, we develop a global parametric model for the temporal drift development and propose an estimation procedure for its parameters based on M-estimation techniques. Since there is no reference image known a priori, the resulting model is semiparametric. We prove consistency and asymptotic normality of the drift estimator as well as the convergence of the estimated image to the true one. The practicability of our method is demonstrated in a simulation study and in an application to PALMIRA fluorescence microscopy.

Acknowledgment. The authors acknowledge financial support from the Deutsche Forschungsgemeinschaft grant SFB 755 and also from DFG HU 1275/2-1 and DFG FOR 916.

OCS10 Dynamic Factor Models

Identification of Background Forces Driving the Fluctuation in the Time Series of an Agricultural Watershed Using Dynamic Factor Analysis

ISTVÁN GÁBOR HATVANI[*,¶], JÓZSEF KOVÁCS[*], LÁSZLÓ MÁRKUS[*], JÁNOS KORPONAI[†], RICHÁRD HOFFMANN[‡], ADRIENNE CLEMENT[§]

[*]Eötvos Loránd University, Budapest,
[†]West Transdanubian Water Authority, Keszthely,
[‡]University of Kaposvár, Kaposvár,
[§]Budapest University of Technology and Economics, Budapest
[¶]email: hatvaniig@gmail.com 181:IstvanGaborHatvani.tex,session:OCS10

Shallow lakes like Lake Balaton are sensitive to environmental changes and anthropogenic effects. To protect it (mainly Keszthely-bay) against elevated nutrient loads brought by the River Zala, a mitigation wetland, the Kis-Balaton Water Protection System (KBWPS) was constructed at the mouth of the River Zala. The aim of the research was to determine exactly which external influences (latent effects) were the dominating ones in determining the long-term scale behavior of the River Zala's processes at its mouth, which is in fact the first sampling site of the KBWPS. In the study 21 water quality response parameters from the River Zala (1978-2006) and six explanatory ones (1980-2006) were examined. At first factor analysis modeling seemed to be a good method to apply, however, after analyzing the mentioned time series structure it became clear that in its conventional form it is not suitable for the purposes described above. Dynamic factor analysis is the proper method to take into account the lagged correlation structure. In case of the study it was concluded, that external changes in the River Zala and the KBWPS' catchment were reflected in the measured (response) river water quality parameters. With the aid of dynamic factor analysis these changes were followed and presumably most of the explanatory parameters driving them were found.

CS12A Hierarchica Bayesian

Using Bayesian Networks to Model Dependencies in Oil Exploration

RAGNAR HAUGE[*,‡], MARITA STIEN[*], MAREN DRANGE-ESPELAND[*], GABRIELE MARTINELLI[†], JO EIDSVIK[†]

[*]Norwegian Computing Center, Oslo, Norway,
[†]Norwegian University of Science and Technology, Trondheim, Norway
[‡]email: Ragnar.Hauge@nr.no 182:RagnarHauge.tex,session:CS12A

A final stage in oil exploration is to decide whether to drill or not at a possible oil reservoir location, called a prospect. The company will have a portfolio of prospects, and for each of these, the probabilities of success and failure, and associated expected revenue and loss are computed. This means that when planning a drilling program, the drilling decision is not only whether to drill a given prospect, but also to decide the order of drilling.

Due to common geological factors, prospects in an area will be correlated. This will have an impact on the optimal sequence, since each prospect also has a value of information relative to the other undrilled prospects. However, this is currently not utilised; correlation between prospects is mainly used after a prospect has been drilled to update success probabilities at remaining prospects. The reason for this is that no good model for prospect dependencies exists, so the correlation is handled in an ad hoc way, based on geological understanding. Doing this in advance becomes too time consuming due to the large number of possibilities.

We show how Bayesian networks can be used to model the geological understanding of an area, thus giving a complete joint model for success probabilities for the prospects in an area. These networks are expert systems, built by geological experts of the area. This is necessary since each area is unique, and each prospect only yields one observation, so there will never be enough data to build the networks. With a Bayesian network, probability updating is fast and simple, so it is easy to evaluate different drilling strategies. The network is built so that each element has a physical interpretation, to simplify both the building and the interpretation of the network. We have successfully used this approach on several real world cases, and present a case based on one of these.

Joint Modeling of Survival Data and Mismeasured Longitudinal Data using the Proportional Odds Model

OCS5
Anal
Complex
Data

JUAN XIONG*, WENQING HE*, GRACE YI†

*University of Western Ontario, London, Ontario, Canada,
†University of Waterloo, Waterloo, Ontario, Canada
‡email: whe@stats.uwo.ca 183:WenqingHe.tex,session:OCS5

Joint modeling of longitudinal and survival data has been studied extensively, where the Cox proportional hazards model has frequently been used to incorporate the relationship between survival time and covariates. Although the proportional odds model is an attractive alternative to the Cox proportional hazards model by featuring the dependence of survival times on covariates via cumulative covariate effects, this model is rarely discussed in the joint modeling context. To fill this gap, we investigate joint modeling of the survival data and longitudinal data which subject to measurement error. We describe a model parameter estimation method based on expectation maximization algorithm. In addition, we assess the impact of naive analyses that fail to address error occurring in longitudinal measurements. The performance of the proposed method is evaluated through simulation studies and a real data analysis.

On a Method for Estimation of Prediction Error Covariance in Very High Dimensional Systems

CS5D
H-D
Inference

HONG SON HOANG*, RÉMY BARAILLE*

*SHOM/HOM Toulouse France

184:HongSonHoang.tex,session:CS5D

Finding an efficient low cost method for well estimating a prediction error covariance (PE-Cov) in very high dimensional systems is a great and exciting open problem, especially in the field of data assimilation in meteorology and oceanography. It is well known that filtering algorithm consitutes a key tool to offer improvement of system forecast in engineering systems. Theoretically, the optimal filtering algorithm like a Kalman filter (KF) is designed to provide the best estimate for the system state based on all available observations. As many engineering problems are expressed mathematically by means of a set of partial differential equations together with initial and/or boundary conditions, their numerical solutions result on system state with very high dimension (order of $10^6 - 10^7$). In this context, even under the ideal conditions of validity of the KF, its practical implementation is impossible due to the fact that at each arrival time of observation, we need to estimate the PE-Cov (with $10^{12} - 10^{14}$ unknown elements).
In this paper we propose a new method for estimating the ECM in such situations. This method is based on: (i) the sampling procedure (SP) to generate the most informative patterns for the PE (they

will develop in the directions of the most rapid growth of the prediction error - Dominant Schur vectors); (ii) the hypothesis on separation of vertical and horizontal correlation structures. It will be shown that the proposed method is a low-cost one to access well a PE-Cov which is easily implementable in practice for very high dimensional systems, and the filter designed on the basis of such PE-Cov (called PE-Filter) can attain nearly the same level of performance as that of the KF. Different numerical experiments, with low and very high dimensional systems, will be provided to show comparable performances of the KF and PE-Filter.

NYA
Not Yet Arranged

Flexible Parametric Adjustment Method for Correcting the Impacts of Exposure Detection Limits in Regression

SHAHADUT HOSSAIN*,†

*UAE University, Al Ain, UAE
†email: shossain@uaeu.ac.ae 185:SHAHADUT_HOSSAIN.tex,session:NYA

For unbiased estimation of the parameters in regression models, it is necessary that the explanatory variables (exposures) X are completely observed along with the outcome variable Y. However, in many fields of applications, measurements on some quantitative exposures can not be observed completely. Some of the measurements on these covariates are below some experimentally determined detection limits (DLs). If not accounted for, the regression analysis involving such limited explanatory variables distorts the association estimates. In this talk, I will discuss a flexible parametric Bayesian adjustment method for eliminating the deleterious impacts arising in the estimates of regression parameters of logistic regression model due to exposure detection limit. The theoretical framework of the proposed adjustment method will be discussed first, followed by the presentation of some simulation results to demonstrate the performance of the proposed method.

Acknowledgment. This research was supported by the UAE University FBE summer research grant 2011.

POSTER
Poster

Use of Bayesian approach to design and evaluation of bridging studies

CHIN-FU HSIAO*,†

*Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan Town, Miaoli County, Taiwan, ROC
†email: chinfu@nhri.org.tw 186:ChinFuHsiao.tex,session:POSTER

In 1998, the International Conference on Harmonization (ICH) published a guidance to facilitate the registration of medicines among ICH regions including European Union, the United States of America, and Japan by recommending a framework for evaluating the impact of ethnic factors on a medicine's effect such as its efficacy and safety at a particular dosage and dose regimen (ICH E5, 1998). The purpose of ICH E5 is not only to evaluate the ethnic factor influence on safety, efficacy, dosage and dose regimen, but also more importantly to minimize duplication of clinical data allow extrapolation of foreign clinical data to a new region. Hsiao et al. (2007) have proposed a Bayesian approach to synthesize the data generated by the bridging study and foreign clinical data generated in the original region for assessment of similarity based on superior efficacy of the test product over a placebo control. However, for Hsiao et al. (2007), even if both regions have positive treatment effect, their effect sizes might in fact be different. That is, their approach could not truly assess the similarity between two regions. Therefore, in this article we develop a Bayesian consistency approach for assessment of similarity between a bridging study conducted in a new region and studies conducted in the original region. Methods for sample size determination for the bridging study are also proposed. Numerical examples illustrate applications of the proposed procedures in different scenarios.

Acknowledgment. This research was supported by the National Science Council, Taiwan, grant No.: 101-2118-M-400-001-.

References

[Hsiao et al. (2007)] Hsiao C.F., Hsu Y.Y., Tsou H.H., Liu J.P., 2007: Use of prior information for Bayesian evaluation of bridging studies, *J. Biopharm. Stat.*, **17(1)**, 109-121.

Penalized Estimation and Selection for Random Effect Spatial Temporal Models

CS7B Spatio-Temp. Stat II.

NAN-JUNG HSU*,†

*Institute of Statistics, National Tsing-Hua University, Hsin-Chu, Taiwan
†email: njhsu@stat.nthu.edu.tw 187:Nan-JungHsu.tex,session:CS7B

This talk is concerned about the inference for random effect spatial temporal models. We consider the penalized maximum likelihood estimation together with a forward selection procedure to determine the important fixed and random effects in the model and estimate the parameters simultaneously. Some numerical results will be provided to demonstrate the advantages of the proposed method.

Simultaneous Clustering and Variable Selection in Regression

CS5C H-D Var. Selection

HSIN-CHENG HUANG*,†

*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
†email: hchuang@stat.sinica.edu.tw 188:Hsin-ChengHuang.tex,session:CS5C

This talk is concerned with high-dimensional linear regression. We consider simultaneous grouping pursuit and feature selection over an undirected graph, where each predictor corresponds to one node over the graph and a connecting edge between two nodes indicates possible grouping between two nodes. In other words, the prior knowledge about grouping is expressed in terms of a graph. To address computational challenges in high-dimensional analysis, we introduce an efficient method that is based on the augmented Lagrange multipliers, coordinate decent, and difference convex methods. This permits accurate selection and pursuit over an arbitrary graph, for large-scale problems. On one hand, the issue of selection instability of feature selection is effectively treated through grouping pursuit, where highly correlated predictors associated with the response are included in a model through estimation of the grouping structure. On the other hand, redundant predictors can be removed during the process of variable selection. As a result, higher predictive performance is realized as compared to feature selection and grouping pursuit alone. Some numerical and theoretical results will be provided to demonstrate the effectiveness of the proposed method.

Consistent Estimations in the Accelerated Failure Time Model with Measurement Errors

OCS15 Ecol. and Biomed. Data

YIHHUEI HUANG*,†

*Department of Mathematics, Tamkang University, New Taipei City, Taiwan
†email: yhhuang@mail.tku.edu.tw 189:YIHHUEI_HUANG.tex,session:OCS15

The accelerated failure time (AFT) model is an attractive alternative to the Cox proportional hazard (ph) model. The AFT model is intuitive in interpretation where the covariate has effect on ex-

panding/contracting the life time. Nevertheless, unlike the Cox ph model, very few methods for the AFT model had been developed when covariate is subject to measurement error or when there are replicate measurements. For example, there is no corrected score or the conditional score for the AFT model while they do exist for the Cox ph model. This is due to some technical difficulties inherent in the AFT model. In fact, it seems that there is no consistent functional method up to date when covariate is measured with error. The method proposed here is a novel and consistent estimation. We will investigate how the naive estimation function is biased which inspires our method, and show that a consistent estimation can be achieved by using surrogate in a tricky way. A surrogate duplication algorithm is also provided to facilitate the estimation. The performance of the proposed estimation was evaluated through a simulation study.

CS9A Model Sel, Lin Reg

Selection of Shrinkage Estimators for Prediction out-of-sample

NINA HUBER*,†, HANNES LEEB*

*Department of Statistics, University of Vienna, Austria

†email: n.huber@univie.ac.at 190:NinaHuber.tex,session:CS9A

In a linear regression model with random design, we consider a family of James–Stein-type shrinkage estimators from which we want to select a 'good' estimator for prediction out-of-sample. We focus on the challenging situation where the number of explanatory variables can be of the same order as sample size and where the number of candidate estimators can be much larger than sample size. We show that an estimator's out-of-sample performance can differ dramatically from its in-sample performance (which is studied extensively in the existing literature). The actual performance of the estimator depends on unknown parameters in a complicated fashion. Using an estimate of the out-of-sample predictive performance, we replace the actual performance by an empirical counterpart and select the empirically best estimator. We show that the empirically best estimator is asymptotically as good as the truly best (oracle) estimator, uniformly over a large class of data-generating processes. Moreover, we show that we can estimate the performances of both estimators in a uniformly consistent fashion. Our main results are explicit uniform finite sample performance bounds for Gaussian data. These findings extend results of Leeb (2008, Bernoulli 14(3):661–690) where the underlying estimators are least-squares estimators.

POSTER Poster

Discrete Valued Mixing AR(1) Model with Explanatory Variables

ŠÁRKA HUDECOVÁ*,†

*Charles University in Prague,Czech Republic

†email: hudecova@karlin.mff.cuni.cz 191:Hudecova.tex,session:POSTER

Discrete valued time series occur in many practical applications, usually as counts in consecutive time points or as states of a system in time. If the values of the series are small integers (or even arbitrary categories), the traditional ARMA framework is not suitable for the modelling. In last several decades, number of different model classes has been proposed, see [McKenzie (2003)] or [Kedem and Fokianos (2002)].

Recently, [Biswas and Song (2009)] suggested a class of models based on mixtures of the distributions of lagged observations, referred to as Pegram's ARMA models or mixing ARMA models, see [Pegram (1980)]. In particular, we say that a random sequence $\{Y_t\}$ follows a mixing AR(1) model if

$$Y_t = (Y_{t-1}, \phi) \star (\varepsilon_t, 1-\phi), \tag{1}$$

where $\phi \in (0,1)$ and $\{\varepsilon_t\}$ is a sequence of iid variables with a discrete distribution on non-negative integers. The operator $\star$ means that the distribution of Y_t is a mixture of the distribution of Y_{t-1} and the distribution of ε_t with weights ϕ and $1-\phi$. Conditionally on Y_{t-1}, the distribution of Y_t is a mixture of a Dirac measures at Y_{t-1} and the distribution of ε_t with weights ϕ and $1-\phi$. Some properties of this class of models have been analyzed in [Biswas and Song (2009)]. Namely, the autocorrelation function (ACF) of a mixing ARMA(p,q) model is the same as the ACF of the standard ARMA(p,q) model.

In many practical applications one needs to work with non-stationary series. The series of interest can exhibit seasonality or generally depend on some external variables (deterministic or stochastic). It is therefore important to allow explanatory variables enter the model. In this contribution we study the mixing AR(1) model, whose parameters possible depend on explanatory variables. Conditional least squares estimators of the model parameters are suggested and their properties are investigated. The performance in finite samples is studied via simulations.

Acknowledgment. The presenter is supported by the Czech Science Foundation project "DYME Dynamic Models in Economics" No. P402/12/G097.

References

[Biswas and Song (2009)] Biswas, A. and Song, P., 2009: Discrete-valued ARMA processes, *Stat. Probab. Lett.*, **79**, 1884–1889.

[Kedem and Fokianos (2002)] Kedem, B. and Fokianos, K., 2002: Regression models for time series analysis. Wiley-Interscience New-York.

[McKenzie (2003)] McKenzie, E., 2003: Discrete variate time series. In Stochastic processes: modeling and simulation, *Handbook of Statist.*, **21**, 573–606.

[Pegram (1980)] Pegram, G., 1980: An autoregressive model for multilag markov chains. *J. Appl. Probab.*, **17**, 350–362.

Test of Independence for Functional Data

CS9D Testing Mod. Structure

LAJOS HORVÁTH*, MARIE HUŠKOVÁ†,‡, GREG RICE*

*University of Utah, Salt Lake City, USA,
†Charles University in Prague, Czech Republic
‡email: huskova@karlin.mff.cuni.cz

192:MarieHuskova.tex,session:CS9D

The contribution will concern testing the null hypothesis that series of functional observations are independent and identically distributed against serial dependence. Our procedure is based on the sum of the L2 norms of the empirical correlation functions. The limit distribution of the proposed test statistic is established under the null hypothesis. Under the alternative the sample exhibits serial correlation, and consistency is shown when the sample size as well as the number of lags used in the test statistic tend to ∞.

A Monte Carlo study illustrates the small sample behavior of the test and the procedure is applied to data sets, Eurodollar futures and magnetogram records.

Monitoring Profiles Based on Proportional Odds Models

OCS15 Ecol. and Biomed. Data

LONGCHEEN HUWANG*,†, YI-HENG HUANG*

*National Tsing Hua University, Hsinchu, Taiwan
†email: huwang@stat.nthu.edu.tw

193:Longcheen_Huwang.tex,session:OCS15

In this talk, the quality of a process or product is represented by a relationship (or profile) between the response variable and one or more explanatory variables, which is characterized better

by a proportional odds model. Two control charting schemes for monitoring such profiles in Phase II study are proposed. Simulation studies are conducted to compare the effectiveness of these two charting schemes. A diagnostic method is utilized to find the change point location of the process and to identify the parameters of change in the profile. An example is also used to illustrate the implementation of the proposed charting schemes and diagnostic method.

Acknowledgment. This research was partially supported by the Taiwanese National Science Council, grant No.: 100-2118-M-007-005-MY2.

OCS15 Ecol. and Biomed. Data

Population Loss Estimation by Occupancy Rates

WEN-HAN HWANG[*,†]

[*]Institute of Statistics, National Chung Hsing University, Taichung 40724, Taiwan

[†]email: wenhan@nchu.edu.tw 194:Wen-Han_Hwang.tex,session:OCS15

The assessment of species' extinction is the crux of conservation ecology. Although the population loss rate can be straightforwardly estimated by the population sizes over two distinct time periods, it is difficult to obtain such information in practice. Instead, the occupancy rate is easily available for most situations. This study attempts to establish the relationship/model between the occupancy rate and the population loss rate. Based the relationship, the population loss rate would be estimated and applied to determine the level of a species' extinction. A simulation study by means of two census plant data turns out that the proposed model is promising to be the bridge between the occupancy rate and the population rate. In addition, we show some theoretical properties that can explain phenomena in the previous study of He (2012).

References

[He (2012)] He, F. (2012). Area-based assessment of extinction risk. *Ecology* **93,** 974–980.

CS2A Stat. Genetics

Comparisons of Normalization Methods for Relative Quantization in Real-Time Polymerase Chain Reaction

YI-TING HWANG[*,§], YU-HUEI SU[*], HARN-JING TERNG[†], HSUN-CHIH KUO[‡]

[*]Department of Statistics, National Taipei University, Taipei, Taiwan,

[†]Advpharma, Inc., New Taipei City, Taiwan,

[‡]National Chengchi University, Taipei, Taiwan

[§]email: hwangyt@gm.ntpu.edu.tw 195:Yi-Ting_Hwang.tex,session:CS2A

The real-time PCR (real-time polymerase chain reaction) is a common technique for evaluating the gene expression. This technique can provide very sensitive and accurate results since it is monitored instantaneously and also performs a quantitative analysis for the target gene. It has become a widespread technique in analyzing gene expressions. There are two methods to quantify the real-time PCR gene expression, relative and absolute quantification. Owing to cost and available sources, the relative quantification is the more commonly used method. However, the relative quantification requires a housekeeping gene as an internal control gene to normalize the target gene expression. Andersen et al. (2004) and Dheda et al. (2004) pointed out the gene expression of housekeeping gene may be unstable not only due to the biological variation, but also different experimental conditions. Hence, we discuss the feasibility of implementing the normalization method for high density oligonucleotide array to the relative quantification in real-time PCR.

Three common normalization methods for high density oligonucleotide array, the scaling normalization (Affymetrix, 2002), the invariant set normalization (Li and Wong, 2001) and the quantile normalization (Bolstad et al. 2003), are discussed. Owing to large differences in data characteristics, Monte Carlo simulations are used to evaluate the performance of these normalizations to the real-time PCR. Four indices are used to assess the performance. Furthermore, a real data is used to illustrate the feasibility of these normalizations to the real-time PCR. We find that instead of using the housekeeping gene, the scaling normalization is a good choice for relative quantification in real-time PCR.

On Estimation for the Fractional Ornstein-Uhlembeck Process and the Yuima Package

OCS6 Asympt. for Stoch Proc.

STEFANO MARIA IACUS*,†

*Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Italy

†email: stefano.iacus@unimi.it 196:StefanoMariaIacus.tex,session:OCS6

In this talk we propose consistent and asymptotically Gaussian estimators for the parameters λ, σ and H of the discretely observed fractional Ornstein-Uhlenbeck process solution of the stochastic differential equation $dY_t = -\lambda Y_t dt + \sigma dW_t^H$, where $(W_t^H, t \geq 0)$ is the fractional Brownian motion. For the estimation of the drift λ, the results are obtained only in the case when $\frac{1}{2} < H < \frac{3}{4}$. This paper also provides ready-to-use software for the `R` statistical environment based on the YUIMA package.

On the Uniqueness of MLEs based on Censored Data for Some Life time Distributions

OCS27 Infer. Censored Sample

ILKAY ALTINDAĞ*,†, YUNUS AKDOĞAN*, AHMET ÇALIK*, COŞKUN KUŞ*, ISMAIL KINACI*

*Selcuk University, Konya, Turkey

†email: ialtindag@selcuk.edu.tr 197:IlkayAltindag.tex,session:OCS27

In this paper, graphical method studied by Balakrishnan and Kateri are discussed for several lifetime distributions. Existence and uniqeness of maximum likelihood estimates of Chen, Gompertz and Burr XII distributions under complete, Type-II censored and progressively Type-II censored sample are discussed by graphical method. A numerical analysis is also implemented in order to show the existence, uniqeness and finding of MLE.

Acknowledgment. This research was partially supported by Selcuk University BAP Office.

Dynamic Modeling and Inference for Ecological and Epidemiological Systems

CS13A Epidem. Models

EDWARD L. IONIDES*,†

*University of Michigan, Ann Arbor, Michigan, USA

†email: ionides@umich.edu 198:Edward_Ionides.tex,session:CS13A

Quantitative models for ecological and epidemiological systems plays roles in forecasting and evaluating the potential effects of interventions, as well as building basic understanding of the mechanisms driving the behavior of the system. Characteristic features of these biological systems include

stochasticity, nonlinearity, measurement error, unobserved variables, unknown system parameters, and even unknown system mechanisms. We will consider the resulting methodological challenges, with particular reference to pathogen/host systems (i.e., disease transmission). We will focus on statistical inference methodology which is based on simulations from a numerical model; such methodology is said to have the plug-and-play property. Plug-and-play methodology frees the modeler from an obligation to work with models for which transition probabilities are analytically tractable. A recent advance in plug-and-play likelihood-based inference for general partially observed Markov process models has been provided by the iterated filtering algorithm. We will discuss the theory and practice of iterated filtering.

CS38A Appl. Multivariate Tech.

Path Analysis and Determining the Distribution of Indirect Effects via Simulation

ÖZNUR İŞÇİ*,†, UĞUR KAYALI*, ATILLA GÖKTAŞ*

*Muğla Sıtkı Koçman University, Faculty of Sciences, Department of Statistics, Muğla/TURKEY
†email: oznur.isci@mu.edu.tr 199:Oznur_ISCI.tex,session:CS38A

The difference between Path analysis and the other multivariate analyses is that Path analysis has the ability to compute the indirect effects apart from the direct effects. The aim of this study is to investigate the distribution of indirect effects that is one of the components of path analysis via generated data using the statistical package Minitab 16. To realize this a simulation study has been conducted with four different sample sizes (50, 100, 250 and 500), three different number of explanatory variables (3, 5 and 7) and finally with three different correlation matrices (low, medium and high). A replication of 1000 has been applied for every single combination of sample size, number of variables and type of correltation matrix. The generated data have been obtained writing six different macros within Minitab 16. Using Lisrel 8.0 the path diagrams have been built for each type of the generated data. According to the results obtained from the generated data, it is found that no matter what the sample size is path coefficients tend to be stable. Morover path coefficients are not affected by correlation types either. Since the replication is a 1000 that is fairly great the indirect effects from the path models have been treated as normal and their confidence intervals have been presented as well. It is also found that path analysis should not be used with three explanatory variables. It is because the indirect effects are not statisticaly significant and hence the path models would convert to the ordinary linear regression model. We think that this study would help scientists who are working in social sciences to determine sample size and different number of variables.

OCS29 Stat. Branching Proc.

Measuring Criticality of Time-Varying Branching Processes with Immigration

MÁRTON ISPÁNY*,†

*University of Debrecen, Debrecen, Hungary
†email: ispany.marton@inf.unideb.hu 200:MartonIspany.tex,session:OCS29

Recently, there has been considerable interest in integer-valued time series models for analyzing data sets which consist of counts of events, objects or individuals. Several integer-valued time series models proposed in the literature are based on the branching model.

Let $\{\xi_{k,j}, \varepsilon_k : k, j \in \mathbb{N}\}$ be independent, non-negative, integer-valued random variables such that $\{\xi_{k,j} : j \in \mathbb{N}\}$ are identically distributed for each $k \in \mathbb{N}$. In the talk, we consider the *time-varying*

branching process with immigration (TVBPI) $(X_k)_{k\in\mathbb{Z}_+}$ defined recursively as

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

The sequence $(X_k)_{k\in\mathbb{Z}_+}$ is also called a branching process with immigration in time varying environment. We can interpret X_k as the size of the k^{th} generation of a population, e.g., the number of traffic accidents, hospital admissons, attacks on computer systems, transactions in transaction processing systems. Assume that, for all $k \in \mathbb{N}$, $m_k := \mathsf{E}\xi_{k,1}$, $\lambda_k := \mathsf{E}\varepsilon_k$, $\sigma_k^2 := \mathsf{Var}\xi_{k,1}$, and $b_k^2 := \mathsf{Var}\varepsilon_k$ are finite.

A TVBPI is called asymptotically critical if $m_n \to 1$ as $n \to \infty$. Suppose that there exists $\alpha \in \mathbb{R}$ such that $m_n = 1 + \alpha n^{-1} + \delta_n$, $n \in \mathbb{N}$, with $\sum_n \delta_n < \infty$. We prove the following classification theorem for the asymptotic behavior of the process. The expectation $\mathsf{E}X_n$ of the n^{th} generation is $O(n)$, $O(n \ln n)$, and $O(n^\alpha)$ as $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$, respectively. The following asymptotical results hold in case of vanishing offspring variance for the random step fuctions $\mathcal{X}_t^n := X_{\lfloor nt \rfloor}$ and $\mathcal{M}_t^n := \sum_{k=1}^{\lfloor nt \rfloor} M_k$, $t \in \mathbb{R}_+, n \in \mathbb{N}$, where $M_k := X_k - \mathsf{E}(X_k \mid \mathcal{F}_{k-1})$ is the associated martingale difference and $\mathcal{F}_k$ is the natural σ-algebra.

Theorem 4. *Suppose that* $\alpha < 1$ *and the sequences* $(\lambda_k)_{k\in\mathbb{Z}_+}$, $(\sigma_k^2 \cdot k)_{k\in\mathbb{Z}_+}$, *and* $(b_k^2)_{k\in\mathbb{Z}_+}$ *converge to* λ, σ^2, *and* b^2, *respectively, in Cesaro sense. Then* $n^{-1}\mathcal{X}^n \xrightarrow{\mathcal{L}} \mu_{\mathcal{X}}$, *as* $n \to \infty$, *where* $\mu_{\mathcal{X}}(t) := \lambda t/(1-\alpha)$. *Moreover, under appropriate Lindeberg conditions,* $n^{-1/2}\mathcal{M}^n \xrightarrow{\mathcal{L}} \sigma_{\mathcal{M}} W$ *as* $n \to \infty$, *where* $(W_t)_{t\in\mathbb{R}_+}$ *is a Wiener process and* $\sigma_{\mathcal{M}}^2 := \lambda\sigma^2/(1-\alpha) + b^2$. *Finally, if* $\alpha < 1/2$ *then* $n^{-1/2}(\mathcal{X}^n - \mathsf{E}\mathcal{X}^n) \xrightarrow{\mathcal{L}} \mathcal{X}$, *as* $n \to \infty$, *where* $(\mathcal{X}_t)_{t\in\mathbb{R}_+}$ *is a time-varying Ornstein-Uhlenbeck process defined by the SDE*

$$\mathrm{d}\mathcal{X}_t = \alpha t^{-1}\mathcal{X}_t \mathrm{d}t + \sigma_{\mathcal{M}} \mathrm{d}W_t.$$

The approximate conditional least squares estimator (CLSE) of the parameter α is given by

$$\widehat{\alpha} = \frac{\sum_{k=1}^n (X_k - X_{k-1} - \lambda_k)(X_{k-1}/k)}{\sum_{k=1}^n (X_{k-1}/k)^2}.$$

Then $\widehat{\alpha}$ can be interpreted as the *quantitative measure of the criticality* of an asymptotically critical TVBPI. We prove the asymptotic normality of this estimator. Applications are presented using real datasets.

Acknowledgment. The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

Fitting Mixed Effects Logistic Regression Model for Binomial Data as a Special Case of Generalized Linear Mixed Models (GLMMs)

CS40A Logistic & Multinom. Distr.

NESLIHAN IYIT*, MUSTAFA SEMIZ*

*Department of Statistics, Faculty of Science, Selcuk University, Konya, Turkey

`201:NeslihanIyit.tex,session:CS40A`

A generalized linear model (GLM) is used to assess and quantify the relationship between a response variable and explanatory variables. In GLM, the distribution of the response variable is chosen from the exponential family such as;

$$f(y) = c(y,\lambda)\exp\left(\frac{y\theta - a(\theta)}{\lambda}\right),\ g(\mu) = x'\beta$$

where λ is a scale parameter, $a(\theta)$ is a function that determines the response distribution, $g(\mu)$ is a link function which is linearly related to explanatory variables.

A generalized linear mixed model $(GLMM)$ is a statistical model that extends the class of generalized linear models $(GLMs)$ by incorporating normally distributed random effects. So $GLMM$ is an extension to the GLM in which the linear predictor contains random effects in addition to the fixed effects. In $GLMMs$ random effects are chosen to control for specific factors which are expected to cause within-subject variation that vary among subjects.

In this study, when the response variable is binary, logistic regression model with a fixed and random effect from $GLMM$ approach which is called mixed effects logistic regression model will be examined on a repeated measurements data set especially with including correlated data.

References

[1] Kuss, O. (2002), "How to Use SAS for Logistic Regression with Correlated Data," Proceedings of the 27th Annual SAS Users Group International Conference (SUGI 27), 261-27.

[2] Allison, P.D. (1999), Logistic Regression Using the SAS System:Theory and Application, Cary, NC, SAS Institute Inc.

[3] Beitler, P.J., Landis, J.R. (1985), "A Mixed-effects Model for Categorical Data", Biometrics, 41, 991-1000.

[4] Breslow, N.R., Clayton, D.G. (1993), "Approximate inference in generalized linear mixed models", Journal of the American Statistical Association, 88, 9-25.

[5] Brown, H., Prescott, R. (1999), Applied Mixed Models in Medicine, John Wiley & Sons, Chichester.

[6] Derr, R.E. (2000), "Performing Exact Logistic Regression with the SAS System", Proceedings of the 25th Annual SAS Users Group International Conference (SUGI 25), 254-25.

[7] Diggle, P.J., Liang, K.-Y., Zeger, S.L. (1994), Analysis of Longitudinal Data, Oxford University Press, Oxford.

[8] Liang, K.-Y., Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", Biometrika, 73, 13-22.

[9] Littell, R.C., Milliken G.A., Stroup W.W., Wolfinger, R.D. (1996), SAS System for Mixed Models, Cary, NC, SAS Institute Inc.

[10] McCullagh, P., Nelder J.A. (1989), Generalized Linear Models, Chapman and Hall, New York.

[11] Wolfinger, R.D. (1997), "Comment: Experiences with the SAS Macro NLINMIX," Statistics in Medicine, 16, 1258-1259.

[12] Wolfinger, R.D. (1999), "Fitting Nonlinear Mixed Models with the new NLMIXED Procedure", Proceedings of the 24th Annual SAS Users Group International Conference (SUGI 24), 287-24.

CS34A Clinical Studies

Enhance Efficiency and Ethics of Clinical Trials Via Bayesian Outcome-Adaptive Randomization and Early Stopping

J. JACK LEE*,†

*University of Texas MD Anderson Cancer Center, Department of Biostatistics

†email: jjlee@mdanderson.org 202:JJack_Lee.tex,session:CS34A

Two major goals for clinical trials are to identify effective treatments and to treat patient best in the trial. A good design should be both efficient and ethical. An efficient design requires a small sample size to achieve the pre-specified type I and type II error rate for testing treatment efficacy. An ethical design allocates patients to the best available treatments to increase the overall treatment success in the trial. Outcome-adaptive randomization (AR) has been proposed in clinical trials to assign more patients to better treatments based on the interim data. When sufficient information accumulates during the trial, early stopping for futility or efficacy can reduce the study sample size.

Bayesian framework provides a platform for continuous learning, hence, is ideal for implementing AR and early stopping. Generally speaking, equal randomization (ER) requires a smaller sample size and yields a smaller number of non-responders than AR to achieve the same type I and type II errors. Conversely, AR produces a higher overall response than ER by assigning more patients to the better treatments as the information accumulates in the trial. ER is preferred when the patient population outside the trial is large. AR is preferred when the difference in efficacy between treatments is large or when limited patients are available outside the trial. The equivalence ratio of outside versus inside trial populations can be computed when comparing the two randomization approaches. Furthermore, AR has the 'self-correction' property even if the initial information is not quite accurate. Under the Bayesian framework, continuous monitoring of the efficacy endpoint can be easily achieved by computing the posterior probability or the predictive probability of treatment success. When properly applied, early stopping efficiently reduce the trial sample size without compromising the type I and type II errors. Dynamic graphics and simulations will be presented to evaluate the relative merits of AR versus ER and the impact of early stopping. A biomarker-based Bayesian adaptive design for selecting treatments, biomarkers, and patients for targeted agent development will be illustrated in the BATTLE-1 and BATTLE-2 trials for patients with non-small cell lung cancer.

References

[1] Lee, JJ., Chen, N., Yin, G., 2012: Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clin Cancer Res*, **18(17)**, 4498-4507

[2] Yin, G.,Chen, N., Lee, JJ., 2012: Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Applied Statistics*, **61(2)**, 219-235

[3] Kim, ES., Herbst, RS., Wistuba, II., Lee JJ., Blumenschein GR., Tsao A., Stewart DJ., Hicks ME., Erasmus, J., Gupta, S., Alden, CM., Liu, S., Tang, X., Khuri, FR., Tran, HT., Johnson, BE., Heymach, JV., Mao, L., Fossella, F., Kies, MS., Papadimitrakopoulou, V., Davis, SE., Lippman, SM., Hong, WK., 2011: The BATTLE Trial: Personalizing Therapy for Lung Cancer. *Cancer Discovery*, **1(1)**, 44-53

Functional Convergence of Linear Processes with Heavy Tail Innovations

CS19A Lim. Thms. Heavy Tails

RALUCA BALAN*, ADAM JAKUBOWSKI†,§, SANA LOUHICHI‡

*University of Ottawa, Canada,
†Nicolaus Copernicus University, Toruń, Poland,
‡Laboratoire Jean Kuntzmann, Grenoble, France
§email: adjakubo@mat.umk.pl

203:AdamJakubowski.tex,session:CS19A

A linear process built upon i.i.d. innovations $\{Y_j\}_{j\in\mathbb{Z}}$ is (for us) a sequence $\{X_n\}_{n\in\mathbb{Z}}$ given by

$$X_n = \sum_{j\in\mathbb{Z}} c_{n-j} Y_j,$$

where the numbers $\{c_j\}_{j\in\mathbb{Z}}$ are such that the series defining X_n is convergent.

Linear processes form the simplest class of dependent models which are suitable for computations and exhibit various interesting phenomena such as clustering of big values and long-range dependence.

We are interested in convergence of partial sum processes

$$S_n(t) = \frac{1}{a_n} \sum_{k=1}^{[nt]} X_k.$$

We shall consider only the case when Y_j's are in the domain of strict attraction of a strictly α-stable and non-degenerate distribution on $\mathbb{R}^1$, $\alpha \in (0,2)$, the numbers c_j are summable:

$$\sum_{j\in\mathbb{Z}} |c_j| < +\infty,$$

and the linear processes are non-trivial, i.e. at least two among c_j's are non-zero.

Since the work by Avram and Taqqu (1992) it is known that in this case the convergence in Skorokhod's J_1-topology cannot hold. Avram and Taqqu (1992) and Louhichi and Rio (2011) obtained functional convergence in Skorokhod's M_1-topology for the case when *all* $c_j \geq 0$ (what implies that X_n's are associated).

In this talk we show functional convergence of $S_n(t)$ in so-called S-topology, introduced by Jakubowski (1997). We give some implications of this fact.

We discuss also convergence of finite dimensional distributions and obtain a complete characterization. It leads to new – and tractable – sufficient conditions in case $0 < \alpha < 1$.

References

[Avram and Taqqu (1992)] Avram, F. and Taqqu, M. Weak convergence of sums of moving averages in the α-stable domain of attraction. *Ann. Probab.* **20 (1992)**, 483-503.

[Jakubowski (1997)] Jakubowski, A. A non-Skorohod topology on the Skorohod space. *Electr. J. Probab.* **2 (1997)**, paper no.4, 1-21.

[Louhichi and Rio (2011)] Louhichi, L. and Rio, E. Functional convergence to stable Lévy motions for iterated random Lipschitz mappings. *Electr. J. Probab.* **16 (2011)**, paper 89.

CS1A Shape & Image

Estimating the Volume of Bird Brain Components from Contact Regions on Endoneurocrania

JIŘÍ JANÁČEK*, DANIEL JIRÁK†, MARTIN KUNDRÁT*

*Institute of Physiology ASCR, Praha, Czech Republic,
†Institute for Clinical and Experimental Medicine, Praha, Czech Republic

204:JirJancek.tex,session:CS1A

Volumes of brain components are informative about cognitive and motor abilities of vertebrates, thus estimation of the volumes from internal surfaces of the skulls of extinct species is of interest in evolutional studies.

Brains of extant birds were segmented to main components (telencephalon, diencephalon, midbrain, cerebellum and pons with medulla oblongata) on MR acquired with high resolution. MR data were compared with their corresponding regions on endoneurocrania segmented from images obtained in micro CT.

The first approach is based on the regression of the volumes on the surface areas and the integral curvatures of the regions. The volumes and the surface areas can also be efficiently estimated by 3D virtual probes from the data volumes.

The second approach uses fitting the deformable models of the brain components (atlases) inside the endoneurocrania.

Acknowledgment. This research was supported by the Czech Science Foundation, grant No.: P302/12/1207.

References

[Kundrát (2007)] Kundrát, M., 2007: Avian-like attributes of a virtual brain model of the oviraptorid theropod Conchoraptor gracilis, *Naturwissenschaften*, **94**, 499 - 504.

[Kubínová and Janáček (1998)] Kubínová, L., Janáček, J., 1998: Estimating surface area by the isotropic fakir method from thick slices cut in an arbitrary direction, *Journal of Microscopy*, **191**, 201 - 211.

Sparse Variable Selection in High-Dimensional Data with and without Shrinkage

CS5C H-D Var. Selection

MAARTEN JANSEN*,•

*Université Libre de Bruxelles, Brussels, Belgium
•email: maarten.jansen@ulb.ac.be

205:MaartenJansen.tex,session:CS5C

The presented work [1] constructs information criteria based on penalized least squares and penalized likelihood for use in variable selection from observations in high-dimensional models. The penalties used in Mallows' Cp and Akaike's information criteria provide an unbiased estimator for the loss or distance w.r.t. the unknown true model. This unbiasedness, however, only holds for the evaluation of the quality of a given model. When the criterion is used to optimize over a set of candidate models, the optimization output is affected by the observational errors, resulting in additional bias. This bias can be reduced or even undone by shrinkage estimators, such as in soft-thresholding or lasso. Shrinkage introduces new bias and leads to an overestimation of the number of nonzeros. Therefore, we present a criterion for minimum loss variable selection without shrinkage.

The loss function is typically the prediction error (in case of Mallows' Cp) or the Kullback-Leibler distance (for AIC). The Optimization of a criterion that estimates this loss function relies on the observations only. From there, the optimization cannot distinguish between large errors that present themselves as good candidates for explaining the data on one hand and true underlying parameters on the other hand. As a result, the optimization of the selection criterion has the tendency to select, next to the very significant effects, explanatory variables that are most dominated by the observational errors. The discrepancy between the appearance of those variables as the best candidates and the reality of being the worst candidates, can be described and visualized as a mirror effect. The description of the mirror relies on a oracular variable selection that does not depend on the observational errors.

It is understood that shrinkage tempers the effect of the false positives. As a negative side effect, the optimization routine becomes more tolerant for false positives, leading to a large overestimation of the model size. This overestimation is a key motivation for the use of minimax, Bayesian, false discovery based methods instead of minimum loss approaches.

By exploration of the mirror effect, we demonstrate, however, that careful minimization of the loss, being aware of the mirror effect, results in much sparser models. Although hard thresholding is harder than soft thresholding (mathematically, computationally), our work illustrates that it yields much better minimum loss results.

Shrinkage driven routines for variable selection are typically implemented as convex optimization problems. They are fast and yet a good approximation for selection without shrinkage, which is a problem of combinatorial complexity. One of the conclusions of our work is that this good approximation holds only for fixed values of the smoothing parameter or — equivalently — the model size. Once the smoothing parameter is optimized, the shrinkage should be compensated for in the information criterion.

References

[1] M. Jansen. Information criteria for variable selection under sparsity. *Under revision (Biometrika)*, 2013.

CS19C Lim. Thms. Sums of RVs

The $m(n)$ out of $k(n)$ Bootstrap for Partial Sums of St. Petersburg Type Games

EUSTASIO DEL BARRIO*, ARNOLD JANSSEN†,‡, MARKUS PAULY†

*Universidad de Valladolid, Facultad de Ciencias, C/ Prado de la Magdalena s/n, Spain,
†University of Düsseldorf, Institute of Mathematics, Universitätsstrasse 1, Germany
‡email: janssena@math.uni-duesseldorf.de 206:Janssen.tex,session:CS19C

Consider an arbitrary i.i.d. sequence $(X_i)_{i\in\mathbf{N}}$ of real valued random variables. At finite sample size $k(n)$ the following question is of high interest. What can be said about the distribution of the partial sum

$$S_{k(n)} = \sum_{i=1}^{k(n)} X_i$$

given the values $X_1, \ldots, X_{k(n)}$? In many cases this question can be attacked by Efron's bootstrap or existing bootstrap modifications which are widely used tools in modern statistics. However, as it is well known, Efron's bootstrap may fail for heavy tailed X_1. In this talk we like to point out that also bootstrap modifications like the $m(n)$ out of $k(n)$ bootstrap cannot solve the problem without further assumptions on X_1. Here it is well known that for non-normal but stable limit laws the $m(n)$ out of $k(n)$ bootstrap can be consistent but the resample size $m(n)$ may depend on the index α of stability. However, continuing the work of del Barrio et al. (2009) we prove that in various cases a whole spectrum of different conditional and unconditional limit laws of the $m(n)$ out of $k(n)$ bootstrap can be obtained for different choices of $\frac{m(n)}{k(n)} \to 0$ whenever X_1 does not lie in the domain of attraction of a stable law. As a concrete example we study bootstrap limit laws for the cumulated gain sequence of repeated St. Petersburg games. For these games the investigation of distributional convergent partial sums have e.g. been investigated by Martin-Löf (1985), Csörgő and Dodunekova (1991), Csörgő (2010) and Gut (2010). Here it is shown that the bootstrap inherits these partial limit laws. In particular, a continuum of different semi-stable bootstrap limit laws occur for classical and generalized St. Petersburg games.

References

[1] Csörgő, S., 2010: Probabilistic approach to limit theorems for the St. Petersburg game, *Acta Universitatis Szegediensis. Acta Scientiarum Mathematicarum*, **76**, 233–350.

[2] Csörgő, S., Dodunekova, R., 1991: Limit theorems for the Petersburg game, in sums, trimmed sums and extremes, vol. **23** of *Progr. Probab., Birkhäuser Boston*, Boston, MA, 285–315.

[3] del Barrio, E., Janssen, A., Matrán, C., 2009: On the low intensity bootstrap for triangular arrays of independent identically distributed random variables, *TEST*, **18**, 283–301.

[4] Gut, A, 2010: Limit theorems for a generalized St Petersburg game, *Journal of Applied Probability*, **47**, 752–760, Corrections (2012).

[5] Martin-Löf, A., 1985: A limit theorem which clarifies the "Petersburg paradox", *Journal of Applied Probability*, **22**, 634–643.

Local Polynomial Fits for Locally Stationary Processes

CS4A Time Series II.

RAINER DAHLHAUS*, CARSTEN JENTSCH†,‡

*University of Heidelberg, Germany,
†University of Mannheim, Germany
‡email: cjentsch@mail.uni-mannheim.de 207:CarstenJentsch.tex,session:CS4A

In this paper we consider the general class of local polynomial estimators for parameter curves of locally stationary processes. Under suitable regularity conditions, we derive explicit expressions for the limiting bias and variance of local polynomial estimators of arbitrary order and prove central limit theorems by applying a general result from empirical spectral process theory. By using these expressions, we construct an adaptive estimator with a plug-in method. In addition, we studied the uniform rate of convergence of these estimators. We also apply the results to the estimation of a time-varying frequency, where we compare the performance of different estimation strategies. In a simulation study, we investigate the finite sample properties of the estimators.

Nonparametric Testing Methods for Treatment-Biomarker Interaction based on Local Partial-Likelihood

OCS5 Anal Complex Data

WENYU JIANG*,†

*Queens University, Kingston, Canada
†email: wjiang@mast.queensu.ca 208:WenyuJiang.tex,session:OCS5

In clinical trials, patients with different biomarker features may respond differently to the new treatments or drugs. In personalized medicine, it is important study the interaction between treatment and biomarkers in order to clearly identify patients that benefit from the treatment. With the local partial likelihood function (LPLE) method proposed by Fan et al. (2006), the treatment effect can be modelled as a flexible function of the biomarker. In this paper, we propose a bootstrap test method for survival outcome data based on the LPLE, for assessing whether the treatment effect is a constant among all patients or varies as a function of the biomarker. The test method is called local partial likelihood bootstrap (LPLB) and is developed by bootstraping the martingale residuals. The test statistic measures the amount of changes in treatment effects across the entire range of the biomarker and is derived based on asymptotic theories for martingales. The LPLB method is nonparametric, and is shown in simulations and data analysis examples to be flexible to identify treatment effects of any form in any biomarker defined subsets, and more powerful to detect treatment-biomarker interaction of complex forms than the Cox regression model with a simple interaction.

Statistical methods for detecting electoral anomalies: The example of Venezuela

NYA Not Yet Arranged

RAÚL JIMÉNEZ*,†

*Universidad Carlos III de Madrid, Spain
†email: rauljose.jimenez@uc3m.es 209:RaulJimenez.tex,session:NYA

Starting with the 2004 recall referendum, an important opposition sector to President Chavez has questioned the integrity of the Venezuelan electoral system. After carrying out a forensic analysis on Venezuelan elections, since the rise of Chavez to power until his recent death, we reach two

controversial and paradoxical conclusions: on one hand, we cannot rule out the hypothesis of fraud in elections run by the current electoral referee. On the other hand, if fraud has been committed, this has hardly been determining. Only under extreme hypothetical scenarios, where almost all the unaudited polling stations were altered, the fraud could overturn the results in referendums of 2004 and 2009, but not in the rest of the study cases.

IS12 Machine Learning

Computational and Statistical Tradeoffs via Convex Relaxation

VENKAT CHANDRASEKARAN[*], MICHAEL I. JORDAN[†,‡]

[*]California Institute of Technology, Pasadena, CA, USA,
[†]University of California, Berkeley, CA, USA
[‡]email: jordan@stat.berkeley.edu 210:MichaelJordan.tex,session:IS12

Modern massive datasets create a fundamental problem at the intersection of the computational and statistical sciences: how to provide guarantees on the quality of statistical inference given bounds on computational resources such as time or space. Our approach to this problem is to define a notion of "algorithmic weakening," in which a hierarchy of algorithms is ordered by both computational efficiency and statistical efficiency, allowing the growing strength of the data at scale to be traded off against the need for sophisticated processing. We illustrate this approach in the setting of denoising problems, using convex relaxation as the core inferential tool. Hierarchies of convex relaxations have been widely used in theoretical computer science to yield tractable approximation algorithms to many computationally intractable tasks. In the current paper we show how to endow such hierarchies with a statistical characterization and thereby obtain concrete tradeoffs relating algorithmic runtime to amount of data.

POSTER Poster

Statistical Inference in Ecology: an Example of Interdisciplinary Work and its Advantages

JORGE ARGAEZ-SOSA[*,‡], CELENE ESPADAS-MANRIQUE[†]

[*]Faculty of Mathematics, University of Yucatan, Mexico,
[†]Center for Scientific tific Research Center of Yucatan, Mexico
[‡]email: argasosa@uady.mx 211:JorgeArgaez-Sosa.tex,session:POSTER

In this paper we present an example of interdisciplinary work, where the interaction between a statistician and an ecologist was essential for the generation of new statistical research. In Ecology, the determination of high potential areas of habitat for species based on environmental information and presence sites (geographical positions where species have been detected) is a very important issue for conservation purposes. For many species, the only available data are of recorded or observed presences, in general, obtained from herbaria and museums. So, inference should be made considering presence only data. In the literature, there are several methods that are used for this problem.

A method that has been applied in several studies with sensible results is called Domain, which is easy to understand and its implementation is not very difficult. Nevertheless, it has two main drawbacks. First, Domain is not based on formal statistical tools, and so a statistical inference process is not possible. Secondly, Domain does not produce a measure to assess the precision of the obtained results. In the first part of this work, we show how to provide Domain with a formal statistical framework for the estimation of the main parameter involved in the Domain method by using the empirical distribution and using extreme value theory. In addition to the inference procedure, we

propose a way to evaluate the precision of the result obtained. After this postulation, during the interaction with an Ecologist, this ecologist realized that the estimated parameter in Domain could have a possible interpretation from the ecological point of view: an estimator of the habitat specificity, which is a very important concept in ecology that is related to the determination of the endangered species. This observation allowed for the formulation of a new statistical problem: the postulation of a way to make statistical inference (hypothesis test, confidence interval) for the parameter related to the specificity of a habitat. For illustration, we give an example of the formalization of the Domain method and a first approximation for the task of making inference for the specificity of a habitat, considering a real endemic endangered species of Yucatan, Mexico.

References

[Besag et al. (1991)] Besag, J., York, J., Mollìe, A., 1991: Bayesian image restoration with two applications in spatial statistics (with discussion), *Ann. Inst. Stat. Math.*, **43**, 1 - 59.

The Ecological Footprint of Taylor's Universal Power Law

CS13B Envtl. & Biol. Stat.

BENT JØRGENSEN[*,¶], WAYNE S. KENDAL[†], CLARICE G.B. DEMÉTRIO[‡], RENÉ HOLST[§]

[*]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark,
[†]Division of Radiation Oncology, University of Ottawa, Ontario, Canada,
[‡]Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brazil,
[§]Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark
[¶]email: bentj@stat.sdu.dk 212:BentJorgensen.tex,session:CS13B

Half a century ago, the English entomologist L.R. Taylor published a short paper in *Nature*, where he proposed a simple power law to describe the relationship between the spatial variance of plant and animal populations and their mean abundance. If Y is the population count with mean μ, then *Taylor's power law* says that $\mathrm{Var}(Y) = a\mu^b$, where a and b are positive parameters. This form of variance function is well-known in statistics, but Taylor's paper marked the beginning of a remarkable development in applied science, characterized by both triumph and controversy. The possibility of controversy is easy to spot, since the power law harnesses but two degrees of freedom of the complex population dynamics of animals, with infinite possibilities for alternative explanations. Yet Taylor's triumph was that within his lifetime, the power law was confirmed empirically again and again to such an extent that it earned the name "universal". As early as 1983, Taylor was able to declare that his law had been observed for 444 different species of birds, moths and aphids sampled over Great Britain. Even more remarkable, the following decades witnessed a development where the power law was observed in an ever expanding range of different circumstances, ranging from, say, the number of sexual partners reported by HIV infected individuals to the physical distribution of genes on human chromosome 7. The apparent lack of a definitive theoretical explanation for Taylor's law has led some authors to make statements such as *Taylor's power law is merely an empirical model which lacks definite theoretical background*. Yet already in 1984, M.C.K. Tweedie, an English radiotherapy physicist and statistician, proposed a class of statistical distributions with power variance functions providing a possible explanation for the power law. We present some of the empirical evidence for Taylor's law and consider the Tweedie distributions as a possible explanation for the power law. We also discuss the possible ramifications for spatial modelling of population abundance data.

CS9B
Model Sel,
Info Crit

Parametric or Nonparametric: The FIC Approach

MARTIN JULLUM[*,†], NILS LID HJORT[*]

[*]University of Oslo, Norway

[†]email: martinju@math.uio.no

213:MartinJullum.tex,session:CS9B

A long-living question is whether one should rely on a parametric or nonparametric model when analyzing a certain data set. This is a question that cannot be answered by classical model selection criteria like AIC and BIC, since the nonparametric model has no likelihood. When performing a statistical analysis, there is often a certain population quantity μ, here coined the focus parameter, that is of main interest. This motivates translating the model selection problem to a question of choosing the best estimator for μ. In the spirit of the parametric vs. nonparametric game: Given an i.i.d. data set, do we use the nonparametric sample quantile, or do we rely on a parametric model to estimate a certain quantile q of the underlying distribution? Similarly, for right-censored data, do we use the Kaplan–Meier estimator or an estimator based on either the exponential, Weibull or Gompertz distribution to estimate the probability that an individual survives at least to a certain time point t?

In general terms, the focused information criterion (FIC) is a model selection criterion which compares and ranks the candidate models based on estimated mean squared errors (MSEs) of the different $\widehat{\mu}$s. As a consequence, the ranking depends not only on the data, but also on the focus parameter of interest. By its nature the actual formulae for the FIC depend on various characteristics of the underlying framework of models under consideration. The FIC idea has earlier been developed for several classes of problems, amongst them for nested parametric models in the i.i.d. case and covariate selection for both the generalized linear models and Cox proportional hazard regression situations. Our FIC machinery is motivated by the same techniques, but is widened in terms of the available candidate models. Our approach compares general non-nested parametric models with a nonparametric alternative. The estimates of the MSEs consist on model robust estimates of squared bias and variance stemming from large-sample results for the different $\widehat{\mu}$s. The approach is carried out for the i.i.d. and censored data situations. Furthermore, a more general model selection criterion allowing for several focus parameters to be considered simultaneously is constructed.

A conceptual advantage of our FIC approach is that one does not need to address a model's general purpose fit, e.g. via goodness-of-fit procedures. The technique is also robust in the sense that when a parametric model is biased, the probability of selecting this model will converge to 0 as the sample size increases. Thus, the nonparametric model will be selected with probability converging to 1 when all of the parametric candidate models produce biased $\widehat{\mu}$s. Desirable results are also obtained for the event of a certain parametric model actually being fully correct.

References

[1] Claeskens, G., Hjort, N. L., 2003: The focused information criterion (with discussion), *Journal of the American Statistical Association*, **98**, 900 - 916.

[2] Claeskens, G., Hjort, N. L., 2008: Model selection and model averaging, *Cambridge University Press*.

[3] Jullum, M, 2012: Focused information criteria for selecting among parametric and nonparametric models, *Master's thesis, Department of Mathematics, University of Oslo*.

Matérn-based Nonstationary Cross-covariance Models for Global Processes

IS20 Space-Time Stat.

MIKYOUNG JUN*,†

*Texas A&M University, College Station, USA
†email: mjun@stat.tamu.edu 214:MikyoungJun.tex,session:IS20

Many physical processes such as climate variables and climate model errors on a global scale exhibit complex nonstationary dependence structure, not only in their marginal covariance but their cross covariance. Flexible cross covariance models for processes on a global scale are critical for accurate description of each physical processes as well as their cross-dependence and for improved prediction. We propose various ways for producing cross covariance models, based on Matérn covariance model class, that are suitable for describing prominent nonstationary characteristics of the global processes. In particular, we seek nonstationary version of Matérn covariance models whose smoothness parameters vary over space, coupled with differential operators approach for modeling large scale nonstationarity. We compare their performances to some of existing models in terms of AIC and spatial prediction in two applications problems: joint modeling of surface temperature and precipitation and joint modeling of errors of climate model ensembles.

General Consistency Results of PCA in High Dimension

OCS19 Multivar. funct. data

SUNGKYU JUNG*, JASON FINE†, J. S. MARRON†

*University of Pittsburgh, Pittsburgh, USA.,
†University of North Carolina at Chapel Hill, Chapel Hill, USA.
‡email: sungkyu@pitt.edu 215:SungkyuJung.tex,session:OCS19

Principal component analysis is a widely used method for dimensionality reduction and visualization of multidimensional data. It becomes common in modern data analytic situation that the dimension d of the observation is much larger than the sample size n. This leads to a new domain in asymptotic studies of the estimated principal component analysis, that is, in terms of the limit of d. A unified framework for assessing the consistency of principal component estimates in a wide range of asymptotic settings is provided. In particular, our result works for any ratio of dimension and sample size, $d/n \to c$, $c \in [0, \infty]$. We apply this framework to two different statistical situations. When applied to a factor model, we obtain a unified view on the sufficient condition for the consistency of principal component analysis. Secondly, we propose to use time-varying principal components to model multivariate longitudinal data with an irregular grid. A sufficient condition for the consistency of the estimates is obtained by the proposed tool. Simulation results and a real data analysis are included.

Modelling the Peptide Microarray Data

POSTER Poster

ENE KÄÄRIK*,‡, TATJANA VON ROSEN†

*Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia
†Stockholm University, Stockholm, Sweden
‡email: ene.kaarik@ut.ee 216:EKaarik.tex,session:POSTER

In immunology, the analysis of a peptide-array data is a multi-step process which starts with a well-defined biological question, goes through several statistical issues such as experimental design

and data quality, and finishes with the data analysis and a biological interpretation.

The immunological studies produce raw data that is not only voluminous but is typically highly skewed with artifacts due to technical issues. Hence, an appropriate data transformation, standardization and normalization is required prior to the statistical analysis. The main objective of normalization is to ensure that measured intensities within and across peptide-slides are comparable; although the correction for systematic differences between samples on the same slide or chip, or between slides or chips, which do not represent true biological variation between samples is of utmost importance.

The mixed model approach is nowadays widely used in the analysis of micro-arrays. These models use the information in so called control spots (negative control and positive controls) and the peptide responses. The replicated blocks on each slide/chip provide an important information for the removal of "noise", i.e. variation due to the measurement process or poor slide processing procedure (Nahtman, 2007). We propose a new normalization algorithm based on half-normal distribution. The half-normal distribution is the probability distribution of the absolute value of a random variable.

In the analysis of peptide-arrays, the ratio of the foreground to background signal on the log-scale is used as a measure of the response (as a response index) at each spot. The response index measures the relative strength of the foreground signal compared to background, using a log-scale to represent simple doubling effects. Using the log-ratio estimates instead of the ratio makes the distribution more symmetrical. An important problem in a high-dimensional data analysis is a reduction of dimensionality or detection of relevant sets of variables. In this work the variables selection method proposed by Läuter et al. (2009) is used to detect relevant sets of peptides some of which could be connected to specific immune system activity. Another problem of interest is to characterize the immune profile in vaccinated persons over time. The following statistical tests concerning profiles will be considered: (1) test of parallelity of profiles; (2) test of equality of profiles; (3) test of constant profiles given that they are parallel. The contribution to the statistical methodology will concern the high-dimensionality regarding the number of profiles.

We will illustrate the methodology applied to the peptide microarray data from a tuberculosis research.

Acknowledgment. This research is supported by Estonian Science Foundation grant No 8294.

References

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

[Nahtman et al. (2007)] Nahtman, T., Jernberg, A., Mahdavifar, S., Zerweck, J., Schutkowski, M., Maeurer, M., Reilly, M., 2007: Validation of peptide epitope microarray experiments and extraction of quality data. *Journal of Immunological Methods*, **328**, 1 - 13.

CS12B Bayesian computing

Various Order of Degeneracies of Markov Chain Monte Carlo for Categorical Data

KENGO KAMATANI*,†

*Graduate School of Engineering Science, Osaka University, Japan

†email: kamatani@sigmath.es.osaka-u.ac.jp 217:KengoKamatani.tex,session:CS12B

The Markov chain Monte Carlo (MCMC) method is an efficient tool for approximation of an integral with respect to a certain type of a probability measure. The method was developed among physics researchers but later, it was implemented to the Bayesian statistics. Since then, the MCMC

method has been one of the most popular tool for the evaluation of the complicated integral with respect to the posterior distribution.

Let $P_\theta(dx) = p_\theta(x)dx$ be a probability measure with the prior distribution $p(d\theta)$. The posterior distribution $p(d\theta|x)$ is proportional to $p_\theta(x)p(d\theta)$ under an observation x. Sometimes the posterior distribution does not have a closed form that usually requires some kind of approximation. The Markov chain Monte Carlo procedure results in a Markov chain $\theta(0), \theta(1), \dots$ with invariant distribution $p(d\theta|x)$. Under mild conditions, $p(d\theta|x)$ is well approximated by the empirical distribution of $\theta(0), \theta(1), \dots, \theta(m-1)$. In many applications this strategy has great advantage compared to other numerical integration methods and we experienced the MCMC revolution in the Bayesian statistics in the past two decades.

The validity for this approximation is guaranteed by ergodicity of the Markov chain defined by the MCMC procedure. Although there have been a lot of efforts to analyze convergence property of MCMC, theories for a practical prediction until convergence are still under developing. The propose of the paper is to add a step for the prediction; we provide a tool for the identification of the performance bottleneck with a matter of degree (ex. mild or severe).

It is not easy to perform theoretical comparison of two MCMC strategies for fixed n such as by their operator norms. However it becomes much simpler if we let the sample size $n \to \infty$. With the similar mind to the effective sample size, we propose the order of degeneracy; we measure the length between $\theta(i+1)$ and $\theta(i)$ from one iteration of a MCMC procedure and compare it to those as if they were from i.i.d. sample from the posterior distribution. The order of degeneracy is the fraction of the two lengths. For regular cases an MCMC procedure should have the order of degeneracy $d_n = 1$. On the other hand this value d_n tends to ∞ for non-regular cases. We apply the order of degeneracy to the model for categorical data and show its efficiency such as simple probit model, the normal ogive model and the cumulative probit model. These example shows that despite the name of "non-regular", it is common in practice.

By this properties, it is possible to identify performance bottlenecks of MCMC procedures. The identification is quite important in practice since it can be drastically better by removing such bottlenecks. In particular we discuss the effect of the fragility of parameter identifiability to the MCMC procedures. Through these examples, we show that the order of degeneracy is useful tool for that purpose.

Acknowledgment. This research was partially supported by Grant-in-Aid for Young Scientists (B) 22740055.

A Fusion Secretary Problem: An Optimal Stopping Rule with Changing Priorities of the Observer

NYA Not Yet Arranged

SIDDHARTHAVINAYAKA P. KANE*,†

*Post Graduate teaching Department of Statistics, Rashtrasant Tukadoji maharaj Nagpur University, Nagpur, INDIA

†email: svpkane@yahoo.co.in 218:SiddharthaKane.tex,session:NYA

Abstract: The present paper deals with the modification in the 'Original' Secretary problem. Here the secretary problem is viewed from a different angle where there are two random variables simultaneously considered, one that is X, representing the real rank of the selected unit and the other that is Y, representing the position at which we stop. The major modification here is to consider two characteristics to be observed by the observer say X1 and X2 instead of only X, which forces the observer to modify the selection process. Moreover the observer mixes the other versions in the same task and goes on changing his priorities of selection rules and the quality of the unit to be selected. With these the rule is developed to achieve the optimization.

OCS10 Dynamic Factor Models

Dynamic Factor Analysis of Environmental Systems II: Challenges and Advances in Complex Systems

DAVID KAPLAN*,†, RAFAEL MUÑOZ-CARPENA*, MIGUEL CAMPO-BESCÓS*, JANE SOUTHWORTH*

*University of Florida, Gainesville, USA
†email: dkaplan@ufl.edu

219:DavidKaplan.tex,session:OCS10

In the past eight years we have applied Dynamic Factor Analysis (DFA) to a diverse set of natural and engineered systems to identify groups of environmental time series that exhibit similar behaviors and ask: "why do response variables (RVs) share similar traits?" Using DFA, we are often able to identify explanatory variables (EVs) that help explain this shared variance, allowing us to build Dynamic Factor Models (DFMs) of the system. DFMs are useful for: 1) generating hypotheses about system behavior and thresholds; 2) informing future monitoring and modeling efforts; and 3) exploration of past and future system behavior. We have found DFA to be a powerful and structured exploratory tool, facilitating multidisciplinary collaboration between engineers and physical/social scientists investigating complex systems. We have recently taken an explicitly spatial approach to assess how relationships between RVs and EVs change over time and space, improving our mechanistic understanding of large and complex environmental systems. In studies of groundwater (Kaplan et al., 2010) and soil moisture (Kaplan and Muñoz-Carpena, 2011) in a forested floodplain, we used DFA to identify how geomorphological characteristics (e.g., flood-plain elevation, distance from river channel, distance upstream from river mouth) dictate the relative importance of hydrological and meteorological forcings on the response variables. Building on this approach, in a study of vegetation change in southern Africa (Campo-Bescos et al., 2013), we applied DFA to ten years of monthly MODIS-derived normalized difference vegetation index (NDVI) data and a suite of environmental covariates. Critically, this analysis pointed to a transition in the importance of environmental drivers on NDVI: precipitation and soil moisture were most important in grass-dominated, semi-arid regions, while fire, evapotranspiration, and temperature were more critical in humid, tree-dominated regions. Ongoing work on similar-scale vegetation changes in the Amazon Basin is further incorporating demographic and economic data (in addition to physical EVs) to describe changes in rainforest cover and structure. Large-scale applications in complex systems have required flexibility in both model application and interpretation, and we have identified three considerations for applying DFA in such systems. First, the large volume of remote sensing data available requires that investigators make judicious and informed decisions on spatial or temporal aggregation of data (i.e., from pixels or seconds to watersheds or months). Often, the appropriate scale of analysis is often not know a priori, making it important for researchers to empirically test how the scale of analysis affects model results. Second, the traditional approach to identifying "optimal" models uses a single selection criterion (e.g., Akaike's or Bayesian Information Criteria, AIC/BIC) to ensure parsimonious model fitting. We have found these criteria to be insensitive or incomplete for large datasets and have developed a multi-criteria objective consisting of: a) parsimony (minimizing AIC or BIC); b) global model goodness-of-fit (GOF); c) improvement of GOF for worst-performing RVs; and d) reduction in importance of CTs in the DFM. Finally, while DFA describes the behavior of a set of RVs, the fundamental issue of correlation vs. causation (central to most statistical approaches) remains. Thus in future work we propose to compare previous DFA results with explicit causality tests (e.g., convergent cross mapping) to assess the appropriateness of the DFMs developed. In all cases, we endorse the utility of a multidisciplinary approach to ensure that DFA results are correctly interpreted and applied in complex systems.

References

[Campo-Bescós, et al.(2013)] Campo-Bescós, M.A., R. Muñoz-Carpena, D.A Kaplan, J. Southworth, L. Zhu; P.R. Waylen. 2013. Beyond precipitation: Physiographic thresholds dictate the relative importance of environmental drivers on savanna vegetation. In review in PLoS ONE (March 2013, PONE-D-13-10679).

[Kaplan, et al. (2010)] Kaplan, D., R. Muñoz-Carpena and A. Ritter, A. 2010. Untangling complex shallow groundwater dynamics in the floodplain wetlands of a southeastern U.S. coastal river. Water Resources Research 46, W08528.

[Kaplan and Muñoz-Carpena (2011)] Kaplan, D.A. and Muñoz-Carpena, R. 2011. Complementary effects of surface water and groundwater on soil moisture dynamics in a degraded coastal floodplain forest. J. of Hydrology 398(3-4):221-234.

Statistical Inference on Grouped Censored Data Based on Divergences

CS9C Model Selection

ILIA VONTA*, ALEX KARAGRIGORIOU†,‡

*National Technical University of Athens, Greece,
†University of Cyprus, Nicosia, Cyprus
‡email: alex@ucy.ac.cy

220:Karagrigoriou.tex,session:CS9C

Measures of divergence are used extensively in statistics in various fields. These measures are classified in different categories and measure the quantity of information contained in the data with respect to a parameter θ, the divergence between two populations or functions, the information we get after the execution of an experiment and other important information according to the application they are used for.

Measures of divergence between two probability distributions have a long history initiated by the pioneer work of Pearson, Mahalanobis, Lévy and Kolmogorov. Among the most popular measures of divergence are the Kullback-Leibler measure of divergence (Kullback and Leibler, 1951) and the Csiszar's φ-divergence family of measures (Csiszár, 1963). A unified analysis has been provided by Cressie and Read (1984) who introduced the power divergence family of statistics that depends on a parameter a and is used for goodness-of-fit tests for multinomial distributions.

Measures of divergence can be used in statistical inference for estimating purposes, in the construction of test statistics for tests of fit or in statistical modeling for the construction of model selection criteria like the Kullback-Leibler measure which has been used for the development of various criteria. In Statistics, the problem of determining the appropriate distribution or the appropriate model for a given data set is extremely important for reducing the possibility of erroneous inference. Additional issues are raised in biomedicine and biostatistics. Indeed, the existence of censoring schemes in survival modelling makes the determination of the proper distribution or model an extremely challenging problem. An important aspect is how to check validity of a specific model assumption. In this work we are focusing on divergence measures that are based on a class of measures known as Csiszar's divergence measures. In particular, we propose a class of goodness of fit tests based on Csiszar's class of measures designed for censored survival or reliability data. Further, we derive the asymptotic distribution of the test statistic under simple and composite null hypotheses as well as under contiguous alternative hypotheses. Simulations are furnished and real data are analyzed to show the performance of the proposed tests for different $\varphi-$divergence measures.

References

[Cressie and Read (1984)] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. R. Statist. Soc*, **5**, 440-454.

[Csiszar (1963)] Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten, *Publ. of the Math. Inst. of the Hungarian Academy of Sc.*, **8**, 84-108.

[Kullback and Leibler (1951)] Kullback, S. and Leibler, R. (1951). On information and sufficiency, *Annals of Math. Statist.*, **22**, 79-86.

CS3A Machine learning

Intrusion as (Anti)social Communication: Characterization and Detection

QI DING*, NATALLIA KATENKA†,§, PAUL BARFORD‡, ERIC KOLACZYK*, MARK CROVELLA*

*Boston University, Boston, USA,
†University of Rhode Island, Kingston, USA,
‡University of Wisconsin, Madison, USA
§email: nkatenka@cs.uri.edu 221:NatalliaKatenka.tex,session:CS3A

A reasonable definition of intrusion in a social network is: entering a community to which one does not belong. This suggests that in a network, intrusion attempts may be detected by looking for communication that does not respect community boundaries. This work evaluates a variety of ways of representing and measuring the social information contained in the Internet network flow data, and shows how to identify traffic sources that engage in (anti)social behaviour.

We find that traditional approaches to community identification operate at too coarse a level to be useful for this problem. In contrast, a more useful approach is a local one, based on the idea of a cut-vertex. We show that one can make the notion of cut-vertex more robust, in our context of intrusion detection, by using metrics from the social network literature: clustering and betweenness. We find that the use of social properties is more powerful than the use of source degree (cardinality), in the sense that it allows detection of malicious low-degree sources that can not be detected by the degree-based detector.

Having identified appropriate graph representations and appropriate socially-oriented metrics, we form and evaluate a detection system for malicious sources. Despite the fact that the resulting system operates on flow data (without content inspection), we show that the resulting system effectively detects the vast majority of sources that the DShield logs identify (using content inspection). This holds promise for the use of socially-based methods on data from core networks where many more flows are visible than at network edges (where most intrusion detection systems operate).

Overall, our results suggest that community-based methods can offer an important additional dimension for intrusion detection systems.

Acknowledgment. We are grateful to Johannes Ullrich and the SANS Institutefor making DShield logs available to us for thepurposes of this study. We also thank Michael Bailey and the IMS project members at the University of Michigan for making their honeynet data available to us. This work was supported in part by NSF grants CNS-0831427, CNS-0905186, CNS-1018266, CNS-1012910, and CNS-1117039. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the NSF. Partial support was also provided by ONR award N000140910654.

OCS2 space-time modeling

Nonstationary Spatial Modeling of Large Global Datasets

MATTHIAS KATZFUSS*,†

*Universität Heidelberg, Germany
†email: katzfuss@gmail.com 222:MatthiasKatzfuss.tex,session:OCS2

In recent years, the global coverage of data collected by satellite instruments has made it possible to analyze environmental processes on a truly global scale. However, this requires spatial statistical

models that are valid on a spherical domain, that are highly flexible to reflect the homogeneity of the globe, and that are computationally feasible to deal with the often massive satellite datasets.

For this purpose, we propose a parameterization of the nonstationary Matérn covariance function that is suitable for the sphere. This covariance function can then be used for the so-called parent process in the full-scale approximation, which combines a low-rank component and a tapered fine-scale component to obtain a computationally feasible model that is close to the parent process. The methodology is illustrated using satellite measurements of CO_2.

Possible Testing Method of Strong Non-Causality in Time Series

OCS14 Hungarian Stat. Assoc.

GALLUSZ ABALIGETI*, DÁNIEL KEHL*,†

*University of Pécs, Hungary
†email: kehld@ktk.pte.hu

223:DanielKehl.tex,session:OCS14

Researchers specializing in philosophy and methodology have been arguing about the unequivocal definition of and about the difference between the terms causality and predictability for decades. The fact that these terms are applied so extensively is due to the popularity of easy-to-operationalize tests that have widely proliferated in practical use. Authors of the related literature mention two categories of the addressed phenomenon, referred to as strong and weak causality.

The source of weak causality, often cited as Granger causality, is the change of the expected value of a time series caused by another time series. As the main difference between the two types of causality, the focal point of our analysis in case of strong causality is not the change in the expected value, but it is the change in the distribution of the process. While it is seldom challenging to directly test weak causality by testing the parameters of the appropriate regression models, the challenge becomes much more complex when it comes to analyzing the latter type of causality. This difficulty is due to our restricted capabilities of testing the continuous distributions' independence. However, if the values of the time series are discrete and have a binary nature, these hurdles are significantly easier to overcome. Facing this particular type of problem by using an appropriate function, as well as a logit or probit model, we obtain estimated conditional probabilities that are required to decide about the existence of strong causality. By testing the parameters of the regression model, the existence of causality (the lack of existence, to be more accurate) can be tested directly. This method of causality testing was used by [Mosconi and Seri (2006)].

The purpose of this research, rooted in and inspired by these two authors, is twofold. First, our goal is to investigate the available means of analyzing causality between time series that have three or more "states". Second, we are looking at the effect of continuous explanatory variables onto the set of applicable methodologies. From another point of view, in certain economic or financial analyses, it is sufficient if not desirable to only focus on more easily conceivable categories. Therefore, in these cases the continuous data set can be reduced to a discrete problem by substituting concrete values with categories such as "increase", "decrease" or "stagnation". This way, the presence of strong causality can be verified or rejected, which would not be feasible if one was strictly committed to working with the original, continuous data.

In this paper, we are introducing the aforementioned transformations that convert continuous data into a testable structure, and we also present the test results of the extended model that we have run on empirical data sets.

Acknowledgment. This research was partially supported by the Social Renewal Operational Programme, grant No.: SROP-4.2.2.C-11/1/KONV-2012-0005, Well-being in the Information Society and the Hungarian Statistical Association (Magyar Statisztikai Társaság). The authors are particularly indebted to Gábor Rappai who contributed to the development of some ideas in the paper.

References

[Mosconi and Seri (2006)] Mosconi, R., Seri, R., 2006: Non-causality in bivariate binary time series, *Journal of Econometrics*, **132 (2)**, 379 - 407.

CS23A Diffusions & Diff. Eq.

Coupling, Local Times, Immersions

WILFRID S. KENDALL*,†

*University of Warwick

†email: w.s.kendall@warwick.ac.uk 224:WilfridKendall.tex,session:CS23A

This talk is based on a case-study of the simple but potentially thematic problem of coupling Brownian motion together with its local time at 0. This problem possesses its own intrinsic interest as well as being closely related to the BKR coupling construction. Attention focusses on a simple and natural immersed (co-adapted) coupling, namely the reflection / synchronized coupling. This coupling can be shown to be optimal amongst all immersed couplings of Brownian motion together with its local time at 0, in the sense of maximizing the coupling probability at all possible times, at least when not started at pairs of initial points lying in a certain singular set. However numerical evidence indicates that the coupling is *not* a maximal coupling, and is a simple but non-trivial instance for which this distinction occurs. It is shown how the reflection / synchronized coupling can be converted into a successful equi-filtration coupling, by modifying the coupling using a deterministic time-delay and then by concatenating an infinite sequence of such modified couplings.

An application of these ideas is made to resolve affirmatively a question of Émery concerning the filtration of the BKR diffusion (BKR=Benes-Karatzas-Rishel), namely whether coupling theory can be applied to give a simple proof of the Brownian-ness of this filtration. The construction of an explicit equi-filtration coupling of two copies of the BKR diffusion follows by a direct generalization, although the proof of success for the BKR coupling requires somewhat more analysis than in the local time case.

CsörgőS Csörgő Mem. Session

Merging in Generalized St. Petersburg Games

PÉTER KEVEI*,†

*MTA–SZTE Analysis and Stochastics Research Group, Bolyai Institute, Szeged, Hungary

†email: kevei@math.u-szeged.hu 225:PeterKevei.tex,session:CsorgoS

In this talk I present some important contributions of Sándor Csörgő to one of his favourite problems, to the St. Petersburg game.

The distribution of a St. Petersburg(α, p) random variable is $P\{X = r^{k/\alpha}\} = q^{k-1}p$, $k \in \mathbb{N}$, where $q = 1 - p$, $p \in (0, 1)$ and $\alpha \in (0, 2)$. Let $X_1, X_2, \dots$ be an iid sequence of St. Petersburg(α, p) random variables, and let S_n denote their partial sum. The classical Doeblin – Gnedenko criterion implies that there is no asymptotic distribution for $(S_n - c_n)/a_n$, in the usual sense, whatever the centering and norming constants are. However, asymptotic distributions do exist along subsequences of $\mathbb{N}$. In the classical case Martin-Löf [5] 'clarified the St. Petersburg paradox', showing that $S_{2^k}/2^k - k$ converge in distribution, as $k \to \infty$. Csörgő and Dodunekova [3] showed that there are continuum different types of asymptotic distributions of $S_n/n - \log_2 n$ along different subsequences. Later Csörgő [1] proved the merging theorem $\sup_{x\in\mathbb{R}} |P\{S_n/n - \log_2 n \le x\} - G_{\alpha,p,\gamma_n}(x)| \to 0$, where $\{G_{\alpha,p,\gamma}\}_{\gamma\in(1/2,1]}$ is the family of possible limit distributions, and $\gamma_n = n/2^{\lceil \log_2 n \rceil}$ is a positional parameter. The optimality of the merge rates was proved by short asymptotic expansions by Csörgő [2]. These expansions are

given in terms of suitably chosen members from the classes of subsequential semistable infinitely divisible asymptotic distribution functions and certain derivatives of these functions, where the classes themselves are determined by the two parameters of the game.

Merging asymptotic expansions are established by Csörgő and Kevei [4] for the distribution functions of suitably centered and normed linear combinations of winnings in a full sequence of generalized St. Petersburg games, where a linear combination is viewed as the share of any one of n cooperative gamblers who play with a pooling strategy. Surprisingly, it turns out that for a subclass of strategies, not containing the averaging uniform strategy, our merging approximations reduce to asymptotic expansions of the usual type, derived from a proper limiting distribution.

References

[1] Csörgő, S., 2002: Rates of merge in generalized St. Petersburg games, *Acta Sci. Math. (Szeged)* **68**, 815–847.

[2] Csörgő, S., 2007: Merging asymptotic expansions in generalized St. Petersburg games, *Acta Sci. Math. (Szeged)* **73**, 297–331.

[3] Csörgő, S., Dodunekova, R., 1991: Limit theorems for the Petersburg game, in: *Sums, Trimmed Sums and Extremes* (M. G. Hahn, D. M. Mason and D. C. Weiner, eds.), Progress in Probability **23**, Birkhäuser (Boston, 1991), pp. 285–315.

[4] Csörgő, S., Kevei, P., 2008: Merging asymptotic expansions for cooperative gamblers in generalized St. Petersburg games. *Acta Mathematica Hungarica* **121 (1–2)**, 119–156.

[5] Martin-Löf, A., 1985: A limit theorem which clarifies the 'Petersburg paradox', *J. Appl. Probab.* **22**, 634–643.

Estimating Utilities from Individual Health Preference Data: a Nonparametric Bayesian Method

CS8B Bayesian Nonpar.

SAMER A KHARROUBI[*,‡], ANTHONY O'HAGAN, JOHN E BRAZIER[†]

[*]University of York, York, UK,
[†]University of Sheffield, Sheffield, UK.
[‡]email: samer.kharroubi@york.ac.uk 226:SamerKharroubi.tex,session:CS8B

A fundamental benefit conferred by medical treatments is to increase the health-related quality of life (HRQoL) experienced by patients. Various descriptive systems exist to define a patient's health state, and we address the problem of assigning a HRQoL value to any given state in such a descriptive system. Data derive from experiments in which individuals are asked to assign their personal values to a number of health states.

We construct a Bayesian model that takes account of various important aspects of such data. Specifically, we allow for the repeated measures feature that each individual values several different states, and the fact that individuals vary markedly in their valuations, with some people consistently providing higher valuations than others. We model the relationship between HRQoL and health state nonparametrically.

We illustrate our method using data from an experiment in which 611 individuals each valued up to 6 states in the descriptive system known as the SF-6D. Although the SF-6D distinguishes 18,000 different health states, only 249 of these were valued in this experiment. We provide posterior inference about the HRQoL values for all 18,000 states.

CS6D Dyn. Response Mod.

New Goodness-of-Fit Diagnostics for Dynamic Discrete Response Models

IGOR KHEIFETS*,‡, CARLOS VELASCO†

*New Economic School, Moscow, Russia,
†Department of Economics, Universidad Carlos III de Madrid, Madrid, Spain
‡email: ikheifets@nes.ruu 227:Kheifets.tex,session:CS6D

This paper proposes new specification tests for dynamic models with discrete responses. In particular, we test the specification of the conditional distribution of multinomial and count data, which is key to apply efficient maximum likelihood methods, to obtain consistent estimates of partial effects and to get appropriate predictions of the probability of future events. The traditional approach is based on a continuation random transformation of discrete data which leads to continuous uniform iid series under the true conditional distribution. Then standard specification testing techniques can be applied to the transformed series, but the extra random noise involved in the continuation may affect power properties of these methods. We investigate in this paper an alternative estimate of a cumulative distribution function based only on discrete data which can be compared directly to a continuous standard uniform cdf. We analyze the asymptotic properties of goodness-of-fit tests based on this new approach and explore the properties in finite samples of a bootstrap algorithm to approximate the critical values of test statistics which are model and parameter dependent. We find that in many relevant cases our new approach performs much better than random-continuation counterparts.

Acknowledgment. This research was supported by the the Spanish Ministry of Economy and Competitiveness, grant No.: SEJ2007-62908.

OCS14 Hungarian Stat. Assoc.

The Economic Spatial Structure Of Europe Considered By A Modelling Approach

ÁRON KINCSES*,†, GÉZA TÓTH*, ZOLTÁN NAGY*

*Hungarian Central Statistical Office, Budapest, Hungary
†email: aron.kincses@ksh.hu 228:AronKincses.tex,session:OCS14

Many theoretical and practical works aim at describing the spatial structure of Europe. The spatial relations have undergone continuous change, so their examination is always considered apposite. In our study, we give an overview of models describing the spatial structure of Europe. We illustrate their variegation by listing, without any claim to completeness a part of them. Our study aims at describing the economic spatial structure of Europe with bidimensional regression analysis based on gravitational model. With help of using the gravity shift-based model, we can clearly justify the veracity of models of different methodological background based.

CS39A Distributio Theory

Some Moment-Indeterminate Distributions from Economics and Actuarial Science

CHRISTIAN KLEIBER*,†

*Universität Basel, Switzerland
†email: christian.kleiber@unibas.ch 229:Kleiber.tex,session:CS39A

The moment problem asks whether or not a given probability distribution is uniquely character-

ized by the sequence of its moments. The existence of moment-indeterminate distributions has been known since the late 19th century; however, first examples were often considered as mere mathematical curiosities. Recent research has shown that the problem is more widespread than previously thought. Specifically, distributions that are not determined by their moments arise in economics, notably the modelling of economic size distributions, and also in mathematical finance and in actuarial science.

Here we present several specific examples of moment-indeterminate distributions: the generalized lognormal distribution, a heavy-tailed distribution arising, for example, in the EGARCH model of asset price volatility and in income distribution modelling, the Benini distribution, again a model for the size distribution of income, and the Benktander distributions, which were proposed as claim size distributions in actuarial science.

Conditions for moment (in)determinacy in these models are investigated, they sometimes depend on values of parameters of the distributions. For those distributions that are moment-indeterminate explicit examples of Stieltjes classes comprising moment-equivalent distributions are presented.

Model Calibration Under Space-Time Misalignment

OCS2 space-time modeling

WILLIAM KLEIBER*,†

*Department of Applied Mathematics, University of Colorado, Boulder, CO, USA
†email: william.kleiber@colorado.edu 230:William_Kleiber.tex,session:OCS2

A common feature of dynamical computer models is physical space displacement of space-time features from field data. Specifically, the computer model may predict accurate spatial patterns that show differences from observations as the result of improper geographical alignment. The traditional approach to correcting model discrepancy is to introduce an additive and/or multiplicative bias. This traditional technique is unable to reward model settings exhibiting correct spatial patterns which may be slightly displaced and thus can be at odds with expert judgement regarding model accuracy. We introduce an alternative approach to model calibration in the presence of space-time displacement discrepancy. Borrowing ideas from the image warping literature, we propose a nonlinear deformation of the computer model that optimally aligns with observed images; probabilistically this manifests as a transformation of model coordinate space with a variational penalty on the likelihood function. We apply the approach to a dynamical magnetosphere-ionosphere computer model that exhibits challenging displacement discrepancies, and successfully identify a region of input parameter space that simultaneously minimizes model error and discrepancy from field data.

Iterative Scaling in Curved Exponential Families

CS35A Discretre Response M.

ANNA KLIMOVA*,‡, TAMÁS RUDAS†

*Institute of Science and Technology (IST) Austria
†Eötvös Loránd University, Budapest, Hungary
‡email: anna.klimova@ist.ac.at 231:Anna_Klimova.tex,session:CS35A

Some areas of statistical learning, among which are text processing, computer tomography, and market basket analysis, employ feature selection procedures. A feature can be seen as a categorical variable, an indicator of a characteristic. A common property of a set of objects can be expressed by a combination of features.

Modeling associations of features of a finite set of objects can be implemented within the relational model framework proposed in [3]. Under such a model, the log probability of a cell (an object) is equal to the sum of the parameters each associated with a combination of cells.

Relational models under which all objects in the sample space have a common characteristic are said to have an overall effect. Such models are regular exponential families, and the standard theory of maximum likelihood estimation applies. In order to compute the MLE in such models, iterative proportional fitting (IPF) or its generalizations, for example, Generalized Iterative Scaling (GIS) [1], and Improved Iterative Scaling (IIS) [2] can be used.

However, under some relational models describing associations of features, there is no characteristic that is common to all objects. The assumption of the overall effect is not feasible, and the lack of presence of overall effect cannot be changed by a re-parameterization. Models for probabilities without the overall effect are curved exponential families. The observed totals of the subsets of objects with a common feature, although sufficient statistics, are preserved by the MLE only up to a constant of proportionality, called the adjustment factor. Neither GIS nor IIS account for the presence of the adjustment factor; if applied to a model without the overall effect, these algorithms do not converge or may converge to a vector of probabilities that do not sum to 1.

In this talk, a generalization of IPF that can be used for models without the overall effect is presented and illustrated with an example. The algorithm complements the IPF core with a stepwise selection of the adjustment factor. The proof of convergence, based on minimizing the Bregman distance, is briefly discussed.

References

[1] Darroch, J., Ratcliff, D., 1972: Generalized iterative scaling for log-linear models, *The Annals of Mathematical Statistics*, **43**, 1470 - 1480.

[2] Della Pietra, S., Della Pietra, V., Lafferty, J.,1997: Inducing features of random fields, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **19**, 283 - 297.

[3] Klimova, A., Rudas, T., Dobra, A., 2012: Relational models for contingency tables, *J. Multivariate Anal.*, **104**, 159 - 173.

CS8A Bayesian Semipar.

An Irregular Semiparametric Bernstein–von Mises Theorem

BAS KLEIJN*, BARTEK KNAPIK†,‡,§

*Korteweg–de Vries Institute of Mathematics, University of Amsterdam, The Netherlands,
†CREST, ENSAE, Malakoff, France,
‡CEREMADE, Université Paris-Dauphine, France
§email: knapik@ceremade.dauphine.fr 232:BartekKnapik.tex,session:CS8A

In recent years, frequentist evaluation of Bayesian approaches to nonparametric models has enjoyed much attention. One of the important aspects studied in the literature is asymptotic efficiency of Bayesian semiparametric methods. The general question concerns a nonparametric model indexed by two parameters: a finite-dimensional parameter of interest, and an infinite-dimensional nuisance parameter; the exclusive interest goes to the estimation of the former. Asymptotically, regularity of the estimator combined with the Cramér–Rao bound in the Gaussian location model that forms the limit experiment fixes the rate of convergence to $n^{-1/2}$ and poses a bound to the accuracy of regular estimators expressed, e.g., through Hajék's convolution and asymptotic minimax theorems.

In Bayesian context, efficiency of estimation is best captured by a so-called Bernstein–von Mises limit. Just like frequentist parametric theory for regular estimates extends quite effortlessly to regular semiparametric problems, semiparametric extensions of Bernstein–von Mises-type asymptotic behavior of posteriors can be obtained without essential problems. Although far from developed fully, some general considerations of Bayesian semiparametric efficiency, as well as model- and/or prior-specific derivations of the Bernstein–von Mises limit exist in the literature. Some authors even consider infinite-dimensional limiting posteriors (notwithstanding the objections raised in the past).

However, not all estimators are regular. The quintessential example calls for estimation of a point of discontinuity of a density: to be a bit more specific, consider an almost-everywhere differentiable Lebesgue density on $\mathbb{R}$ that displays a jump at some point $\theta \in \mathbb{R}$; estimators for θ exist that converge at rate n^{-1} with exponential limit distributions.

We consider estimation of a parameter of interest θ based on a sample $X_1, \ldots, X_n$, distributed i.i.d. according to some unknown P_{θ_0,η_0}, where η_0 denotes the nuisance parameter. We shed some light on the behavior of marginal posteriors for the parameter of interest in semiparametric, irregular estimation problems, through a study of the Bernstein–von Mises phenomenon. The models considered in this talk exhibit a weakly converging expansion of the likelihood called *local asymptotic exponentiality (LAE)*, to be compared with local asymptotic normality in regular problems. This type of asymptotic behavior of the likelihood is expected to give rise to a (negative-)exponential marginal posterior satisfying the irregular Bernstein–von Mises limit:

$$\sup_A \Big| \Pi_n\big(h \in A \;\big|\; X_1, \ldots, X_n \big) - \mathrm{Exp}^-_{\Delta_n, \gamma_{\theta_0,\eta_0}}(A) \Big| \xrightarrow{P_0} 0,$$

where $h = n(\theta - \theta_0)$, $\mathrm{Exp}^-_{\Delta,\gamma}$ denotes the negative exponential distribution supported on $(-\infty, \Delta]$ with the scale parameter γ, and the random sequence Δ_n converges weakly to exponentiality. Like in the regular case, the above limit allows for the asymptotic identification of credible sets with confidence intervals associated with the maximum likelihood estimator. The constant γ_{θ_0,η_0} determines the scale in the limiting exponential distribution and, as such, the width of credible sets. In this talk, we present general sufficient conditions on model and prior to conclude that the above limit obtains.

The main theorem is applied in two semiparametric LAE example models. The first is an extension of a problem of estimation of the shift parameter in the family of exponential distributions with a fixed scaled parameter, and the latter includes a problem of estimation of the scale parameter in the family of uniform distributions $[0, \lambda]$, $(\lambda > 0)$.

Bayesian Inference in Cyclostationary Time Series Model with Missing Observations

OCS26 Resampling Nonstat T.S.

OSKAR KNAPIK*,†

*Department of Statistics, Cracow University of Economics, Cracow, Poland
†email: knapiko@uek.krakow.pl 233:OskarKnapik.tex,session:OCS26

In recent years, there is a growing interest in modelling nonstationary time series. Periodically correlated, almost periodically correlated, cyclostationary time series form important examples of such models. The survey of Gardner, Napolitano, Paura (2006) is quoting over 1500 different papers recently published that are dedicated to cyclostationarity. Almost periodically correlated (APC) time series have found applications in various areas such as econometrics, signal processing, communications and biology. The purpose of this paper is to provide Bayesian inference for such signals within parametric statistical model. The statistical inference is based on the Markov Chain Monte Carlo (MCMC) methods. In particular, data augmentation technique supported by the Metropilis-Hastings within Gibbs sampler algorithm is proposed to conduct Bayesian inference for unknown quantities of the model. The whole is complemented with simulations.

CS6B Funct. Est., Regression

Regularizing Linear Programming Estimation for Nonregular Regression Models

KEITH KNIGHT*,†

*University of Toronto, Toronto, Canada

†email: keith@utstat.toronto.edu 234:KeithKnight.tex,session:CS6B

We will consider estimation in regression models where the errors are non-regular; examples include models with positive errors, bounded errors, and errors whose densities have jump discontinuities. In such models, estimation based on linear programming arises naturally and, under appropriate regularity conditions, the resulting estimators typically converge in distribution to the solution of a linear program whose constraints are a random set determined by a Poisson process. However, if the errors are sufficiently non-homogeneous, linear programming estimation will break down in the sense that it will not be able to attain its optimal convergence rate. In this talk, we will discuss regularizing linear programming estimation using simple quadratic penalization in order to recover the optimal convergence rate as well as some computational methods.

CS5C H-D Var. Selection

Oracle Inequalities for High-Dimensional Panel Data Models

ANDERS BREDAHL KOCK*

*Aarhus University, Denmark

†email: akock@creates.au.dk 235:AndersKock.tex,session:CS5C

When building a statistical model one of the first decisions one has to make is which variables are to be included in the model and which are to be left out. Since high-dimensional data sets are becoming increasingly available, the last 10-15 years has witnessed a great deal of research into procedures that can handle such data sets. In particular, a lot of attention has been given to penalized estimators. The Lasso is probably the most prominent of these procedures and a lot of research has focussed on investigating the theoretical properties of it. The Lasso-type procedures have become popular since they are computationally feasible and perform variable selection and parameter estimation at the same time.

However, most focus in the literature has been on the standard linear regression model. In this paper we shall investigate the properties of the Lasso and the adaptive Lasso in the linear fixed effects panel data model

$$y_{i,t} = x'_{i,t}\beta^* + c_i^* + \epsilon_{i,t}, \quad i = 1, \dots, N,\ t = 1, \dots, T \tag{1}$$

where $x_{i,t}$ is a $p_{N,T} \times 1$ vector of covariates where $p_{N,T}$ is indexed by N and T to indicate that the number of covariates can increase in the sample size (N and T). In the sequel we shall omit this indexation. N is often interpreted as the number of individuals sampled while T is the number of periods in which each person is sampled. The individuals will be assumed independent while no restrictions are made on the dependence of the variables over time for each individual. The c_is are the unobserved time homogeneous heterogeneitys while $\epsilon_{i,t}$ are the error terms. $x_{i,t}$ might be a very long vector – potentially much longer than the sample size. On the other hand, only a few of the variables in $x_{i,t}$ might be relevant for explaining $y_{i,t}$ meaning that the vector β^* is sparse.

Oftentimes the unobserved heterogeneity $c_{i,t}$ is simply removed by a differencing or demeaning procedure. However, just like β^*, $c^* = (c_1, \dots, c_N)$ might be a (sparse) vector and hence it is our goal

to investigate the properties of the Lasso for fixed effects panel models. We shall see that the Lasso can estimate the two parameter vectors almost as precisely as if the true sparsity pattern had been known and only the relevant variables had been included from the outset. In particular, we

1. provide *nonasymptotic* oracle inequalities for the estimation error of the Lasso for β^* and c^* under different sets of assumptions on the covariates and the error terms. More precisely, for a given sample size we provide upper bounds on the estimation error which hold with at least a certain probability.
2. show that our bounds are optimal in the sense that they can at most be improved by a multiplicative constant.
3. use the nonasymptotic bounds to give a set of sufficient conditions under which the Lasso estimates β^* and c^* consistently in high-dimensional models. We also give conditions under which Lasso does not discard any relevant variables.
4. establish nonasymptotic lower bounds on the probability with which the adaptive Lasso unveils the correct sparsity pattern.
5. use the nonasymptotic bounds to give conditions under which the adaptive Lasso detects the correct sparsity pattern asymptotically.

Estimation of Integrated Covariances in the Simultaneous Presence of Nonsynchronicity, Noise and Jumps

OCS6 Asympt. for Stoch Proc.

YUTA KOIKE*,†

*University of Tokyo, Japan
†email: kyuta@ms.u-tokyo.ac.jp

236:yutakoike.tex,session:OCS6

Let Z^1 and Z^2 be two Itô semimartingales. In general the quadratic covariation of Z^1 and Z^2 consists of two sources; the continuous martingale parts and the co-jumps of the semimartingales. In this talk we will focus on disentangling these two components of the quadratic covariation of Z^1 and Z^2 by using high-frequency observation data.

Let $(S^i)_{i\in\mathbb{Z}_+}$ and $(T^j)_{j\in\mathbb{Z}_+}$ each are observation times of Z^1 and Z^2 respectively. We assume that the observation data $\mathsf{Z}^1 = (\mathsf{Z}^1_{S^i})_{i\in\mathbb{Z}_+}$ and $\mathsf{Z}^2 = (\mathsf{Z}^2_{T^j})_{j\in\mathbb{Z}_+}$ of Z^1 and Z^2 are contaminated by some noise, that is, $\mathsf{Z}^1_{S^i} = Z^1_{S^i} + U^1_{S^i}$ and $\mathsf{Z}^2_{T^j} = Z^2_{T^j} + U^2_{T^j}$. Here the observation noise $(U^1_{S^i})_{i\in\mathbb{Z}_+}$ and $(U^2_{T^j})_{j\in\mathbb{Z}_+}$ are known as market microstructure noise in econometrics. Our aim is to estimate the integrated covariance of Z^1 and Z^2 (i.e. the quadratic covariation of the continuous martingale parts of them) from the observation data Z^1 and Z^2. There are mainly three difficulties in this problem: separating the jumps, removing the noise and dealing with the nonsynchronicity of observation times. On separating the jumps there exist two techniques popular among the literature: one is the *bipower technique* (proposed in [1]) and the other is the *thresholding technique* (proposed in [3] and [5]). In the present situation, however, the former cannot be applied due to the nonsynchronicity. The latter also cannot be applied *directly* because the observed returns are too noisy. To overcome this issue, we first implement the *pre-averaging procedure* (proposed in [4]) for removing the noise, and after that we apply the thresholding technique for separating the jumps. Finally, by using the *Hayashi-Yoshida method* (proposed in [2]) that enables us to deal with the nonsynchronicity, we construct a new estimator for the integrated covariance.

We will show the asymptotic mixed normality of this estimator under some mild conditions allowing infinite activity jump processes with finite variations, some dependency between the sampling times and the observed processes as well as a kind of endogenous observation errors.

References

[1] O. E. Barndorff-Nielsen, N. Shephard, Power and bipower variation with stochastic volatility and jumps, Journal of Financial Econometrics 2 (2004) 1–37.

[2] T. Hayashi, N. Yoshida, On covariation estimation of non-synchronously observed diffusion processes, Bernoulli 11 (2005) 359–379.

[3] C. Mancini, Disentangling the jumps of the diffusion in a geometric jumping Brownian motion, Giornale dell'Istituto Italiano degli Attuari 64 (2001) 19–47.

[4] M. Podolskij, M. Vetter, Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps, Bernoulli 15 (2009) 634–658.

[5] Y. Shimizu, Estimation of diffusion processes with jumps from discrete observations, Master's thesis, University of Tokyo, 2003.

OCS1 Longitudin Models

Joint Modelling of Longitudinal Outcome and Competing Risks

RUWANTHI KOLAMUNNAGE DONA*,†

*University of Liverpool

†email: Ruwanthi.Kolamunnage-Dona@liverpool.ac.uk

237:RuwanthiKolamunnageDona.tex,session:OCS1

Available methods for joint modelling of longitudinal and survival data typically have only one failure type for the time to event outcome. We extend the methodology to allow for competing risks data. When there are several reasons why the event can occur, or some informative censoring occurs, it is known as 'competing risks'. We fit a cause-specific hazards sub-model to allow for competing risks, with a separate latent association between longitudinal measurements and each cause of failure. The method is applied to data from the SANAD (Standard And New Antiepileptic Drugs) trial of anti-epileptic drugs (AEDs), as a means of investigating the effect of drug titration on the relative effects of AEDs Lamotrigine and Carbamazepine on treatment failure.

IS4 Empirical Proc.

On the Problem of Optimality in Aggregation Theory

LECUÉ GUILLAUME*,†

*CNRS, Ecole Polytechnique

†email: guillaume.lecue@cmap.polytechnique.fr

238:Lecue.tex,session:IS4

Consider the bounded regression framework with random design and a finite class F of regression functions, called a dictionary. The aim in aggregation theory is to construct procedures having a risk at least as good as the best element in the dictionary. For that, we want to prove exact oracle inequalities : given a risk function $R(\cdot)$, we want to construct a procedure $\hat{f}_n$ such that with high probability

$$R(\hat{f}_n) \leq \min_{f \in F} R(f) + \text{residue}$$

where the residual term should be as small as possible. For the quadratic risk, the residual term should be of the order of $(\log |F|)/n$. So far there has been two optimal aggregation procedures achieving this rate (in deviation):

1. the "selection-convexification" method of G. Lecué and S. Mendelson (cf. [3]),

2. the "star" algorithm of J.-Y. audibert (cf. [1]).

In this talk, I will in particular prove that the "Q-aggregation" method of P. Rigollet (cf. [6, 2]) initially introduced in the Gaussian regression model with fixed design provides a third optimal way of aggregating experts in our setup. This optimality result holds for any risk associated with a Lipschitz and strongly convex loss function. I will also prove that a prior probability measure can be assigned to all of the elements in the dictionary and that the Q-aggregation procedure can take into account this prior (cf. [4]).

I will also speak about the convex aggregation problem where the aim is to mimic the best element in the convex hull of F. For this problem, I will prove that empirical risk minimization is optimal (cf. [5]).

We will then finish the talk by introducing some open problems related to the optimality of empirical risk minimization and the geometry of the model.

References

[1] AUDIBERT, J.-Y. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems (NIPS)* (2007).

[2] DAI, D., RIGOLLET, P., AND ZHANG, T. Deviation optimal learning using greedy q-aggregation. *Ann. Statist.* (March 2012). arXiv: 1203.2507.

[3] LECUÉ, G., AND MENDELSON, S. Aggregation via empirical risk minimization. *Probab. Theory Related Fields 145*, 3-4 (2009), 591–613.

[4] LECUÉ, G., AND RIGOLLET, P. Optimal learning with Q-aggregation. *Submitted 2013*

[5] LECUÉ, G. Empirical risk minimization is optimal for the convex aggregation problem. *To appear in Bernoulli journal.*

[6] RIGOLLET, P. Kullback–leibler aggregation and misspecified generalized linear models. *Ann. Statist. 40*, 2 (2012), 639–665.

Sensitivity Analysis of Stochastic Biochemical Systems. Inference, Experimental Design and Information Processing

OCS31 Stoch. Molecular Biol.

MICHAŁ KOMOROWSKI*,†

*Institute of Fundamental Technological Research, Polish Academy of Sciences
†email: mkomor@ippt.gov.pl 239:MichalKomorowski.tex,session:OCS31

Investigating how stochastic systems respond to parameter perturbations is an important element to utilise stochastic models in biomolecular sciences as they constitute a natural a framework to represent intracellular dynamics. I will present a statistical framework that allows for inference of kinetic parameters of such systems, efficient computation of the Fisher Information and quantification of their information transmission capacity. The framework is based on the Linear Noise Approximation. It allows for a comprehensive analysis of stochastic models by means of ordinary differential equations only. The main aim of the talk will be to present developed statistical theory , and present examples demonstrating the potential of these methods.

Asymptotic Behavior of CLSE for 2-type Doubly Symmetric Critical Branching Processes with Immigration

OCS29 Stat. Branching Proc.

KRISTÓF KÖRMENDI*,†, GYULA PAP*

*University of Szeged, Hungary
†email: kormendi@math.u-szeged.hu 240:KristofKormendi.tex,session:OCS29

The model is as follows. For each $k, j \in \mathbb{Z}_+$ and $i, \ell \in \{1, 2\}$, the number of individuals of type

i in the k^{th} generation will be denoted by $X_{k,i}$, the number of type ℓ offsprings produced by the j^{th} individual who is of type i belonging to the $(k-1)^{\text{th}}$ generation will be denoted by $\xi_{k,j,i,\ell}$, and the number of type i immigrants in the k^{th} generation will be denoted by $\varepsilon_{k,i}$. Then

$$\begin{bmatrix} X_{k,1} \\ X_{k,2} \end{bmatrix} = \sum_{j=1}^{X_{k-1,1}} \begin{bmatrix} \xi_{k,j,1,1} \\ \xi_{k,j,1,2} \end{bmatrix} + \sum_{j=1}^{X_{k-1,2}} \begin{bmatrix} \xi_{k,j,2,1} \\ \xi_{k,j,2,2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{k,1} \\ \varepsilon_{k,2} \end{bmatrix}, \qquad k \in \mathbb{N}.$$

We will suppose that the offspring mean matrix has the form

$$\begin{bmatrix} \mathbb{E}(\xi_{1,1,1,1}) & \mathbb{E}(\xi_{1,1,1,2}) \\ \mathbb{E}(\xi_{1,1,2,1}) & \mathbb{E}(\xi_{1,1,2,2}) \end{bmatrix} := \begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}.$$

We call the process critical if the offspring mean matrix has spectral radius 1, which in our case is equivalent to $\alpha + \beta = 1$. Our aim is to find the asymptotic behavior of the conditional least squares estimators (CLSE) of the parameters α, β and the criticality parameter $\varrho = \alpha + \beta$. We find that the asymptotic behavior is different in 3 distinct cases, if the total number of offsprings or the difference between the number of type 1 and type 2 offsprings is degenerate then we find asymptotic normality, however in the third case the asymptotic distribution of the estimators can only be expressed with stochastic integrals of a diffusion process. We also consider the analogous question for continuous time branching processes with immigration.

Acknowledgment. This research have been supported by the Hungarian Chinese Intergovernmental S & T Co-operation Programme for 2011-2013 under Grant No. 10-1-2011-0079 and by the Hungarian Scientific Research Fund under Grant No. OTKA T-079128.

References

[1] Ispány, M., Körmendi, K. and Pap, G. (2012) Statistical inference for 2-type doubly symmetric critical Galton–Watson processes with immigration., arXiv: 1210.8315

POSTER
Poster

On Some Aspects of Fisher Information as a Measure of Neural Stimulus Optimality

LUBOMIR KOSTAL*,†, PETR LANSKY*, STEVAN PILARSKI*

*Institute of Physiology AS CR, Czech Republic

†email: kostal@biomed.cas.cz

241:Kostal_poster.tex,session:POSTER

The most frequently employed methods for determining the stimulus optimality conditions are those of information theory and statistical estimation theory. In our contribution we concentrate on the latter, with Fisher information playing the key role. Although the Fisher information is employed ubiquitously, there are situations where the results should be interpreted with care. We describe a simple biologically motivated system for which the Fisher information does not provide useful insight on the efficiency of stimulus estimation. Furthermore, we show that in some situations, the scale of stimulus may affect the inference on the optimal stimulus level. In addition, we compare the actual performance of maximum likelihood and moment estimators with the theoretical bound as determined by Fisher information.

Acknowledgment. This research was by the Institute of Physiology RVO: 67985823, Centre for Neuroscience P304/12/G069 and the Grant Agency of the Czech Republic projects P103/11/0282 and P103/12/P558.

Stochastic Inference of Dynamic System Models: From Single-molecule Experiments to Statistical Estimation

IS24 Single Molecule Exp.

SAMUEL KOU*,†

*Harvard University, Cambridge, MA, USA
†email: kou@stat.harvard.edu 242:SamuelKOU.tex,session:IS24

Dynamic systems, often described by coupled differential equations, are used in modeling diverse behaviors in a wide variety of scientific disciplines. In this talk we will consider their assessment and calibration in light of experimental/observational data. For the assessment, we explore how the deterministic dynamic system models reconcile with stochastic observations, using recent single-molecule experiments on enzymatic reactions as an example, where the single-molecule data reveal clear departure from the classical Michaelis-Menten model. For the calibration, we will propose a new inference method for the parameter estimation of dynamic systems. The new method employs Gaussian processes to mirror a dynamic system and offers large savings of computational time while still retains high estimation accuracy. Numerical examples will be used to illustrate the estimation method.

Goodness-of-Fit Tests for Long Memory Moving-Average Marginal Density

OCS25 Long-mem. Time Ser.

HIRA L. KOUL*, NAO MIMOTO*, DONATAS SURGAILIS†

*Michigan State University, East Lansing, USA, †Vilnius Institute of Mathematics and Informatics, Lithuania

243:Koul.tex,session:OCS25

In this talk we will discuss the problem of fitting a known d.f. or density to the marginal error density of a stationary long memory moving-average process when its mean is known and unknown. When the mean is unknown and estimated by the sample mean, the first-order difference between the residual empirical and null distribution functions is known to be asymptotically degenerate at zero. Hence, it cannot be used to fit a distribution up to an unknown mean. However, we shall show that by using a suitable class of estimators of the mean, this first order degeneracy does not occur. We also present some large sample properties of the tests based on an integrated squared-difference between kernel-type error density estimators and the expected value of the error density estimator based on errors. The asymptotic null distributions of suitably standardized test statistics are shown to be chi-square with one degree of freedom in both cases of known and unknown mean. This is totally unlike the i.i.d. errors set-up where suitable standardizations of these statistics are known to be asymptotically normally distributed.

Generalized Logistic Distributions and their Applications in Finance

CS40A Logistic & Multinom. Distr.

VASILEIOS KOUTRAS*, KONSTANTINOS DRAKOS*, MARKOS V. KOUTRAS†,‡

*Department of Accounting and Finance, Athens University of Economics and Business, Athens, Greece,
†Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece
‡email: mkoutras@unipi.gr 244:MARKOSKOUTRAS.tex,session:CS40A

The aim of the present work is to introduce families of distributions which include as special

case the classical Logistic distribution and offer quite remarkable adaptability in real data arising in finance. The key point for the development of our models is the well known property of the Logistic distribution that the logit transformation of its cdf, i.e. $log(F(x)/(1-F(x)))$ is a linear function of x.

The introduction of new families of distributions which provide flexibility and adaptability in fitting well to real and empirical data is an important statistical issue. The capability of the Logistic distribution to describe satisfactory real data, has initiated considerable research on producing generalizations more flexible than the classical model, with obvious aim to establish better data fitting, see Balakrishnan (1992), Gupta and Kundu (2010) and references therein. A plausible generalization of the Logistic distribution arises quite naturally by considering a family of distributions having respective logit transformation of Polynomial type (see Koutras *et al.* (2013)). Apparently, when the need for fitting real data in a probability model arises, the new family is a more powerful choice than the classical Logistic family since there exist additional parameters which can be appropriately adjusted to improve the fit.

In the present article we discuss several generalizations of the Logistic distribution which provide excellent fit in financial data where the presence of skewness and asymmetry is quite common, and the use of models with heavy tails is indispensable. As an illustration we construct appropriate distributional models for the Euro foreign exchange reference rates provided by the European Central Bank.

Acknowledgment. This research was partially supported by the University of Piraeus Research Center and the Athens University of Economics and Business Research Center

References

[Balakrishnan (1992)] Balakrishnan, N., 1992 : *Handbook of the Logistic Distribution*, Statistics: Textbooks and Monographs, vol. 123, Marcel Dekker, New York.

[Gupta, R. D. and Kundu, D. (2010)] Gupta, R. D. and Kundu, D., 2010: Generalized logistic distributions, *Journal of Applied Statistical Sciences*, **18**,51-66.

[Koutras, V, Drakos, K. and Koutras, M. V. (2013)] Koutras, V, Drakos, K. and Koutras, M. V., 2013: A Polynomial Logistic distribution and its applications in finance, *Communications in Statistics: Theory and Methods* (accepted for publication).

CS25B Risk Mgment

A Migration Approach for USA Banks' Capitalization

VASILEIOS KOUTRAS*,†, KONSTANTINOS DRAKOS*

*Department of Accounting and Finance, Athens University of Economics and Business, Athens, Greece

†email: vkoutras@aueb.gr 245:VASILEIOSKOUTRAS.tex,session:CS25B

Since the early 90's the Basel agreement has put in place a capitalization threshold, relative to risk-weighted assets, above which banks are advised to operate. In particular, banks were expected to maintain a Capital Ratio (CAR), defined as Tier1+Tier2/Risk-Weighted Assets, of at least 8%. The Federal Deposit Insurance Corporation (FDIC), the US banking sector regulator, has adopted a finer classification of capitalization using 5 buckets according to the value of CAR , which, at least for regulatory purposes, is preferred to the simple dichotomous Basel classification.

In the present study we employ a panel dataset whose cross section consists of all US commercial banks and its time series dimension of the period 1992-2009 (annual intervals). Adopting the FDIC bucketing approach, we introduce a Markov Chain setup whose states correspond to the FDIC buckets, and estimate the transition probabilities between capitalization buckets, including the default state (Ross (1996)) for two sub periods - the 90's and the 00's. The estimated transition probability matrices are then used to compare the two decades in terms of persistence and mobility based on

appropriate indices (Jafry and Schuermann (2004), Shorrocks (1978)).The benefit of this approach is that it allows the quantification of migration probabilities of banks across the capitalization spectrum, among which those related to the default state are of special interest. Moreover, the analysis permits estimating direction-specific migration probabilities, i.e improvements vs. deteriorations of capitalization. This information is particularly useful both for market participants and regulators.

An additional comparison of the two periods is carried out by exploiting appropriate distance metrics (Trück and Rachev (2009)). Essentially we compare the two sub-periods by evaluating the distances of the transition probability matrices of each period to the identity matrix in order to assess their proximity to complete persistence (perfect immobility). Based on a battery of mobility indices we document a substantial mobility increase in the post 2000 era. What is more important is that this mobility takes the form of increased probability of capitalization deterioration. That is, during the 00's in comparison to the 90's, the US banking sector has exhibited greater mobility, with capitalization deteriorations been more likely and more abrupt.

Future research could explore possible higher order time dependence of the Markov Chain. Additionally, one could allow for geographic and size dependent transition probabilities.

Acknowledgment. This research was partially supported by the Athens University of Economics and Business Research Center

References

[Jafry, Y., and Schuermann, T. (2004)] Jafry, Y., and Schuermann, T., 2004: Measurement, estimation and comparison of credit migration matrices, *Jour. of Banking and Finance* , **28**, 2603-2639.

[Ross, S., (1996)] Ross, S., 1996 : *Stochastic Processes* (2nd Edition), John Wiley & Sons Inc., NY.

[Shorrocks, A., (1978)] Shorrocks, A., 1978: The measurement of mobility, *Econometrica* , **46**, 1013-1024.

[Trück, S. and Rachev, T. (2009)] Truck, S., and Rachev, T., 2009: *Rating Based Modeling of Credit Risk Theory and Applications of Migration Matrices* , (2nd Edition), Elsevier Inc.

On a Mixed-Moments Method of Estimation

CS37A Estim. Methods

IOANNIS A. KOUTROUVELIS*,†

*Department of Engineering Sciences, University of Patras, Greece

†email: koutrouv@upatras.gr 246:Koutrouvelis.tex,session:CS37A

Maximum likelihood estimation for certain parametric families of distributions is faced with theoretical/computational difficulties or requires the evaluation of special functions. Examples of such families are provided by the three-parameter gamma distribution and the Poisson-exponential distribution. This paper examines the application of a mixed-moments procedure for fitting certain parametric families of distributions having a closed form for their moment generating function. The procedure uses the mean of the original data and a number of fractional moments of the exponentially transformed data. It combines the method of moments with a regression method based on the empirical moment generating function. The asymptotic normal distribution of the resulting parameter estimators is derived. The performance of the procedure is illustrated for fitting various skewed distributions having applications in stochastic hydrology including gamma, inverse Gaussian and PE distributions.

References

[Koutrouvelis and Canavos (1999)] Koutrouvelis, I. A. and Canavos, G. C., 1999: Estimation in the Pearson type 3 distribution, *Water Resources Research*, **35**, 2693-2704.

[Koutrouvelis et al. (2005)] Koutrouvelis, I. A., Canavos, G. C. and Meintanis, S. G., 2005: Estimation in the three-parameter inverse Gaussian distribution, *Computational Statsitics & Data Analysiss*, **49**, 1132-1147.

POSTER
Poster

Bayesian Interpolation of Random Point Events: A Path Integral Analysis

SHINSUKE KOYAMA*,†

*The Institute of Statistical Mathematics, Tokyo, Japan
†email: skoyama@ism.ac.jp

247:Koyama.tex,session:POSTER

We try to capture the rate of event occurrence, such as neuronal spikes, as a function of time. In order to estimate an underlying rate function from a sequence of events, it is necessary to perform temporal smoothing over a short time window at each time point. In Bayesian analysis, in which the assumption on the smoothness is incorporated in the prior probability of underlying rate, the timescale of the temporal average, or the degree of smoothness, may be optimized by maximizing the marginal likelihood. Here, the marginal likelihood is obtained by marginalizing the complete-data likelihood over all possible latent rate processes. We carry out this marginalization by a path integral method. It is shown that there exists a lower bound for the "degree" of rate variation below which the optimal smoothness parameter diverges. This implies that no inference on the temporal variation of the rate can be made even with an infinitely long sequence of events.

Let $\{t_i\} := \{t_1, t_2, \ldots, t_n\}$ be a sequence of event times occurred in the time interval $[0, T]$. We assume that $\{t_i\}$ is derived from an inhomogeneous Poisson process with the rate function $\lambda(t)$. The probability density of $\{t_i\}$, conditioned on $\{\lambda(t)\} := \{\lambda(t) : 0 \leq t \leq T\}$ is given by

$$p(\{t_i\}|\{\lambda(t)\}) = \exp\left[-\int_0^T \lambda(t)dt\right] \prod_{i=1}^{n} \lambda(t_i). \tag{1}$$

In order to estimate $\{\lambda(t)\}$ from $\{t_i\}$, we employ a Bayesian method. We introduce a "roughness-penalizing prior" on $\{\lambda(t)\}$ as

$$p(\{\lambda(t)\}|\beta) = \frac{1}{Z(\beta)} \exp\left[-\beta \int_0^T \left(\frac{d^\alpha \lambda(t)}{dt^\alpha}\right)^2 dt\right], \tag{2}$$

where β is the hyperparameter controlling the smoothness of $\lambda(t)$, and $\alpha \in \{1, 2, \ldots\}$ is the order of smoothness (which is typically taken to be α =1 or 2 [Ramsay and Silverman (2010)]). The inference of $\lambda(t)$ is then made with the posterior probability of $\lambda(t)$, given $\{t_i\}$, which is obtained by the Bayes' theorem.

The smoothness hyperparameter β is determined by maximizing the marginal likelihood function:

$$p(\{t_i\}|\beta) = \int p(\{t_i\}|\{\lambda(t)\}) p(\{\lambda(t)\}|\beta) \mathcal{D}\{\lambda(t)\}, \tag{3}$$

where $\int \mathcal{D}\{\lambda(t)\}$ represent a path integral over all possible rate processes. In this work, we develop a path integral method [Kleinert (2009)] to perform this marginalization over the functional space.

Acknowledgment. The author was supported by JSPS KAKENHI Grant Number 24700287.

References

[Kleinert (2009)] Kleinert, H., 2009: Path Integrals in Quantum Mechanics, Statistics, Polymer Physics and Financial Markets, World Scientific Publishing Company, 5th edition.

[Ramsay and Silverman (2010)] Ramsay, J., Silverman, B. W., 2010: Functional Data Analysis, Springer, 2nd edition.

A New Estimator of Mean in Randomized Response Models

CS6D Dyn. Response Mod.

NURSEL KOYUNCU*,†

*Department of Statistics, Hacettepe University, Ankara, Turkey
†email: nkoyuncu@hacettepe.edu.tr 248:NURSEL_KOYUNCU.tex,session:CS6D

Surveys on highly personnel and stigmatizing questions such as gambling, abortion, alcoholism, and many others are affected unreliable and misleading answers. Respondents may refuse to answer the sensitive questions or give untruthful answers. To overcome this difficulty, Warner (1965) introduced the method of randomized response (RR) for surveying human populations for obtaining information on variables of sensitive protecting the anonymity of the respondents. This pioneering method aims to ensure efficient and unbiased estimation protecting a respondent's privacy. In this method respondents are instructed to give partially misclassified responses by using a randomization device such as deck of cards. Many new models are developed for sensitive question which require "yes" or "no" response. In the recent studies, Diana and Perri (2011) give attention to sensitive question results in a quantitative variable. Koyuncu et al. (2013) introduced exponential estimators to improve the efficiency of mean estimator based on randomized response technique. In this talk we have proposed a new estimator in randomized response models. The expression of mean square error has been obtained by theoretical way. The gain in efficiency over the existing estimators has been shown with a simulation study.

References

[Koyuncu et al. (2013)] Koyuncu, N., Gupta, S., Sousa, R., 2013: Exponential Type Estimators of the Mean of a Sensitive Variable in the Presence of Non-Sensitive Auxiliary Information, *Communications in Statistics: Simulation and Computation*, (article in pres)

[Diana and Perri (2011)] Diana G., Perri P.F., 2011: A class of estimators for quantitative sensitive data, *Statistical Papers*, **52**, 633 - 650.

[Warner (2011)] Warner, S. L., 1965: Randomized response: a survey technique for eliminating evasive answer bias, J. *Amer. Statist. Assoc.*, **60**, 63 - 69.

Using Definitive Screening Designs to Get More Information from Fewer Trials

CS32A Nonparame

VOLKER KRAFT*,†, TOM DONNELLY*

*SAS Institute, JMP Division, Germany
†email: volker.kraft@jmp.com 249:VolkerKraft.tex,session:CS32A

This presentation will focus on recent Design of Experiments screening methods,not only for efficiently screening factors but for using the screening data to more rapidly develop second-order predictive models. Definitive Screening designs will be shown to not only detect main effects but also curvature in each factorŰand do so in fewer trials than traditional fractional-factorial designs that when a center point is included can only detect curvature globally.

These new designs when first published in 2011 could support only continuous factors. Improvements published in March of 2013,now enable them to support categorical factors with two levels. When the number of significant factors is small, a Definitive Screening design can collapse into a Şone-shotŤ design capable of supporting a response-surface model with which accurate predictions can be made about the characterized process.

A case study will be shown in which a 10-factor process is optimized in just 24 trials. Checkpoint trials at predicted optimal conditions show the process yield increased by more than 20collapse into a one-shot design, the existing trials can economically be augmented to support a response-surface model in the important factors. Graphical comparisons between these alternative methods and traditional designs will show the new ones to yield more information in often fewer trials.

CS6C Func. Est., Smoothing

Smoothing Parameter Selection in Two Frameworks for Spline Estimators

TATYANA KRIVOBOKOVA[*,†]

[*]Georg-August-Universität Göttingen, Germany
[†]email: tkrivob@uni-goettingen.de 250:TatyanaKrivobokova.tex,session:CS6C

In contrast to other nonparameteric regression techniques, smoothing parameter selection for spline estimators can be performed not only by employing criteria that approximate the average mean squared error (e.g. generalized cross validation), but also by making use of the maximum likelihood (or empirical Bayes) paradigm. In the later case, the function to be estimated is assumed to be a realization of some stochastic process, rather than from a certain class of smooth functions. Under this assumption both smoothing parameter selectors for spline estimators are well-studied and known to perform similar. A more interesting problem is the properties of smoothing parameter estimators in the frequentist framework, that is if the underlying function is non-random. In this talk we discuss the asymptotic properties of both smoothing parameter selection criteria for general low-rank spline smoothers in the frequentist framework and give insights into their small sample performance.

NYA Not Yet Arranged

Goodness-of-Fit Test for Gaussian Regression with Block Correlated Errors

SYLVIE HUET[*], ESTELLE KUHN[*,†]

[*]INRA MIA UR 0341, Domaine de Vilvert, 78352 JOUY-EN-JOSAS Cedex, FRANCE
[†]email: estelle.kuhn@jouy.inra.fr 251:estellekuhn.tex,session:NYA

Let us consider a multivariate Gaussian regression model of size nJ and assume that its covariance matrix has a block diagonal structure composed of n squared blocks of size J. This assumption means that the n blocks of size J behave independently but that there exists possible correlation among the observations within the same block. The expectation and the covariance matrix of the Gaussian model are known up to a fixed number of real parameters and can depend on some covariates. This framework appears for example when considering n independent observations of a Gaussian multiresponse regression model of size J with heteroscedastic error. Many models of various application fields are also covered by our framework. Let us quote for example mixed effects models with correlated within subject errors and autocorrelated errors models for longitudinal data.

Our aim is to test the null hypothesis that the expectation of the Gaussian vector belongs to a linear subspace V of $\mathbb{R}^{nJ}$ against the alternative that it does not. This issue belongs to the usually so called goodness-of-fit or lack-of-fit testing procedures. It has been adressed already by several authors under several model assumptions. However to our best knowledge there exists no goodness-of-fit testing procedure adapted to the case of multivariate Gaussian regression model with unknown block diagonal covariance matrix.

Therefore we propose a new goodness-of-fit test for testing that the expectation of a Gaussian vector of size nJ with block diagonal covariance matrix composed of n blocks of size J belongs to a specified linear subspace V of $\mathbb{R}^{nJ}$. Since we aim at considering a nonparametric alternative, we base our test on multiple testing procedure that is to say a global procedure that involves several tests againts several parametric alternatives. Each of these alternatives is characterized by a linear subspace orthogonal to the linear subspace V. To get an efficient testing procedure we consider an alternative composed of several linear subspaces whose dimension may grow with the number of observations n. We state the existence and the consistency of the estimate of the parameters involved both in the expectation and in the covariance matrix of the model under the null hypothesis and under the alternative hypotheses in order to get asymptotic properties when n goes to infinity. Then we prove that our test as well as its bootstrap version achieve the nominal level and are consistent. We prove also that the test is consistent against local alternative approaching the null hypothesis at the $1/\sqrt{n}$ rate up to a factor $\sqrt{\log\log(n)}$. We illustrate the behaviour of our procedure for finite sample size on the basis of a simulation study.

Modeling Corporate Exits Using Competing Risks Models

OCS5 Anal Complex Data

REG KULPERGER*

University of Western Ontario *

252:Kulperger.tex,session:OCS5

We discuss modeling corporate exits from a publicly traded system, specifically publicly traded US corporations. There are two types of exits, merger and acquisition, and bankruptcy. Each company files financial statements on a quarterly basis, thus giving information on the health of the individual company. Thus each corporation has many covariates available on its health. A natural model is thus a competing risks exit model, specifically a log linear hazards model. Conditional on the covariates, for a given period (quarter year) the corporations are conditionally independent. This is similar to a biostatistical model, but with one additional feature, namely calendar time. This allows one to naturally consider both constant and time evolving baseline hazards. As with such models one also needs to somehow measure whether the models fits the data. These various aspects are discussed.

Jittered Phase Diagrams for Seasonal Patterns in Time Series

CS4A Time Series II.

ROBERT M. KUNST*,†,‡

*Institute for Advanced Studies Vienna, Austria,
†University of Vienna, Austria
‡email: kunst@ihs.ac.at

253:RobertMKunst.tex,session:CS4A

In the analysis of seasonal time series, an issue of main concern is the discrimination of deterministic patterns (shape reversion) and seasonal unit roots (permanent shape transition). We suggest complementing the customary hypothesis tests with jittered phase diagrams on a discretized set of basic seasonal shapes that represent the dynamic transition between the patterns. Discretization relies on upward and downward movements between seasons, which yields eight classes for quarterly data. These jittered phase diagrams provide a convenient visualization that supports the discrimination among the main classes of potential seasonal data-generating processes.

For several process classes of interest, characteristic patterns are determined. Data-generating processes with seasonal unit roots tend to produce clearly recognizable saltire shapes in the charts.

Transitions across classes are rare but tend to become permanent. By contrast, deterministic seasonal variation is reflected in blurred or irregular patterns that prefer specific classes for their entire lifetimes. Periodic models typically generate contaminated saltire shapes and indications of cyclical movement among several particular shape classes. Weak or stationary stochastic seasonal variation entails frequent and unsystematic moves across all shape classes.

Whereas we mainly aim at a visualization device, we also consider a nonparametric hypothesis test constructed from the charts. The suggested test statistic is a weighted average $\xi_1 + \lambda\xi_2$ of the basic statistics ξ_1 and ξ_2. ξ_1 measures the distance between the actual shape in the chart and a pure saltire cross. It is related to the range unit-root test of APARICIO *et al.* (2006). ξ_2 measures the frequency of changing the shape class and is related to the zero-crossings counts statistic of BURRIDGE *et al.* (1996). Some simulations explore the optimal weight λ. Generally, the power properties of the nonparametric test are found to be satisfactory, even for simple parametric designs where parametric tests are known to dominate.

The framework is mainly tuned to quarterly data that admit the most characteristic visualization. Variants for the important monthly case, however, are also investigated. The considerable increase in the number of seasonal shape classes from eight to 2^{11} precludes the visualization of the within-class shapes and emphasizes the inspection of transition patterns between classes.

The procedure is applied to exemplary economic variables, such as national accounts data and the unemployment rate, and also to some other variables, such as temperature and precipitation measurements. The applications confirm the traded wisdom that empirically observed seasonality is primarily deterministic. They also convey the impression that larger samples are needed to discriminate safely among the basic seasonal features than are often available.

References

[Aparicio et al.] Aparicio, F., Escribano, A., Sipols, A.E., 2006: Range unit-root (RUR) tests: robust against nonlinearities, error distributions, structural breaks and outliers, *Journal of Time Series Analysis*, **27**, 545–576.

[Burridge et al.] Burridge, P., and Guerre, E., 1996: The Limit Distribution of Level Crossings of a Random Walk, and a Simple Unit Root Test. *Econometric Theory*, **12**, 705–723.

CS5B H-D Distribution

Pseudo-Gibbs Distribution and Its Application on Multivariate Two-Sample Test

KUN-LIN KUO*,†

*Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan

†email: klkuo@nuk.edu.tw 254:KunLinKuo.tex,session:CS5B

We say that the given conditional distributions are compatible if they can be derived from a joint distribution. Generally, we implement Gibbs sampler via compatible conditional distributions. Under compatibility, different scan orders will lead to the same stationary distribution. A Gibbs sampler defined by an ensemble of incompatible conditional models is called a pseudo-Gibbs sampler, whose stationary distribution is termed a pseudo-Gibbs distribution. Unlike the conventional Gibbs sampler, different scan orders will produce different pseudo-Gibbs distributions, but they are not entirely different. In the literature, little is discussed about the limiting distributions of a incompatible Gibbs sampler. We will provide some interesting properties for these pseudo-Gibbs distributions.

Testing whether two samples are consistent with a single unknown distribution is a task that occurs in many areas of research. The classical two-sample test does not have a natural extension to comparing two multivariate populations. In this study, we apply the properties of the pseudo-Gibbs distributions to address the multivariate two-sample problem.

Acknowledgment. This research was partially supported by the National Science Council of Taiwan under grant NSC 101-2118-M-390-001.

Dynamic Factor Analysis of Environmental Systems: III. Applications in Environmental Management and Decision

OCS10 Dynamic Factor Models

YI-MING KUO*,‖, HONE-JAY CHU†, HWA-LUNG YU‡, TSUNG-YI PAN‡, CHENG-SHIN JANG§, HSING-JUH LIN¶

*Ming Dao University. Chang-Hua, Taiwan,
†National Cheng Kung University, Tainan, Taiwan,
‡National Taiwan University, Taiwan,
§Kainan University, Taiwan,
¶National Chung Hsing University, Taichung, Taiwan
‖email: airkuo@ntu.edu.tw 255:YiMing_Kuo.tex,session:OCS10

Dynamic factor analysis (DFA) is a dimension-reduction technique especially designed for time-series data. DFA is able to identify underlying common trends (unexplained variability) between multivariate time series and can evaluate interactions with selected potential explanatory variables. This paper summarized the applications of DFA to (1) investigate the source contributions of PM2.5 by monitoring data collected at the four aerosol supersites in Southern Taiwan throughout 2009; (2) determine the main factors regulating temporal and spatial variations in the water quality in the Kaoping River Estuary over a 9-year period (2003–2011); (3) examine environmental factors which are most responsible for the 8-year temporal dynamics of the intertidal seagrass Thalassia hemprichii in southern Taiwan; (4) investigate common trends in annual extreme precipitation for 24-h duration (AM24 h) and annual maximum rainfall depths for 1-h duration (AM1 h) of the same event at 16 rainfall stations in the Kaoping River watershed, Southern Taiwan; (5) identify common trends that represent unexplained variability in ground water arsenic (As) concentrations of decommissioned wells and to investigate whether explanatory variables (total organic carbon, As, alkalinity, ground water elevation, and rainfall) affect the temporal variation in ground water As concentration; (6) determine the key factors regulating temporal and spatial variations of phytoplankton abundance at three monitoring sites in Shihmen Reservoir during five-year (2006–2010) observations. DFA enhances our understanding of the influences of environmental variables on our concerning variables. The significant variables and common trends determined from DFA can be applied in environmental management and decision. Decision maker can use above information for setting water quality criteria, increasing monitoring variables, regulating hydrological conditions, and determining pollution prevention plans in order to sustainably design our environment.

References

[1] Kuo, Y.M., Jang, C.S., Yu, H.L., Chen, S.C., Chu, H.J., 2013. Identifying nearshore groundwater and river hydrochemical variables influencing water quality of Kaoping River Eestuary using dynamic factor analysis. Journal of Hydrology. Published on line.

[2] Kuo, Y.M., Chu, H.J., Pan, T.Y., 2013. Temporal Precipitation Estimation from nearby Radar Reflectivity using Dynamic Factor Analysis in the Mountainous Watershed - a case during Typhoon Morakot, Hydrological Processes. DOI: 10.1002/hyp.9639.

[3] Chu, H.J., Lin, C.Y., Liau, C.J., Kuo, Y.M., 2012. Identify controlling factors of ground-level ozone levels over southwestern Taiwan using a decision Tree. Atmospheric Environment 60: 142-152.

[4] Kuo, Y.M., Wang S.W., Jang, C.S., Yeh, N.C., Yu, H.L., 2011. Identifying the factors influencing PM2.5 in southern Taiwan using dynamic factor analysis. Atmospheric Environment 45: 7276-7285.

[5] Kuo, Y.M., Chu, H.J., Pan, T.Y., Yu, H.L., 2011. Investigating common trends of annual maximum rainfalls during heavy rainfall events in southern Taiwan. Journal of Hydrology 409: 749-758.

[6] Kuo, Y.M., Chang, F.J., 2010. Dynamic factor analysis for estimating groundwater arsenic trends. Journal of Environmental Quality. 39: 176-184.

[7] Kuo, Y.M., Lin, H.J., 2010. Dynamic factor analysis of long-term growth trends of the intertidal seagrass Thalassia hemprichii in southern Taiwan. Estuarine, Coastal and Shelf Science 86: 225-236.

NYA
Not Yet
Arranged

Temporal Data Processing

MICHAL KVET*,†, KAROL MATIASKO*,‡

*University of Zilina, Faculty of Management Science and Informatics, Slovakia
email: †Michal.Kvet@fri.uniza.sk, ‡Karol.Matiasko@fri.uniza.sk
256:MichalKvet.tex,session:NYA

Massive development of data processing requires new functions to provide fast, faithful and easy approach to the data attributes representing the state of an object during the time.

Database systems are fundamental part of information systems; they belong to the most important parts of the information technologies and are used not only in standard applications, but also in critical applications for energetics, traffic, medicine or industry.

Most of data in a database represent the current state. The change of the attribute value causes update of the row and the old data are deleted. However, each value has its own history; progress, which can be necessary to be stored and monitored. That is the reason of creating new concept of the database.

Backups and log files were the keystone of the historical data processing in the past. Most of the users and IT professionals used to ignore them. If they had to be used, it meant that there was a serious problem and the last valid state of the database have to be restored.

Data processing requires keeping actual, historical and future valid data in the database. Conventional tables store only actual valid objects.

Uni-temporal table uses composite primary key (ID, BD, ED), which does not consist of only identifier (ID), but also time of the validity. It causes that each object is represented by different number of rows. BD represents the begin date validity, ED reflects the end date of the attributes validity. They offer storing whole states of the object during its existence, even after logical delete. This paper decribes the problems and defines solutions; special methods, structures, operations to provide easy access to these data. Also future valid data should be stored; if the begin time of the new validity occurs, state of the object is automatically updated without user interaction.

Acknowledgment. This contribution is the result of the project implementation: Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research and Development Operational Programme funded by the ERDF. Podporujeme vyskumne aktivity na Slovensku/Projekt je spolufinancovany zo zdrojov EU.

References

[1] C. J. Date, H. Darwen, N. A. Lorentzos - Temporal data and the relational model, Morgan Kaufmann, 2003. ISBN: 1558608559

[2] Ch. S. Jensen, R. T. Snodgrass - Temporally Enhanced Database Design

[3] T. Johnson, R. Weis - Managing Time in Relational Databases, Morgan Kaufmann, 2010. ISBN: 9780123750419

[4] M. Kvet, A. Lieskovsky, K. Matiasko - Temporal data modelling, 2013.(IEEE conference ICCSE 2013, 4.26. - 4.28.2013) In press

[5] J. Patel - Temporal DB System R. T. Snodgrass - Developing Time-Oriented Database Applications in SQL, Morgan Kaufmann, 1999. ISBN: 9781558604360

Diffusion Approximation of Neuronal Models Revisited

POSTER
Poster

PETR LANSKY[*,†], JAKUB CUPERA[*]

[*]Institute of Physiology AS CR, Prague, Czech Republic
[†]email: lansky@biomed.cas.cz 257:PetrLansky_poster.tex,session:POSTER

Stochastic neuronal models are a common tool for investigation of information transfer within the nervous system. Among them, the diffusion models play an important role for their relatively simple tractability. However, models with reversal potentials have several alternative diffusion approximations. In our contribution, the probability distributions of the first-passage time for the original model and its diffusion approximations are numerically compared in order to find which of the approximations is the most suitable one. The properties of the random amplitudes of postsynaptic potentials are discussed. It is shown on a simple example that the quality of the approximation depends directly on them.

Bayesian Modelling of Root and Leaf Transfer and Phytotoxicity of Metals From Particulate Matter

POSTER
Poster

THOMAS PUECHLONG[*,†], CHRISTOPHE LAPLANCHE[*], TIANTIAN XIONG[*], ANNABELLE AUSTRUY[*], CAMILLE DUMAT[*]

[*]Université de Toulouse ; INP, UPS, CNRS ; EcoLab (Laboratoire Ecologie Fonctionnelle et Environnement); ENSAT, Avenue de l'Agrobiopole, F-31326 Castanet Tolosant, France.
[†]email: thomas.puechlong@univ-tlse3.fr 258:Christophe.Laplanche.tex,session:POSTER

The impact of particulate matter (PM) root and shoot uptake is evaluated on edible plants. Plants are cultivated in the laboratory, different amount of PM are introduced (by leaf or soil solution deposition). To study the kinetics of metals transfer and storage, a measure of metals concentrations present in the leaves and roots is performed 5, 10, and 15 days after exposure. In order to have a quantification of physiological responses depending on the concentration absorbed by the organism, the lipid composition, leaf and root biomass and water content are measured. In the case of leaves, additional measures are carried out, namely gas exchange and chlorophyll concentrations. Concentration and effect measurements are destructive and performed at the individual level. This difficulty is apprehended by modeling individual variability of transfer and effect process as random variability of transfer and effect parameters. The model variables are well defined on 3 levels (duration, exposure condition, and individual). In order to link the metal transfer kinetics in the plant with physiological response a dynamic model of transfer and effect is developed. This model should be able to predict the intensity of physiological responses depending of the metal, its concentration, the exposure condition (foliar and/or root) and the duration of exposure. It is inspired by three compartment pharmacokinetics/pharmacodynamics [Antic. (2009)] Bayesian model [Lunn et al. (2002)], this type of model has been historically developed for mammals which drugs are administered. Nevertheless, it is adapt to plants that administers PM *via* different absorption, transfer or effect models that has been developed to measure the heavy metals impacts on vegetation. The model is implemented in the free software OpenBUGS. OpenBUGS, from a statistical description model (a priori distribution of parameters of interest, measured variables distribution and relationships between variables), performs a posteriori sampling of model parameters using a MCMC (Markov Chain Monte Carlo) method. The Bayesian approach advantages is able to consider the parameters priors distribution from the literature, distributions variables measured of different natures and non-linear relationships between variables. Estimates of marginal a posteriori distributions parameters as well as assessing

adequacy penalized by the complexity of model are two guides that will draw conclusions on the process driving transfer and effect of PM on plants.

Acknowledgment. This research was supported by ADEME project DIMENSION.

References

[Lunn et al. (2002)] Lunn, L., Best, N., Thomas, A., Wakefield, J., Spiegelhalter, D., 2002: Bayesian analysis of population PK/PD models : General concepts and software., *Journal of Pharmacokinetics and Pharmacodynamics*, **29(3)**, 271-307.

[Antic. (2009)] Antic, J., 2009: Méthodes non-paramétriques en pharmacocinétiques et/ou pharmacodynamie de population, Doctoral dissertation, *Université de Toulouse, Université Toulouse III-Paul Sabatier*.

CS12A Hierarchica Bayesian

Hierarchical Bayesian Modelling of Brown Trout (*Salmo trutta*) Growth: A Tool for Sustainable Fishery Management in Navarra, Northern Spain

CHRISTOPHE LAPLANCHE[*,†], JOSÉ ARDAÍZ[‡], PEDRO M. LEUNDA[§]

[*]Université de Toulouse; INP, UPS; EcoLab (Laboratoire Ecologie Fonctionnelle et Environnement); ENSAT, Avenue de l'Agrobiopole, 31326 Castanet Tolosan, France,
[†]CNRS; EcoLab; 31326 Castanet Tolosan, France,
[‡]Gobierno de Navarra, Departamento de Desarrollo Rural y Medio Ambiente, c/ Gonzalez Tablas 9, 31005 Pamplona/Iruña, Navarra, Spain,
[§]Gestión Ambiental de Navarra S.A., c/ Padre Adoaín 219 Bajo, 31013 Pamplona/Iruña, Navarra, Spain
email: christophe.laplanche@ensat.fr 259:LaplancheChristophe.tex,session:CS12A

Brown trout (Salmo trutta) has strict habitat and water quality demands, and nowadays, most wild populations are subjected to severe anthropogenic disturbances (e.g. global warming, water regulation, overfishing). A key environmental variable driving brown trout life history is water temperature and a key descriptor of brown trout populations is growth rate. We have constructed a statistical model which simulates brown trout growth in order to relate fish length to air temperature (Ruiz and Laplanche 2010, Lecomte and Laplanche 2012). The model is built as a Bayesian Hierarchical model (HBM) in order to cope with the non-linearities of the model links, random effects, and the nested structure of the model variables. The model is run with water/air temperature data and fish length data collected over 19 years on 61 points distributed over a 120x80 km area in the Navarra province, Northern Spain. The model allows the exploration of the growth rate under the form of a spatio-temporal map over the hydrologic network, reconstructing data of the past and predicting trends for the next future. The Bayesian model of the growth will be a useful tool for sustainable fishery management of wild brown trout populations in Navarra.

References

[Ruiz and Laplanche (2010)] Ruiz, P., Laplanche, C., 2010: A hierarchical model to estimate the abundance and biomass of salmonids by using removal sampling and biometric data from multiple locations, *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 2032-2044.

[Lecomte and Laplanche (2012)] Lecomte, J.-B., Laplanche, C., 2012: A length-based hierarchical model of brown trout (Salmo trutta fario) growth and production, *Biometrical Journal*, **54**, 108-126.

Numerical Approximation of the Stochastic Heat Equation Based on a Space-Time Variational Formulation

OCS22 Numeric SPDE

STIG LARSSON*,†, MATTEO MOLTENI*

*Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE–41296 Gothenburg, Sweden
†email: stig@chalmers.se 260:LarssonStig.tex,session:OCS22

We consider the linear heat equation perturbed by colored multiplicative noise. The equation is formulated using a weak space-time variational formulation, where the equation is multiplied by a smooth test function depending on the spatial and temporal variables. By a partial integration over space-time, the first order time and space derivatives are placed on the test function. In the deterministic case it is known that the resulting problem has a unique solution by using an abstract inf-sup theorem. In the stochastic case we need to allow the test functions to be stochastic processes and we integrate also with respect to the probability measure. We extend the inf-sup theorem to the stochastic case and obtain unique solvability.

Spatial discretization by a standard finite element method is also analyzed in the same weak variational framework. Mean square error estimates are proved.

A lot of the literature on numerical methods for stochastic evolution problems is based on the semigroup approach of Da Prato and Zabczyk. This work is a first attempt to develop analysis techniques for numerical methods for the stochastic heat equation, which do not rely on the semigroup formulation.

Linear-Rational Term Structure Models

IS29 Stoch. in Finance

MARTIN LARSSON*,†,, DAMIR FILIPOVIĆ*,, ANDERS TROLLE*

*Swiss Finance Institute, Ecole Polytechnique Fédérale de Lausanne
†email: martin.larsson@epfl.ch 261:Larsson.tex,session:IS29

Linear-Rational term structure models are pricing models based on factor processes with affine drift, in conjunction with a state price density modeled as an affine function of the state. This combination leads to flexible term structure models that are able to combine several attractive features within a single framework: the short rate is guaranteed to stay nonnegative; many key quantities are available in closed form; swaption prices can be computed very efficiently; and unspanned stochastic volatility can be incorporated in a simple and natural way. The classical affine models are unable to achieve this combination. In this talk I will describe the Linear-Rational framework and its various features, and show how this enables us to calibrate the model simultaneously to the dynamics of swap rates and implied swaption volatilities.

Variance Bounding Markov Chain Monte Carlo Methods for Bayesian Inference with Intractable Likelihoods

IS1 Bayesian Comp.

ANTHONY LEE*,†

*University of Warwick
†email: anthony.lee@warwick.ac.uk 262:AnthonyLee.tex,session:IS1

The continued development and practical success of Markov chain Monte Carlo methodology has contributed to the proliferation of the use of Bayesian inference across a number of application

domains. This is at least partially due to the generality with which quantities of interest for diverse models can be estimated via partial sums of a Markov chain simulated on a computer. Motivated by such successes, and by modern statistical applications, there has been interest amongst practitioners to use increasingly sophisticated models to describe data. In some cases the likelihood function arising is in some sense intractable, precluding use of standard methods. The use of recently proposed "pseudo-marginal" Markov kernels have alleviated in some cases the difficulties that arise in this context, and have laid a benchmark for the evaluation of new kernels. We show that in a number of scenarios, these kernels, while useful, can fail to be variance bounding and hence geometrically ergodic. However, we propose a new and related class of Markov kernels that can systematically inherit variance bounding from an appropriately defined "idealized" kernel under verifiable regularity conditions.

Acknowledgment. This talk is derived from two projects. One is joint work with Christophe Andrieu (University of Bristol) and Arnaud Doucet (University of Oxford), and the other with Krzysztof Łatuszyński (University of Warwick).

CS9B Model Sel, Info Crit

Model Selection via Bayesian Information Criterion for Quantile Regression Models

EUN RYUNG LEE[*,§], HOHSUK NOH[†], BYEONG U. PARK[‡]

[*]University of Mannheim, Mannheim, Germany,
[†]Université catholique de Louvain, Louvain-la-Neuve, Belgium,
[‡]Seoul National University, Seoul, Korea
[§]email: elee@mail.uni-mannheim.de 263:EunRyungLee.tex,session:CS9B

Bayesian Information Criterion (BIC) is known to identify the true model consistently as long as the predictor dimension is finite. Recently, its moderate modifications have been shown to be consistent in model selection even when the number of variables diverges. Those works have been done mostly in mean regression, but rarely in quantile regression. The best known results about BIC for quantile regression are for linear models with a fixed number of variables. In this paper, we investigate how BIC can be adapted to high-dimensional linear quantile regression and show that a modified BIC is consistent in model selection when the number of variables diverges as the sample size increases. We also discuss how it can be used for choosing the regularization parameters of penalized approaches that are designed to conduct variable selection and shrinkage estimation simultaneously. Moreover, we extend the results to structured nonparametric quantile models with a diverging number of covariates. We illustrate our theoretical results via some simulated examples and a real data analysis on human eye disease.

CS5A H-D Dim. Reduction

On the Conditional Distributions of Low-Dimensional Projections from High-Dimensional Data

HANNES LEEB

University of Vienna
email: hannes.leeb@univie.ac.at 264:HannesLeeb.tex,session:CS5A

This talk presents results of Leeb (2013). We study the conditional distribution of low-dimensional projections from high-dimensional data, where the conditioning is on other low-dimensional projections. To fix ideas, consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. Moreover, consider two projections defined by unit-vectors α and β,

namely a response $y = \alpha' Z$ and an explanatory variable $x = \beta' Z$. It has long been known that the conditional mean of y given x is approximately linear in x, under some regularity conditions; cf. Hall and Li (1993). However, a corresponding result for the conditional variance has not been available so far. We here show that the conditional variance of y given x is approximately constant in x (again, under some regularity conditions). These results hold uniformly in α and for most β's, provided only that the dimension of Z is large. In that sense, we see that most linear submodels of a high-dimensional overall model are approximately correct. Our findings provide new insights in a variety of modeling scenarios. We discuss several examples, including sliced inverse regression, sliced average variance estimation, generalized linear models under potential link violation, and sparse linear modeling.

References

[1] P. Hall and K.-C. Li. On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**:867–889, 1993.

[2] H. Leeb. On the conditional distribution of low-dimensional projections from high-dimensional data. *Ann. Statist.*, forthcoming, 2013.

Reduction of Chemical Reaction Networks II: Michaelis-Menten and Beyond

OCS31 Stoch. Molecular Biol.

ANDRÉ LEIER*, MANUEL BARRIO†, TATIANA T. MARQUEZ-LAGO‡

*Okinawa Institute of Science and Technology Graduate University, Onna-son, Japan,
†Departamento de Informática, Universidad de Valladolid, Valladolid, Spain,
‡Integrative Systems Biology Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Japan

265:AndreLeier.tex,session:OCS31

We will discuss several expansions to the abridgment method using delay distributions presented previously in 'Reduction of Chemical Reaction Networks I'. In particular, we will show how such delay distributions can be derived for chemical reaction systems that include other types of monomolecular reactions such as constitutive synthesis, degradation, or backward and forward bypass reactions. We will argue why for some scenarios one must adopt a numerical approach for obtaining accurate stochastic representations, and propose two alternatives for this. In such cases, the accuracy depends on the respective numerical sample size. Additionally, we will address binary and Michaelis-Menten type reactions and the conditions under which our method can still yield good approximations.

In general, our model reduction methodology yields significantly lower computational costs while retaining accuracy. Quite naturally, computational costs increase alongside network size and separation of time scales. Thus, we expect our model reduction methodologies to significantly decrease computational costs in these instances.

References

[Barrio et al. (2013)] Barrio, M., Leier, A., Marquez-Lago, T.T., 2013: Reduction of chemical reaction networks through delay distributions, *J. Chem. Phys.*, **138**, 104114.

OCS24 Random Graphs

Order Statistics and the Length of the Best-of-n-of-$(2n-1)$ Competition

TAMÁS LENGYEL*,†

*Occidental College, Los Angeles, USA

†email: lengyel@oxy.edu

266:TamasLengyel.tex,session:OCS24

World Series (best-4-of-7) type competitions are popular playoff formats to decide the champion in most North American professional sports and they give rise to interesting questions regarding the winning probability, the expected length of the generalized series (cf. [Lengyel (1993)]), the limit distribution of the length, and the statistical inference about the closeness of fit to real data. We refer to the best-of-n-of-$(2n-1)$ series as game A. In each single game Teams 1 and 2 win with probability p and $q = 1-p$, respectively. The series ends when one of the teams, the winning team, accumulates n wins for the first time. Unconditionally extending the number of games to $2n-1$ results in game B. In game B the team which accumulates more wins than the other is called the winner. The extension offers a nice way to calculate the series winning probabilities in game A. If $X \sim Binomial[2n-1, p]$ is the number of wins by Team 1 in game B then the probability of winning can be easily represented by $P(X \geq n)$ in both games. Similar methods for calculating the expected length and its variance of the original series have not been known. We prove the following

Theorem 5. *Let $X \sim Binomial[2n-1, p]$, $X' \sim Binomial[2n, p]$, and $X'' \sim Binomial[2n+1, p]$ be the number of wins in a series of $2n-1$, $2n$, and $2n+1$ games, respectively, and Y be the length of the best-n-of-$(2n-1)$ series, each with single game winning probability p, then*

$$E(Y) = n\left(1 + \frac{p}{q}P(X \leq n-2) + \frac{q}{p}P(X \geq n+1)\right)$$

and

$$E(Y) = n\left(\frac{1}{q}P(X' \leq n-1) + \frac{1}{p}P(X' \geq n+1)\right).$$

For the variance, we have that

$$\mathrm{var}(Y) = n\left(\frac{n+1}{q^2}P(X'' \leq n-1) + \frac{n+1}{p^2}P(X'' \geq n+2) - \frac{1}{q}P(X' \leq n-1) - \frac{1}{p}P(X' \geq n+1)\right) - (E(Y)^2).$$

If $p = q = 1/2$ then $E(Y) = 2n\left(1 - \binom{2n}{n}/2^{2n}\right)$ and $\mathrm{var}(Y) = 2n\left(1 - \binom{2n}{n}/2^{2n} - 2n\left(\binom{2n}{n}/2^{2n}\right)^2\right)$.

Note that all of the above are closed formulas in terms of the distribution function of various binomial distributions. The nature of the result for $p = q = 1/2$ calls for a derivation using 0-1 variables, however, it requires a somewhat unexpected approach based on order statistics (cf. [Arnold et al. (2008)]), conditioned on the number of wins by Team 1 in game B.

References

[Arnold et al. (2008)] Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N., 2008: A First Course in Order Statistics, SIAM Classics in Applied Mathematics, 54.

[Lengyel (1993)] Lengyel, T., 1993: A combinatorial identity and the World Series, *SIAM Review*, **35**, 294–297.

Estimation of Deformations between Distributions by Minimal Wasserstein Distance

CS5B H-D Distribution

HÉLÈNE LESCORNEL*,†, JEAN-MICHEL LOUBES*

*Institut de Mathématiques de Toulouse, France
†email: helene.lescornel@math.univ-toulouse.fr 267:HeleneLescornel.tex,session:CS5B

We study a model where observations are coming from several deformations of the same distribution. More precisely, we assume that we have at hand observations of the following random variables

$$X_j = \varphi_{\theta_j^\star}(\varepsilon), \quad 1 \leqslant j \leqslant J.$$

We denote by μ the structural distribution of the model, that is the law of the variable ε. The deformation corresponds to the function $\varphi_{\theta_j^\star}$.

Here we consider that the shape of the deformation is available through the function φ but that its amount represented by the parameter $\theta_j^\star \in \mathbb{R}^d$ is unknown . Our aim is to propose estimators for the quantities $\theta^\star = (\theta_1^\star, \ldots, \theta_J^\star) \in \Theta \subset \mathbb{R}^{Jd}$ and μ, and to present their asymptotic properties.

For that, we assume that we have at hand the following i.i.d. observations : $X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij})$, $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant J$.

The idea is to align the distributions of the random variables X_j. For that, for $\theta = (\theta_1, \ldots, \theta_J) \in \Theta$ we consider the random variables

$$Z_{ij}(\theta) = \varphi_{\theta_j}^{-1}(X_{ij}) = \varphi_{\theta_j}^{-1} \circ \varphi_{\theta_j^\star}(\varepsilon_{ij}) \sim \mu_j(\theta) \quad \forall 1 \leqslant i \leqslant n.$$

By varying the parameter θ, we aim to align the distributions $\mu_j(\theta)$. Indeed, we have that $\mu_j(\theta^\star) = \mu$, $\forall 1 \leqslant j \leqslant J$. To quantify the alignment, we consider their Wasserstein distance of order 2 which has the following expression

$$W_2^2(\mu_{j-1}(\theta), \mu_j(\theta)) = \int_0^1 \left(\left(F_{j-1}^\theta\right)^{-1}(t) - \left(F_j^\theta\right)^{-1}(t) \right)^2 dt$$

with F_j^θ the distribution function of $\mu_j(\theta)$.

Hence this leads us to consider the criterion $M(\theta) = \frac{1}{J-1} \sum_{j=2}^J W_2^2(\mu_{j-1}(\theta), \mu_j(\theta))$ which permits a characterization of the parameter of interest : $M(\theta^\star) = 0 = \min_{\theta \in \Theta} M(\theta)$.

The data permits to approach this theoritical criterion by replacing the laws by their empirical version. Hence we obtain the following criterion, expressed through the order statistics of the samples $(Z_{ij}(\theta))_{1 \leqslant i \leqslant n}$, for j from 1 to J.

$$M_n(\theta) = \frac{1}{J-1} \sum_{j=2}^J W_2^2\left(\mu_{j-1}^n(\theta), \mu_j^n(\theta)\right) = \frac{1}{J-1} \sum_{j=2}^J \frac{1}{n} \sum_{i=1}^n \left[Z_{(i)j-1}(\theta) - Z_{(i)j}(\theta)\right]^2.$$

Finally we consider the M-estimator

$$\widehat{\theta^n} \in \arg\min_{\theta \in \Theta} M_n(\theta)$$

for the deformation parameters which permits to define an estimator $\widehat{\mu}_n$ of the density μ.

We present consistency results for these estimators, which are obtained mainly under assumptions of regularity about deformation functions. The proofs follow the guidelines of M-estimation theory and remain valid under weak assumptions on the structural distribution μ.

In a second time, we present a result of convergence in distribution for the estimator of the deformation parameter. Its proof is based on a Delta-Method and requires stronger assumptions about the distributions.

OCS26 Resampling Nonstat T.S.

Resampling Methods for Nonstationary Time Series

JACEK LESKOW*,†

*Institute of Mathematics, Cracow University of Technology, Cracow, Poland

†email: jleskow@pk.edu.pl 268:leskow.tex,session:OCS26

One of the fundamental problems in dealing with nonstationary time series is to establish the statistical inference results for the mean and autocovariance. The classical results, based on central limit theorem have only a limited use, since usually the asymptotic gaussian law has a very complicated autocovariance structure. Since at least ten year in such situation one can resort to resampling techniques such as bootstrap and subsampling. The purpose of the talk will be to present recent results concerning consistency of subsampling and bootstrap for time series with periodically varying mean and autocovariance. The focus will be both on theory and applications. From methodological point of view it will be shown how asymptotic independence assumptions are influencing the consistency of proposed resampling procedures. The applications of such results are wide and range from analysis of vibromechanical signals through biological signals and financial time series.

CS19D Lim. Thms. Processes

Asymptotics and Bootstrap for Degenerate von Mises Statistics under Ergodicity

ANNE LEUCHT*,‡, MICHAEL H. NEUMANN†

*Universität Mannheim, Germany, †Friedrich-Schiller-Universität Jena, Germany

‡email: anne.leucht@uni-mannheim.de 269:Anne_Leucht.tex,session:CS19D

Von Mises- (V-) and related U-statistics play an important role in mathematical statistics. In the case of hypothesis testing, major interest is on degenerate statistics of this type since numerous test quantities can be reformulated as or approximated by statistics of this type under the null hypothesis. Well-known examples are the Cramér-von Mises and the χ^2 statistics.

In the case of i.i.d. random variables the limit distribution can be derived invoking a spectral decomposition of the kernel if the latter is squared integrable. This method has been adopted for mixing and associated random variables, respectively. However, in the dependent case this approach requires some care. Most of the results in the literature have been derived under restrictive assumptions on the associated eigenvalues and eigenfunctions; see e.g. [1] and [2]. However, their validity is quite difficult or even impossible to verify for many concrete examples in statistical hypothesis testing. In this talk, new limit theorems for degenerate U- and V-statistics under ergodicity are presented. Here, we avoid any of these high-level assumptions which is achieved by a restriction to positive semidefinite kernels.

The asymptotic distributions cannot be used directly since they depend on certain parameters, which in turn depend on the underlying situation in a complicated way. Therefore, problems arise as soon as critical values for test statistics of von Mises-type have to be determined. The bootstrap

offers a convenient way to circumvent these problems. We derive consistency of general bootstrap methods for these statistics again under easily verifiable assumptions.

The results are then applied to construct goodness-of-fit tests for GARCH processes and for Poisson count processes.

Acknowledgment. This research was funded by the German Research Foundation DFG, projects NE 606/2-1 and NE 606/2-2.

References

[1] Dewan, I. and Prakasa Rao, B. L. S. (2001). Asymptotic normality for U-statistics of associated random variables. *J. Statist. Plann. Inference* **97**, 201-225.

[2] Huang, W. and Zhang, L.-X. (2005). Asymptotic normality for U-statistics of negatively associated random variables. *Statist. Probab. Lett.* **76**, 1125-1131.

Estimation of Inhibitory Response Latency

POSTER
Poster

MARIE LEVAKOVA[*,‡], PETR LANSKY[†]

[*]Faculty of Science, Masaryk University, Brno, Czech Republic,
[†]Institute of Physiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[‡]email: xlevakov@math.muni.cz

270:MarieLevakova_poster.tex,session:POSTER

Response latency is a frequently studied aspect of neural spiking activity. We define it, in agreement with usual approach, as the time period from the stimulus onset to the change in the neural firing rate evoked by the stimulation. Many methods of latency estimation have been proposed so far, however, they have been mostly based on the assumption of the rate increase (e.g. in [1], [2]). Despite some specifics of the rate decrease response, which cause that methods of latency estimation may fail, the latency for this type has not been investigated.

Two types of models of a spike train are introduced. Models of the first type assume a constant latency across trials, whereas in models of the second type the latency is considered to be a random variable varying across trials. In the second case, statistical properties are evaluated.

All the presented estimation methods are based on observations of the time from the stimulus onset to the occurrence of the first spike after the stimulus (forward recurrence time) in n independent trials (the same approach is used in [3]). The mean, the variance, the probability density function of the forward recurrence time and its Laplace transform are derived. Then, either standard estimation methods are applied or some alternatives are proposed. Namely, a method based on the Laplace transform of the probability density function of the forward recurrence time and a semiparametric comparison of the theoretical cumulative distribution function of the forward recurrence time derived for the model in absence of stimulation with its empirical counterpart obtained from data.

Acknowledgment. This research was supported by Grant Agency CR, grant No.: P103/11/0282.

References

[1] Baker, S. N., Gerstein, G. L., 2001: Determination of Response Latency and Its Application to Normalization of Cross-Correlation Measures, *Neural Computation*, **13**(**6**), 1351 - 1377.

[2] Friedman, H. S., Priebe, C. E., 1998: Estimating stimulus response latency, *Journal of Neuroscience Methods*, **83**, 185 - 194.

[3] Tamborrino, M., Ditlevsen, S., Lansky, P., 2012: Identification of noisy response latency. *Physical Review E*, **86**, 021128.

IS10 High-Dim. Inference

Mixed and Covariate-Dependent Graphical Models

ELIZAVETA LEVINA*

*University of Michigan, Ann Arbor, Michigan, USA

†email: elevina@umich.edu 271:LizaLevina.tex,session:IS10

Graphical models are a popular tool for understanding dependency structure of multivariate data, and sparse graphical models can provide an informative and interpretative summary of high-dimensional data. The commonly used graphical models are usually one of two types: Gaussian graphical models (for continuous data) and the Ising models or Markov networks (for binary and discrete data). However, in practice both types of variables are frequently present in the same dataset, creating the need for mixed graphical models. Some models for these were developed in the earlier literature, but none of them scale to high dimensions. We propose a novel graphical model for mixed data, which is simple enough to be suitable for high-dimensional data, yet flexible enough to represent all possible graph structures, and develop a computationally efficient algorithm for fitting the model. We will also discuss another extension of the graphical model that allows the graph to depend on additional covariates, and apply it to data on genetic instability in tumor samples.

CS6A Funct. Est., Kernel Meth.

Estimation in Semiparametric Single-index Model with Nonignorable Missing Data

BO LI*,†

*Tsinghua University, Beijing, China

†email: libo@sem.tsinghua.edu.cn 272:BoLI.tex,session:CS6A

Consider a single-index model

$$y = g(x\beta) + \varepsilon \tag{1}$$

where the covariates x is a d-dimensional vector which can always be observed whereas the response y is subject to missingness. Both the Euclidean parameter β and the univariate function g are unknown. For the purpose of identification, we assume the first component of β is 1 WLOG. Let r be the response indicator for y, where $r = 1$ if y is observed and $r = 0$ otherwise. We denote $\pi(x, y) = P(r = 1|x, y)$. Ignorable missing mechanism assumes the probability $\pi(x, y)$ to be independent of y, which can not be tenable in many practical circumstances. In this paper, we follow Kim and Yu (2011) to consider an exponential tilting model for the missing mechanism. We adopt a similar semiparametric nonignorable missing mechanism assumption

$$\pi(x, y) = P(r = 1|x, y) = \frac{\exp(h(x\beta) + \phi y)}{1 + \exp(h(x\beta) + \phi y)} \tag{2}$$

where h is an unknown function as well. Namely, we assume both the regression equation and the missing mechanism depend on x through the same 1-d index $x\beta$. Our model formulation can be viewed as a multivarite extension of Kim and Yu (2011), while we use linear index to achieve dimension reduction. When $\phi = 0$, the model reduces to the typical ignorable missing response case. Our aims are to estimate the parameters β, g and the unconditional mean of y, $\theta = E(y)$.

Let $(x_i, y_i, r_i), i = 1, \cdots, n$ be i.i.d observations, and denote $\gamma = -\phi$ for convenience. We assume γ is known a priori (We'll use validation sample to obtain a preliminary estimator for it in practice, see e.g. (23) in Kim and Yu (2011)). We first follow (10) in Kim and Yu (2011) to define $\hat{E}(y|x\beta, r = 0) = \sum_{k=1}^{n} w_{k0}(x\beta, x_k\beta)y_k$, where $w_{k0}(x\beta, x_k\beta) = \frac{r_k K_h(x\beta, x_k\beta)exp(\gamma y_k)}{\sum_{j=1}^{n} r_i K_h(x\beta, x_j\beta)exp(\gamma y_j)}$ for some kernel function $K_h(\cdot, \cdot)$.

Then we can impute $y_i^*(\beta) = r_i y_i + (1-r_i)\hat{E}(y|x_i\beta, r=0)$, which allows us to obtain a kernel estimator for $g(x\beta) = E(y|x\beta)$: $\hat{g}(x\beta) = \hat{E}(y|x\beta) = \sum_{i=1}^n w_{i1}(x\beta, x_i\beta)y_i^*(\beta)$, with $w_{i1}(t) = \frac{K_h(x\beta, x_i\beta)}{\sum_{j=1}^n K_h(x\beta, x_j\beta)}$. The estimator $\hat{\beta}$ can then be estimated by minimizing

$$\sum_{i=1}^{n}[y_i^*(\beta) - \hat{g}(x_i\beta)]^2 \tag{3}$$

With the estimator $\hat{\beta}$, we can obtain a feasible imputed y, $y_i^{**} = r_i y_i + (1-r_i)\hat{E}(y|x_i\hat{\beta}, r=0)$. The resulting estimator of θ can thus be defined as $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n y_i^{**} = \frac{1}{n}\sum_{i=1}^n \{r_i y_i + (1-r_i)\hat{E}(y|x_i\hat{\beta}, r=0)\}$.

The asymptotic properties of our estimator are established and corroborated by our extensive numerical studies. In the numerical studies, we use cross-validation based on the empirical loss defined by (3) to choose the bandwidth h.

References

[Kim and Yu (2011)] Kim, J.K. and Yu, C., 2011: A Semiparametric Estimation of Mean Functionals With Non-ignorable Missing Data. , *JASA*, **106**, 157 - 165.

Functional Data Classification via Covariate Adjusted Subspace Projection

OCS20 H-D Longitudinal Data

PAI-LING LI*,†

*Department of Statistics, Tamkang University, Taiwan
†email: plli@stat.tku.edu.tw

273:Pai-Ling_Li.tex,session:OCS20

A covariate adjusted subspace projected functional data classification (SPFC) method is proposed for curves or functional data classification with accommodating additional covariate information. Based on the framework of subspace projected functional data clustering, curves of each cluster are embedded in the cluster subspace spanned by a mean function and eigenfunctions of the covariance kernel. We assume that the mean function may depend on covariates, and curves of each cluster are represented by the covariate adjusted functional principal components analysis (FPCA) model or covariate adjusted Karhunen-Loève expansion. Under the assumption that all the groups have different mean functions and eigenspaces, an observed curve is classified into the best predicted class by minimizing the distance between the observed curve and predicted functions via subspace projection among all clusters based on the covariate adjusted FPCA model. The proposed covariate adjusted SPFC method that accommodates additional information of other covariates is advantageous to improving the classification error rate. Numerical performance of the proposed method is examined by simulation studies, with an application to a data example.

Acknowledgment. This research was supported in part by the grant from National Science Council (NSC 101-2118-M-032-004).

POSTER Poster

A Bayesian Non-Parametric Approach for Mapping Dynamic Quantitative Traits

ZITONG LI[*,§], MIKKO J. SILLANPÄÄ[†,‡]

[*]University of Helsinki, Helsinki, Finland,
[†]University of Oulu, Oulu, Finland,
[‡]Biocenter, Oulu, Finland
[§]email: zitong.li@helsinki.fi 274:Zitong.tex,session:POSTER

Dynamic or longitudinal quantitative traits are those containing the repeated phenotype measurements over time. In quantitative genetics, the multivariate varying-coefficient linear regression model or so called the functional mapping model

$$y_i(t_r) = \beta_0(t_r) + \sum_{j=1}^{p} x_{ij}\beta_j(t_r) + e_i(t_r),$$

for individuals $i = 1, \ldots, n$, and time points $t_r = t_1, \ldots, t_k$ can be used to detect the association between the dynamic phenotype data (responses $y_i(t_r)$) and the biological markers (covariates x_{ij}) for $j = 1, \ldots, p$. The time dependent coefficients $\beta_j(t_r)$ for marker j are often modeled as a smoothing function or curve in order to introduce the smoothing into the model.

Because of the recent development of modern high-thoughput genotyping and phenotyping techniques, big dimensional data with either a large number of markers or a large number of time points are often produced, which is challenging from the both statistical analysis and computation point of view. Motivated by these challengings, we propose an efficient Bayesian non-parametric functional mapping method on the basis of the above mentioned multivariate regression model. From the modelling point of view, the coefficients $\beta_j(t_r)$ of marker j are modeled by B-splines, and a random walk penalty prior is used in order to avoid the overfitting problem. In addition, we also assume a first order autoregressive (AR(1)) covariance in the residual error terms $e_i(t_r)$ in order to model the time dependency of the phenotype data due to non-genetic (*i.e.*, environmental) factors. From the computational perspective, we use a fast deterministic approximation algorithm called! Variational Bayes (VB) for parameter estimation. VB does not only provide posterior mean and uncertainty estimation (*i.e.*, standard error) for each parameter involved in the model, but also a lower bound estimate to the marginal likelihood, which can be used to guide variable selection. Based on the lower bound estimate, a pursuit matching like algorithm is then used to perform variable selection - search a best subset of markers which may highly associate with the dynamic trait among large data panels. We demonstrate the efficiency of our methods on both real and simulated data sets.

CS6B Funct. Est., Regression

A New Flexible Nonparametric Estimator For Regression Functions

ECKHARD LIEBSCHER[*,†]

[*]University of Applied Sciences Merseburg, Merseburg, Germany
[†]email: eckhard.liebscher@hs-merseburg.de 275:Eckhard_Liebscher.tex,session:CS6B

The aim of the talk is to discuss small sample and asymptotic properties of a new nonparametric estimator for the regression function in a fixed-design regression model. Let us consider the model

$$Y_{ni} = r(x_{ni}) + \varepsilon_i \quad (i = 1, \ldots, n),$$

where $0 \leq x_{n1} < \ldots < x_{ni} < x_{n,i+1} < \ldots < x_{nn} \leq 1$ are the nonrandom design points, $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with $\boldsymbol{E}(\varepsilon_i) = 0$, $\boldsymbol{Var}(\varepsilon_i) = \sigma^2$. We assume that r is smooth and has at least three continuous derivatives. In a first step we consider the following linear estimator for r:

$$\hat{r}_n(x) = \sum_{j=l}^{m} w_{nj}(x)\, Y_{nj}$$

where $l, m \in \{1, \ldots, n\}$ such that $x_{l-1} < x - b \leq x_l, x_m < x + b \leq x_{m+1}$, $x_0 = -\infty, x_{n+1} = +\infty$. Here b is the bandwidth, and the $w_{nj}(x)$'s are the weights. In the second step an estimator for the mean square error is established and then minimised w.r.t. to the weights w_{nj} depending on x. The optimised weights are plugged in the formula of $\hat{r}$ which leads to the estimator. Contrary to the usual approach, no kernel structure is assumed for w_{nj}. Our approach is similar to the so-called direct weight optimisation approach (Sacks and Ylvisaker (1978), Roll, Nazin and Ljung (2005)) but we use an estimator of the mean square error for the optimisation.

We have proven rates of almost sure convergence for the regression estimator. The results are provided in the talk. The new estimator is compared with usual estimators such as the Gasser-Müller estimator and the local linear one. An interesting feature of the performance of the new estimator is that boundary problems do not occur because of automatic correction.

In the last part of the talk results of a comprehensive simulation study are presented. It turns out that the new estimator has a significantly smaller average mean square error in the considered situations compared with the Gasser-Müller estimator and the local linear one.

Optimal Outlier Robust Estimation for Normal Populations

CS12B Bayesian computing

VOLKMAR LIEBSCHER[*,†]

[*]Ernst Moritz Arndt University Greifswald, Institute of Mathematics and Computer Science, Walther-Rathenau-Str.47, 17487 Greifswald, Germany
[†]email: volkmar.liebscher@uni-greifswald.de 276:VolkmarLiebscher.tex,session:CS12B

We present a decision theoretic framework for the problem to determine good estimates of a normal mean with possible outliers present in a small-size sample. From simulations, it turns out that classical robust estimates like median, Huber- or Hampel-type estimators are at least slightly outperformed by Bayesian W-estimates derived from priors which are (improper) Gaussian mixtures.

The latter turn out to be redescending estimators themselves. This shows the relevance of redescending estimators as well as robustness of certain Bayesian estimators with improper but nonflat prior.

Genotype Copy Number Variations using Gaussian Mixture Models

OCS20 H-D Longitudinal Data

CHANG-YUN LIN[*,‡], YUNGTAI LO[†], KENNY Q. YE[†]

[*]Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taiwan,
[†]Albert Einstein College of Medicine
[‡]email: chlin6@nchu.edu.tw 277:Chang-Yun_Lin.tex,session:OCS20

Copy number variations (CNVs) are important in the disease association studies and are usually targeted by most recent microarray platforms developed for GWAS studies. However, the probes targeting the same CNV regions could vary greatly in performance, with some of the probes carrying little information more than pure noise. In this paper, we investigate how to best combine

measurements of multiple probes to estimate copy numbers of individuals under the framework of Gaussian mixture model (GMM). First we show that under two regularity conditions and assume all the parameters except the mixing proportions are known, optimal weights can be obtained so that the univariate GMM based on the weighted average gives the exactly the same classification as the multivariate GMM does. We then developed an algorithm that iteratively estimates the parameters and obtains the optimal weights, and uses them for classification. The algorithm performs well on simulation data and two sets of real data, which shows clear advantage over classification based on the equal weighted average.

References

[Korn et al. (2008)] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. 2008: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, *Nature Genetics*, **40(10)**, 1253-1260.

CS33A Longitudin Data

A Goodness-of-Fit Test of Cumulative Logit Random-Effect Models for Longitudinal Ordinal Responses

KUO-CHIN LIN*,†

*Tainan University of Technology, Tainan, Taiwan
†email: t20053@mail.tut.edu.tw 278:KUOCHINLIN.tex,session:CS33A

Longitudinal studies involving categorical responses are widely applied in a variety of fields, and are often fitted by generalized linear mixed models (GLMMs) as well as generalized estimating equations (GEE) approach. GLMMs extend generalized linear models (GLMs) and the heterogeneity between subjects can be taken into account through incorporation of random-effects that follow some random distributions and are routinely assumed to be normally distributed. The model fitting algorithm for GLMMs includes the penalized quasi-likelihood (PQL) and Gauss-Hermite quadrature approaches, which can be conducted by the commonly used packages: glmmPQL, glmmML and SarebeR in R software. The familiar tactics for analyzing ordinal responses are proportional odds models, adjacent category models and continuation ratio models. Among of them, the cumulative logit model with proportional odds assumption is the most popular approach for the analysis of ordinal data.

Suppose that a longitudinal study consists of ordinal responses with C categories and p-dimensional covariate vectors, $(Y_{ij}, \mathbf{X}_{ij})$, where Y_{ij} represents the jth response for subject i, and the vector of covariates $\mathbf{X}_{ij}$ can be discrete or continuous. Under the assumption of identical odds ratios across the $C-1$ cutoffs, a generalized linear mixed model with logit link function for the conditional cumulative probabilities, $\eta_{ij}^{(k)} = P(Y_{ij} \le k \,|\, \mathbf{b}_i)$, is given by $\text{logit}[\eta_{ij}^{(k)}] = \lambda_k - [\,\mathbf{X}'_{ij}\,\beta + \mathbf{Z}'_{ij}\,\mathbf{b}_i\,]$, where $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are covariate vectors for the fixed and random effects, respectively, the intercepts $\lambda_1, \ldots, \lambda_{C-1}$ satisfy $\lambda_1 \le \cdots \le \lambda_{C-1}$, and the subject-specified random effects $\mathbf{b}_i$ are often assumed to follow a multivariate normal distribution, $\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{D})$.

The aim of this research work is to develop a goodness-of-fit test for assessing the adequacy of the GLMMs with cumulative logit and proportional odds structure. The construction of the proposed test is based on the unweighted sum of squared residuals using bootstrapping approach to obtain the bootstrapping residuals and to estimate the marginal probability. The power of the proposed test is evaluated by simulation studies. Also, a data set from longitudinal study is utilized to illustrate the application of the proposed test.

Acknowledgment. This research was partially supported by the National Science Council, Taiwan, grant No.: NSC 101-2118-M-165-001.

Fast ML Estimation in Mixtures of t-Factor Analyzers via an Efficient ECM Algorithm

OCS20 H-D Longitudinal Data

TSUNG-I LIN*,†

*Institute of Statistics, National Chung Hsing University, Taichung 40724, Taiwan
†email: tilin@nchu.edu.tw 279:TsungILin.tex,session:OCS20

Mixture of t factor analyzers (MtFA; McLachlan et al., 2007) have been shown to be a sound model-based tool for robust clustering of high-dimensional data. This approach, which is deemed to be one of natural parametric extensions with respect to normal-theory models, allows for accommodation of potential noise components, atypical observations or data with longer-than-normal tails. In this paper, we propose an efficient expectation conditional maximization (ECM) algorithm for fast maximum likelihood estimation of MtFA. The proposed algorithm inherits all appealing properties of the ordinary EM algorithm such as its stability and monotonicity, but has a faster convergence rate since its CM steps are governed by a much smaller fraction of missing information. Numerical experiments based on simulated and real data show that the new procedure outperforms the commonly used EM and AECM algorithms (Depmster et al., 1977; Meng and van Dyk 1997) substantially in most of the situations, regardless of how the convergence speed is assessed by the computing time or number of iterations.

Acknowledgment. This work was supported by the National Science Council under grant number NSC101-2118-M-005-006-MY2 of Taiwan.

References

[Dempster et al. (1977)] Dempster, A.P., Laird, N.M and Rubin, D.B., 1977: Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39,** 1–38.

[McLachlan et al. (2007)] McLachlan G.J., Bean R.W. and Jones B.T., 2007: Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Comput Stat Data Anal* **51**, 5327–5338

[Meng and van Dyk (1997)] Meng, X.L. and van Dyk, D., 1997: The EM algorithm – an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59**, 511–567.

Singular Behavior of the Stochastic Heat Equation on a Polygonal Domain

OCS22 Numeric SPDE

FELIX LINDNER*,†

*Technische Universität Dresden, Germany
†email: felix.lindner@tu-dresden.de 280:FelixLindner.tex,session:OCS22

In this talk, a regularity result concerning the solution to the stochastic heat equation on a bounded polygonal domain $\mathcal{O} \subset \mathbb{R}^2$ is presented. We consider semilinear equations with multiplicative noise of the form

$$\mathrm{d}X(t) = \big(\Delta_{\mathcal{O}}^{\mathrm{D}} X(t) + F(X(t))\big)\,\mathrm{d}t + B(X(t))\,\mathrm{d}W(t), \quad X(0) = X_0, \quad t \in [0,T],$$

where $\Delta_{\mathcal{O}}^{\mathrm{D}}$ is the $L_2(\mathcal{O})$-Laplacian with zero Dirichlet boundary condition, $F : L_2(\mathcal{O}) \to L_2(\mathcal{O})$ is a Lipschitz nonlinearity, $B : L_2(\mathcal{O}) \to \mathcal{L}_{\mathrm{HS}}(L_2(\mathcal{O}))$ is a Lipschitz mapping with values in the space of Hilbert-Schmidt operators on $L_2(\mathcal{O})$ and W is a cylindrical Wiener process on $L_2(\mathcal{O})$. Based on a classical result for deterministic elliptic equations by P. Grisvard, it is shown that the solution X can be decomposed into a regular part X_{R} with maximal spatial L_2-Sobolev regularity and a singular part

X_S whose spatial L_2-Sobolev regularity is limited due to the shape of the domain. As a consequence of the time irregularity of the Wiener process W the decomposition takes place in a space of random generalized functions with negative Sobolev smoothness in time.

Our considerations are motivated by the question whether adaptive approximation methods for the solutions to SPDEs pay off in the sense that they admit better convergence rates than nonadaptive (uniform) approximation methods. It is known that the approximation rate that can be achieved by uniform methods is determined by the Sobolev regularity of the exact solution, whereas the approximation rate of adaptive methods is determined by the regularity in a specific scale of Besov spaces. The presented result complements recent results on the Besov regularity of the solutions to SPDEs, cf. the talk by S. Dahlke.

Acknowledgment. This research was supported by the Deutsche Forschungsgemeinschaft (grant SCHI 419/5-1,2).

CS9D Testing Mod. Structure

Weighted Average ML Estimation for Generalized Linear Models

ANTTI LISKI*, ERKKI P. LISKI†,‡

*Tampere University of Technology, Finland,
†University of Tampere, Finland
‡email: Erkki.Liski@uta.fi

281:ErkkiLiski.tex,session:CS9D

In model selection one attempts to use data to find a single "winning" model, according to a given criterion, whereas with model averaging one seeks a smooth compromise across a set of competing models. However, the proper treatment of model uncertainty is still a demanding task. In this paper, our framework is a model where are $p + m$ potential explanatory variables available. There are p focus variables that we want to keep in the model on theoretical or other grounds, and m auxiliary variables that are included in the model only if they are supposed to improve estimation of the coefficients of the focus variables. So, we have the "smallest" model, say the narrow model, having p focus variables and the "biggest" model, say the wide model, having also m additional auxiliary variables. The narrow model is a submodel of the wide model. Further, various submodels are obtained by including only a subset auxiliary variables in the model.

Magnus et al. (2010) introduced a model averaging technique, called weighted average least squares (WALS) estimator, within the linear regression. They claimed that WALS is superior to standard Bayesian model averaging, both theoretically and practically. We consider the problem of averaging across maximum likelihood estimates of competing generalized linear models. Existing model averaging methods usually require estimation of a weight for each candidate model. However, in applications the number of candidate models may be huge which makes such an approach computationally infeasible. Utilizing a connection between shrinkage estimation and model weighting we present a computationally efficient model averaging estimation method which avoids estimation of single model weights. To the best of our knowledge this extension to generalized linear models is knew, although Heumann and Grenke (2010) have considered an extension of WALS to logistic regression. The performance of estimators is displayed also in simulation experiments which utilize a realistic set up based on real data.

References

[Heuman and Grenke (2010)] Heuman, C. and Grenke, M., 2010: An efficient model averaging procedure for logistic regression models using a Bayesian estimator with Laplace prior. In: Kneip, T. and Tutz, G. (Ed.) *Statistical Modelling and Regression Structures*. Festschrift in honour of Ludwig Fahrmeir. Heidelberg, Physica-Verlag.

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

Breaking the Noise Floor in Diffusion MRI, a Bayesian Data Augmentation Approach

CS1A Shape & Image

DARIO GASBARRA*, JIA LIU†,§, JUHA RAILAVO‡

*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland,
†Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland,
‡Helsinki University Central Hospital(HUS), Helsinki, Finland
§email: jia.liu@jyu.fi 282:jia_liu.tex,session:CS1A

By measuring in vivo the microscopic diffusion of water molecules, diffusion Magnetic Resonance Imaging (dMRI) is the only noninvasive technique which can detect diffusion anisotropies, which correspond to nervous fibers in the living brain. Spectral data from the displacement distribution of water molecules is collected by a magnetic resonance scanner. From the statistical point of view, inverting the Fourier transform from sparse and noisy spectral measurements is a non-linear regression problem. Diffusion tensor imaging (DTI) is the simplest modeling approach postulating a Gaussian displacement distribution at each volume element (voxel). Usually diffusion tensor estimation is based on a linearized log-normal regression model that fits dMRI data at low frequency (b-value). This approximation fails to fit the high b-value measurements which contain information about the details of the displacement distribution but have a low signal to noise ratio (SNR). In this paper, we directly work with Rice noise model for the full range of b-values. Using data augmentation to represent the likelihood, the non-linear regression problem is reduced to the framework of generalized linear models. We construct a Bayesian hierarchical model, in order to perform simultaneously estimation and regularization of the tensor field. The Bayesian paradigm is implemented by Markov chain Monte Carlo with Gibbs-Metropolis updates.

References

[Leemans A et al. (2009)] Leemans, A., Jeurissen, B., Sijbers, J., Jones, D.K., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Proc. Intl Soc. Mag. Reson. Med, pp. 3537, Hawaii, USA.

[Mori S. (2007)] Mori S., 2007. Introduction to Diffusion Tensor Imaging. Elsevier Science.

[Pajevic S, Basser P.J. (2003)] Pajevic S, Basser P.J., 2003. Parametric and non-parametric statistical analysis of DT-MRI data. J. Magn. Reson. 161 (1), 1-14.

[Zhu H. et al. (2007)] Zhu H., Zhang H., Ibrahim J.G., Peterson B.S., 2007. Statistical analysis of diffusion tensors in diffusion-weighted Magnetic resonance imaging Data. JASA 102 (480), 1085-1102.

Nonparametric Combination of Multiple Inferences Using Data depth, Bootstrap and Confidence distribution

IS29 Stoch. in Finance

DUNGANG LIU*, REGINA LIU†,‡, MINGE XIE†

*Yale University, Connecticut, USA,
†Rutgers University, New Jersey, USA
‡email: rliu@stat.rutgers.edu 283:ReginaLiu.tex,session:IS29

We apply the concepts of confidence distribution and data depth together with bootstrap to develop a new nonparametric approach for combining inferences from multiple studies for a common hypothesis. A confidence distribution (CD) is a sample-dependent distribution function that can be used to estimate unknown parameters. It can be viewed as a “distribution estimator" of the parameter of interest. Examples of CDs include Efron’s bootstrap distribution and Fraser’s significance

function (also referred to as p-value function). CDs have shown high potential to be effective tools in statistical inference. We discuss a new nonparametric approach to combining the test results from several independent studies for testing a common hypothesis. Specifically, in each study we apply data depth and bootstraps to obtain a p-value function for the common hypothesis. The p-value functions are then combined under the framework of combining confidence distributions. This approach has several advantages. First, it allows us to resample directly from the empirical distribution, rather than from the estimated population distribution satisfying the null constraints. Second, it enables us to obtain test results directly without having to construct an explicit test statistic and then establish or approximate its sampling distribution. The proposed method provides a valid inference approach for a broad class of testing problems involving multiple studies where the parameters of interest can be either finite or infinite dimensional. The method will be illustrated using simulation data and aircraft landing data collected from airlines.

Acknowledgment. This research was partially supported by research grants from National Science Foundation (DMS-1107012, DMS-1007683, DMS-0915139), and National Health Institute (R01- DA016750-09).

IS22 Stat. Neuronal Data

Estimating the Number of Neurons in Multi-Neuronal Spike Trains

WEI-LIEM LOH*

*National University of Singapore, Singapore

284:WeiLiemLoh.tex,session:IS22

A common way of studying the relationship between neural activity and behavior is through the analysis of neuronal spike trains that are recorded using one or more electrodes implanted in the brain. Each spike train typically contains spikes generated by multiple neurons. A natural question that arises is "what is the number of neurons ν generating the spike train?". This talk proposes a method-of-moments technique for estimating ν. This technique estimates the noise nonparametrically using data from the silent region of the spike train and it applies to isolated spikes with a possibly small, but non-negligible, presence of overlapping spikes. Conditions are established in which the resulting estimator for ν is shown to be strongly consistent. To gauge its finite sample performance, the technique is applied to simulated spike trains as well as to actual neuronal spike train data.

CS19E Lim. Thms.

On the Ruin Probability

VLADIMIR LOTOV*,†,‡

*Novosibirsk State University, Novosibirsk, Russia,

†Sobolev Institute of Mathematics, Novosibirsk, Russia

‡email: lotov@math.nsc.ru

285:VladimirLotov.tex,session:CS19E

Let X_1, X_2, ... be i.i.d. random variables, $EX_1 < 0$, and let

$$S_0 = 0, \quad S_n = X_1 + \ldots + X_n, \quad S = \sup_{n \geq 0} S_n.$$

For $a \geq 0$ and $b > 0$, we introduce the first exit time from the interval $(-a, b)$:

$$N = N_{a,b} = \inf\{n \geq 1 : S_n \notin (-a, b)\}.$$

We give new asymptotic representations for the ruin probability $P(S_N \geq b)$ as $b \to \infty$ under various conditions on the rate of decreasing of $P(X_1 \geq t)$, $t \to \infty$, including the case of heavy tails.

The results are based on the well-known asymptotics of $P(S \geq b)$, $b \to \infty$, and the distribution of the excess over negative barrier.

Graph Property Testing

SIL
Spec.
Invited
Lecture

LÁSZLÓ LOVÁSZ*,†

*Eötvös Loránd University, Budapest, Hungary
†email: lovasz@cs.elte.hu

286:Laszlo_Lovasz.tex,session:SIL

For very large networks, a natural method (often the only way) to obtain information about their structure is sampling. A theory of these methods was initiated by computer scientists in the 1990's; recently, this theory has been connected to the theory of graph limits and of distributed algorithms.

There are several similar, but different goals we may want to achieve, which are analogous to basic tasks in classical statistics: we may want to estimate a parameter of the graph (say, the density of triangles, or the density of the largest cut), or we may want to decide whether the graph has a certain property (say, is it planar?). Some new tasks also arise, due to the different kind of structure of the object we are sampling from: we may want to compute some additional structure on top of the given network, like a maximum matching.

In this talk we sketch the theoretical foundations for such sampling algorithms, and show some connections to the theory of graph limits and extremal graph theory.

Ratio-of-Uniforms for Unbounded Distributions

POSTER
Poster

LUCA MARTINO*, DAVID LUENGO†, JOAQUÍN MÍGUEZ*

*Universidad Carlos III de Madrid, Leganés, Spain,
†Universidad Politécnica de Madrid, Madrid, Spain
‡email: luca@tsc.uc3m.es, david.luengo@upm.es, jmiguez@tsc.uc3m.es

287:David_Luengo_GRoU.tex,session:POSTER

In this work (see also [Martino et al. (2012)]), we investigate the relationship among three classical sampling techniques: the inverse-of-density (a.k.a. Khintchine's theorem) [Khintchine (1938)], the transformed rejection sampling (TRS) [Wallace (1976)], and the generalized ratio-of-uniforms (GRoU) [Wakefield et al. (1991)].

Given a monotonic probability density function (PDF), we show that the transformed area obtained using the generalized ratio-of-uniforms method can be found equivalently by applying the transformed rejection approach to the inverse function of the target density. Then we provide an extension of the classical inverse-of-density idea, showing that it is completely equivalent to the GRoU method for monotonic densities.

We also discuss how the previous results can be extended to a generic PDF, dividing its support domain into a collection of intervals where it is either monotonically increasing or decreasing. Finally, we apply these considerations to design different variants of the GRoU method, introducing, for instance, a novel GRoU strategy to handle unbounded target densities.

Acknowledgment. This work has been partly financed by the Spanish government, through the ALCIT (TEC-2012-38800-C03-01), COMPREHENSION (TEC2012-38883-C02-01) and DISSECT (TEC2012-38058-C03-01) projects, as well as the CONSOLIDER-INGENIO 2010 Program (Project CSD2008-00010).

References

[Khintchine (1938)] A. Y. Khintchine, 1938: On unimodal distributions. Izvestiya NauchnoIssledovatelskogo Instituta Matematiki i Mekhaniki, 2:1-7.

[Martino et al. (2012)] L. Martino, D. Luengo and J. Míguez, 2012: On the Generalized Ratio of Uniforms as a Combination of Transformed Rejection and Extended Inverse of Density Sampling. arXiv: 1205.0482.

[Wakefield et al. (1991)] J. C. Wakefield, A. E. Gelfand, and A. F. M. Smith, 1991: Efficient generation of random variates via the ratio-of-uniforms method. Statistics and Computing, 1(2):129-133.

[Wallace (1976)] C.S. Wallace 1976: Transformed rejection generators for gamma and normal pseudo-random variables. Australian Computer Journal, 8:103-105.

IS4 Empirical Proc.

Detecting Positive Correlations in a Multivariate Sample

GÁBOR LUGOSI[*,†]

[*]ICREA and Department of Economics, Pompeu Fabra University
[†]email: gabor.lugosi@upf.edu 288:Lugosi_Gabor.tex,session:IS4

In this joint work with Sébastien Bubeck (Princeton) and Ery Arias-Castro (UCSD), we consider the problem of testing whether a correlation matrix of a multivariate normal population is the identity matrix. We focus on sparse classes of alternatives where only a few entries are nonzero and, in fact, positive. We derive a general lower bound applicable to various classes and study the performance of some near-optimal tests. We pay special attention to computational feasibility and construct near-optimal tests that can be computed efficiently. We apply our results to prove new lower bounds for the clique number of high-dimensional random geometric graphs.

NYA Not Yet Arranged

Semi-parametric Analysis of the Return to Education of China

JI LUO[*,†], XU WENWEN[*]

[*]Zhejiang University of Finance and Economics, Hangzhou, China
[†]email: luoji0409@126.com 289:Ji_Luo.tex,session:NYA

In the paper, Semi-parametric Additive Mincer wage equation model was built on the basis of the expansion studies of the traditional Mincer wage equation, in which the parameters and non-parametric items were estimated by penalized least squares method. The explanatory factors and the applicability of the model were discussed and analyzed. China's overall rate of return on education was estimated through the China Nutrition and Health Survey (CHNS) data, which is equal to 22.25%. It is worth mentioning that this value has a comparative significance in the international, unlike other research results. In order to compare the return to education in different region and gender, we estimated the return to education respectively, according to the region and gender. At last, the paper concluded the differences in the rate of return on education, and gave some conclusions and recommendation.

Acknowledgment. This research was partially supported by the National Natural Science Foundation of China, Grant No.: 11271317; China Statistical Research Project, Grant No.: 2012LY129; Zhejiang Provincial Natural Science Foundation, Grant No.: LY12A01017.

References

[1] Brauw, A., Rozelle, S., 2008: Reconciling the Returns to Education in Rural China, *Review of Development Economics*, **12**(1), 57-71.

[2] Fleisher, B. M., Wang, X., 2005: Returns to Schooling in China under Planning and Reform, *Journal of Comparative Economics*, **33**, 265-277.

[3] Li, H., 2003: Economic Transition and Returns to Education in China, *Economics of Education Review*, **22**, 317-328.

[4] Li, H., Luo, Y., 2004: Reporting Errors, Ability Heterogeneity, and Returns to Schooling in China, *Pacific Economic Review*, **9**(3), 191-207.

[5] Li, H., Liu, P. W., Ma, N., Zhang, J., 2007: Does Education Pay in Urban China? Estimating Returns to Education Using Twins, *The Economic Journal*, **117**, 1504-1520.

[6] MaurerFazio, M. and Dinh, N., 2004: Differential Rewards to and Contributions of Education in Urban China's Segmented Labor Markets, *Pacific Economic Review*, **9**, 173-189.

[7] Wu, X., Xie, Y., 2003: Does the Market Pay Off? Earnings Inequality and Returns to Education in Urban China, *American Sociological Review*, **68**, 425-442.

[8] Yang, D. T., 2005: Determinants of Schooling Returns Durning Transition: Evidence from Chinese Cities, *Journal of Comparative Economics*, **33**, 244-264.

[9] Zhang, J., Zhao Y., 2002: Economic Returns to Schooling in Urban China[D], *Chinese University of Hong Kong*.

Learning from Data, Predicting its Effect

SIL Spec. Invited Lecture

GREG GYURKO*, TERRY LYONS*,†, HAO NI*

*Oxford Man Institute for Quantitative Finance,
Mathematical Institute, University of Oxford
†email: terry.lyons@oxford-man.ox.ac.uk

290:Terry_Lyons.tex,session:SIL

A relatively new area of mathematics, known as rough path theory [1] considers the evolution of controlled systems and identifies in a stratified and quantified way the features in the control that determine the system response. The theory has been very successful at explaining the behaviour of controlled differential systems where the control is highly oscilliatory leading to new thoretical tools and numerical methods. Through the remarkable work [2] of Hairer, it is now having a significant influence on the ways that one thinks about stochastic PDEs and a number of classical problems have been addressed using the new technology he is developing from the original theory.

In this talk we will present an effective application of this apparently abstract and mathematical theory to the non-parametric analysis of time series. I will explain how it provides powerful non-parametric approaches to the recognition of patterns and introduce joint work [4] with Hao Ni that introduces the signature [3] of this series as a basic non-parametric tool for learning and prediction. We illustrate with simple examples of autoregressive type. I will also mention joint work [5] with Greg Gyurko and Mark Kontkowski where this methodolgy underpins effective efforts to identify, analyse, and predict complex features of markets in a totally non-parametric "learn from data" approach.

In modern langauge, we will explain how the rough path theory gives powerful tools that are designed from the bottom up to analyse big sequential data.

Acknowledgment. Supported by EPSRC grant EP/H000100/1, ERC grant 291244, and by the Oxford-Man Institute

References

[1] Lyons, Terry J. "Differential equations driven by rough signals." Revista Matemática Iberoamericana 14, no. 2 (1998): 215-310.

[2] Hairer, Martin. "Rough stochastic pdes." Communications on Pure and Applied Mathematics 64, no. 11 (2011): 1547-1585.

[3] Hambly, Ben, and Terry Lyons. "Uniqueness for the signature of a path of bounded variation and the reduced path group.(English summary)." Ann. of Math.(2) 171, no. 1 (2010): 109-167.

[4] Lyons, Terry, and Ni, Hao. Learning from the past, predicting the statistics of the future, learning an evolving system, preprint.

[5] Gyurko, Greg. Progress report, Slippage-Ito project, Oxford-Man Institute.

CS25B Risk Mgment

Economic Crisis and the Need for Technical Efficiency Analysis

PEDRO MACEDO*,†, MANUEL SCOTTO*

*Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, Portugal

†email: pmacedo@ua.pt 291:PedroMacedo.tex,session:CS25B

Technical efficiency analysis is a fundamental tool to measure the performance of production activity (e.g., industries, universities, hospitals, banks, etc.). Nowadays, a wide range of methodologies to measure technical efficiency are available being the choice of a specific approach always controversial, since different choices lead to different results. Recently, an increasing interest with the state-contingent production frontiers has emerged in the literature. This interest is mainly due to the fact that uncertainty in economics is best interpreted within a state-contingent framework. However, this increasing interest has not yet been reflected in an increase of empirical applications, since empirical models with state-contingent production frontiers are usually ill-posed. In particular, these empirical models are affected by severe collinearity.

In this talk will be discussed new maximum entropy procedures in the estimation of technical efficiency with state-contingent production frontiers under severe empirical conditions. Simulation studies and a real application on wine production in Portugal will be used to illustrate that those maximum entropy estimators are powerful alternatives to the maximum likelihood estimator usually used in the efficiency analysis literature.

Acknowledgment. This work was supported by *FEDER* funds through *COMPETE*–Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

References

[Macedo et al. (2012)] Macedo, P., Silva, E., Scotto, M., 2012: Technical efficiency with state-contingent production frontiers using maximum entropy estimators, *J. Prod. Anal.* (to appear).

CS30A Inf. on Distributions

Goodness-of-Fit Test for The Skew-t Distribution

MOHAMMAD MAHDI MAGHAMI*,†

*Department of Statistics, University of Isfahan, Isfahan, Iran

†email: maghami8@gmail.com 292:Mohammad_Mahdi_Maghami.tex,session:CS30A

In this manuscript goodness-of-fit test is proposed for the Skew-t distribution based on properties of the family of these distributions and the sample correlation coefficient. The critical values for the test can be achieved by Monte Carlo simulation method for several sample sizes and levels of significance. The power of the proposed test can be specified for different sample sizes and considering diverse alternatives.

References

[1] Azzalini A. (1985): A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, Vol. 12, pp. 171-178.

[2] Cook, R. D., Weisberg, S. (1994): *An Introduction to Regression Graphics*, Wiley, New York.

[3] Perez Rodriguez, P., Villasenor Alva, J.A. (2010): On testing the skew normal hypothesis, *Journal of Statistical Planning and Inference*, Vol. 140, pp. 3148-3159.

Nonparametric Tests for Regression Quantiles

IS8 Function Estim.

ENNO MAMMEN[*,§], INGRID VAN KEILEGOM[†], KYUSANG YU[‡]

[*]Department of Economics, University of Mannheim, Germany,
[†]Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium,
[‡]Department of Applied Statistics, Konkuk University, Seoul, Korea
[§]email: emammen@rumms.uni-mannheim.de 293:EnnoMammen.tex,session:IS8

We discuss nonparametric tests for parametric specifications of regression quantiles. The test is based on the comparison of parametric and nonparametric fits of the regression quantiles. The nonparametric fit is a Nadaraya-Watson quantile smoothing estimator.

An asymptotic treatment of the test statistic requires the development of new mathematical arguments. An approach that makes only use of plugging in a Bahadur expansion of the nonparametric estimator is not satisfactory. It requires too strong conditions on the dimension and the choice of the bandwidth.

Our alternative mathematical approach requires the calculation of moments of Bahadur expansions of Nadaraya-Watson quantile regression estimators. This calculation is done by inverting the problem and application of higher order Edgeworth expansions. The moments allow estimation bounds for the accuracy of Bahadur expansions for integrals of kernel quantile estimators.

Another application of our method gives asymptotic results for the estimation of weighted averages of regression quantiles.

Quantitative Central Limit Theorems for Angular Polyspectra

OCS3 Spectral Analysis

DOMENICO MARINUCCI[*,‡], IGOR WIGMAN[†,§]

[*]Department of Mathematics, University of Rome Tor Vergata,
[†]Department of Mathematics, King's College London
email: [‡]marinucc@mat.uniroma2.it, [§]igor.wigman@kcl.ac.uk
294:Marinucci.tex,session:OCS3

We use the Stein-Malliavin approach recently developed by Nourdin and Peccati (2009-2012) to provide quantitative Central Limit Theorems for the angular polyspectra of arbitrary order for spherical random fields. We consider both total variation and Wasserstein distances; for the latter, we provide results also concerning the area of spherical excursion sets.

A Resampling Method to Compare Inter-Industry Financial Ratios

CS25C Stoch. Finance II.

MARCO MAROZZI[*,†]

[*]University of Calabria, Rende (CS), Italy
[†]email: marco.marozzi@unical.it 295:Marozzi.tex,session:CS25C

Multivariate analysis of variance (MANOVA) and multiple discriminant analysis (MDA) are the most commonly used methods for comparing firm financial ratios. They are quite always used for inferential purposes without checking carefully whether underlying assumptions are fulfilled or not in

the data. This question is extremely important because typical financial data do not fulfill MANOVA nor MDA assumptions. Typical financial data are taken from a database of publicly traded firms. All the firms that have no missing data are considered, there is no random sampling. Moreover many financial ratios are highly skewed and heavy tailed. Financial ratios are not normally distributed because most of them are restricted from taking on values below zero but can be very large positive values. If it is not possible to use random samples of firms, a solution is to work within the resampling framework. Our aim is to provide managers, auditors, shareholders, lenders and potential investors with an effective method to compare firm financial ratios. We follow a descriptive point of view and then our method does not require any particular assumptions. More precisely our method does not require random sampling nor normal distribution and it is devised to explicitly consider the possible difference in variances. It is a sort of measure of difference between groups of firms which takes also into account the dependence among the financial ratios.

Let ${}_lX_{ij}$ denote the value of financial ratio l for firm j of group $i = 1, 2$. We say that the two groups are not different so far as ${}_lX$ is concerned if both means $M({}_lX_i)$ and variances $VAR({}_lX_i)$ of ${}_lX$ in the two groups are equal. To grade the difference between groups we compute ${}_lC = {}_lU^2 + {}_lV^2 - 2\rho_l U_l V$, where ${}_lU$ and ${}_lV$ statistics are respectively the standardized sum of squared ranks and squared contrary ranks of the first group and $\rho = corr({}_lU, {}_lV)$. Note that the ${}_lC$ statistic is a combination of ${}_lU$ and ${}_lV$ statistics taking into account their negative correlation ρ. If both means and variances of ${}_lX$ in the two groups are equal then ${}_lC = 0$ and it increases as the difference between groups increases. For comparing the grade of difference of various financial ratios, a resampling method is used to normalize the ${}_lC$ statistic to lay between 0 and 1.

Univariate analysis of financial ratios may be misleading therefore we should combine several financial ratios for a complete picture of the firm. It is very important to underline that the combination procedure of ${}_lC$, $l = 1, ..., L$ is resampling based and devised just to take into account the dependence among the financial ratios.

The method is applied to inter-industry comparison of the financial ratios of Chinese and Japanese publicly traded firms. We consider the following valuation ratios: P/E = price to earnings ratio, P/B = price to book equity ratio, P/S = price to sales ratio, $EV/EBITDA$ = enterprise value to EBITDA ratio, where EBITDA stands for earnings before interest, taxes, depreciation and amortization, EV/C = enterprise value to capital ratio, EV/S = enterprise value to sales ratio. The traditional method to address the problem at hand is MANOVA but several assumptions underlying it are not fulfilled in the data. Therefore, in place of MANOVA it is preferable to use our method. It does not require random sampling because it is a descriptive method, is robust against skewness and heavy tailness and takes explicitly into account the difference in variability as well as in central tendency between groups as well as the dependence relations among financial ratios.

We found that industry sectors of Japanese firms are generally more different than those of Chinese firms. In general, the difference between industry sectors is high. The rankings of financial ratios from the most to the least different one are very similar for Japanese and Chinese firms. The most different and the least different ratio are respectively P/S and P/E for both Chinese and Japanese firms.

Anticipating Linear Stochastic Differential Equations Driven by a Lévy Process

POSTER Poster

JORGE A. LEÓN*, DAVID MÁRQUEZ-CARRERAS†,‡, JOSEP VIVES†

*Departamento de Control Automático, Cinvestav-IPN, México D.F., Mexico,
†Departament de Probabilitat, Lògica i Estadística, Facultat de Matemàtiques, Universitat de Barcelona, Barcelona, Catalunya
‡email: davidmarquez@ub.edu

296:DavidMarquez-Carreras.tex,session:POSTER

In this paper we study the existence of a unique solution for linear stochastic differential equations driven by a Lévy process, where the initial condition and the coefficients are random and not necessarily adapted to the underlying filtration. Towards this end, we extend the method based on Girsanov transformation on Wiener space and developped by Buckdahn to the canonical Lévy space.

Reduction of Chemical Reaction Networks I: Chains of Reactions and Delay Distributions

OCS31 Stoch. Molecular Biol.

TATIANA T. MARQUEZ-LAGO*, ANDRE LEIER†, MANUEL BARRIO‡

*Integrative Systems Biology Unit, Okinawa Institute of Science and Technology,
†Okinawa Institute of Science and Technology,
‡Departamento de Informatica, Universidad de Valladolid

297:TatianaMarquez-Lago.tex,session:OCS31

Accurate modelling of dynamic cellular events requires an adequate description of key chemical reactions. Quite importantly, simulation of such chemical events needs to take place over reasonable time spans, fully describing underlying dynamics of interest. In order to achieve this, computational costs have to be kept as low as possible. In fact, more often than not, it is the associated computational costs which actually limit our capabilities of representing complex cellular behaviour. Thus, efficient simulation strategies and model reduction methodologies become essential.

In this talk, I will present a novel methodology aimed at representing chains of chemical reactions by much simpler, reduced models [1]. The abridgement is achieved by generation of model-specific delay distribution functions, consecutively fed to a delay stochastic simulation algorithm. In this first part of our composite presentation in model reduction methodologies, I will show how such delay distributions can be analytically described whenever the system is solely composed of consecutive first-order reactions, yielding an exact reduction. More complicated scenarios will be covered in the second part of our composite presentation (Reduction of Chemical Reaction Networks II). Our model reduction methodology yields significantly lower computational costs while retaining accuracy. Quite naturally, computational costs increase alongside network size and separation of time scales. Thus, we expect our model reduction methodologies to significantly decrease computational costs in these instances.

We anticipate the use of delays in model reduction will greatly alleviate some of the current restrictions in simulating large sets of chemical reactions, largely applicable in pharmaceutical and biological research.

References

[1] Manuel Barrio, Andre Leier and Tatiana Marquez-Lago (2013). Reduction of chemical reaction networks through delay distributions. J Chem Phys 138, 104114.

POSTER
Poster

Bayesian Analysis of Misclassified Polychotomous Response Data

JACINTO MARTÍN[*,†], LIZBETH NARANJO[*], CARLOS J. PÉREZ[*]

[*]University of Extremadura, Spain
[†]email: jrmartin@unex.es 298:JacintoMartin.tex,session:POSTER

All the models for polychotomous outcomes (ordinal or nominal) or correlated binary outcomes depending on the covariates focus on the estimation of the regression parameters. The conventional approach to modeling these data is to assume that the outcomes are measured without error. However, in many important applications, especially related to epidemiology and toxicological studies, measurement error problems may arise. In these contexts, additional parameters are necessary to correct the bias yielded by the use of misclassified data. If the misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non optimal decision making. Therefore, statistical models should address misclassification.

Bayesian analysis for modeling polychotomous response data (ordinal or nominal) or correlated binary response data that is subject to misclassification is considered. Misclassification in the categorical response is considered through the development of Bayesian regression models for ordered, nominal or correlated binary categories in the response variable. Probit link has been used to model the data, however, the misclassification models are defined in such way that it is possible to use other link functions. The computational difficulties have been avoided by using data augmentation frameworks. The idea of using data augmentation is exploited to derive efficient Markov chain Monte Carlo methods. A simulation based example illustrates the model performance when comparing with standard methods that do not consider misclassification.

Acknowledgment. This research has been partly funded by the European Union (FEDER funds), the Spanish Government Board (National Research Projects) and the Regional Government Board (Junta de Extremadura), by means of grants reference MTM2011-28983-C03-02 and GR10110.

References

[1] Albert, P. S., 2009: Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard, *Statistics in Medicine*, **28**, 780-797.

[2] Albert, P. S., Hunsberger, S. A., Biro, F. M., 1997: Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation, *Journal of the American Statistical Association*, **92**(440), 1304-1311.

[3] McGlothlin, A., Stamey, J. D., Seaman, J. W. Jr., 2008: Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach, *Biometrical Journal*, **50**(1), 123-134.

[4] Mwalili, S. M., Lesaffre, E., Declerck, D., 2005: A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study, *Applied Statistics*, **54**(1), 77-93.

[5] Paulino, C. D., Silva, G., Achcar, J. A., 2005: Bayesian analysis of correlated misclassified binary data, *Computational Statistics and Data Analysis*, **49**, 1120-1131.

POSTER
Poster

Modelling Misclassified Polychotomous Response Data: A Bayesian approach

JACINTO MARTÍN[*], LIZBETH NARANJO, CARLOS J. PÉREZ

University of Extremadura, Spain
[*]email: jrmartin@unex.es 299:JacintoMartn.tex,session:POSTER

Models for polychotomous outcomes (ordinal or nominal) or correlated binary outcomes de-

pending on the covariates focus on the estimation of the regression parameters. The conventional approach to model these data assumes that the responses are measured without error. However, in many important applications, especially those related to epidemiological and toxicological studies, classification errors may arise. In these contexts, additional parameters are necessary to correct the bias yielded by the use of misclassified data. If the misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non optimal decision making. Therefore, statistical models should address misclassification.

Bayesian analyses for modeling polychotomous response data are considered when data are subject to misclassification. Specifically, misclassification in the categorical response is considered through the development of Bayesian regression models for ordered, nominal and correlated binary categories in the response variable. Probit link has been used to model the data. However, the misclassification models have been defined so that other link functions can be used. The computational difficulties have been avoided by using data augmentation frameworks. The idea of using data augmentation is exploited to derive efficient Markov chain Monte Carlo methods. A simulation-based example illustrates the model performance when comparing with standard methods that do not consider misclassification.

Acknowledgment. This research has been partly funded by the European Union (FEDER funds), the Spanish Government Board (National Research Projects) and the Regional Government Board (Junta de Extremadura), by means of grants reference MTM2011-28983-C03-02 and GR10110.

References

[1] Albert, P. S., 2009: Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard, *Statistics in Medicine*, **28**, 780-797.

[2] Albert, P. S., Hunsberger, S. A., Biro, F. M., 1997: Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation, *Journal of the American Statistical Association*, **92**(440), 1304-1311.

[3] McGlothlin, A., Stamey, J. D., Seaman, J. W. Jr., 2008: Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach, *Biometrical Journal*, **50**(1), 123-134.

[4] Mwalili, S. M., Lesaffre, E., Declerck, D., 2005: A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study, *Applied Statistics*, **54**(1), 77-93.

[5] Paulino, C. D., Silva, G., Achcar, J. A., 2005: Bayesian analysis of correlated misclassified binary data, *Computational Statistics and Data Analysis*, **49**, 1120-1131.

Comparison of Stochastic Claims Reserving Models in Insurance

CS25B Risk Mgment

MIKLÓS ARATÓ*, MIKLÓS MÁLYUSZ*, LÁSZLÓ MARTINEK*,†

*Eötvös Loránd University, Budapest, Hungary

†email: martinek@cs.elte.hu 300:LaszloMartinek.tex,session:CS25B

The appropriate estimation of incurred but not reported (IBNR) claims is crucial to preserve the solvency of insurance institutions, especially in casualty and property insurance. A general assumption is that claims related to policyholders, and occur in accounting year i, are reported to the insurance company in the subsequent years, sometimes many years later. Thus, reserves have to be made to cover these arising expenses.

Data are represented by so called run-off triangles. These are $(n+1)\times(n+1)$ matrices, where element X_{ij} represents the claim amount incurred in year i, and reported with a delay of j years, or may indicate aggregate values. Elements for indices $i+j>n+1$ are unknown, have to be predicted. If X_{ij} denotes an incremental value in the triangle, then the IBNR reserve is $\sum_{1\le i,j,\le n+1} X_{ij} - \sum_{i+j\le n+1} X_{ij}$.

The topic has a wide actuarial literature, describing development models and evaluation techniques. Some important publications are [England, Verrall (2002), Merz, Wuthrich (2008)]. They contain deterministic estimation models, like the classical chain-ladder method, and also nondeterministic models, bootstrap methods, for instance.

The cardinal aim of our present work is the comparison of appropriateness of several stochastic estimation methods, supposing different distributional development models. On the one hand, in most cases a Monte Carlo type evaluation is needed, because the explicit description of ultimate claim value is not possible. For instance, if this value is the sum of random variables from log-normal distribution. On the other hand, the comparison is based on score values for empirical distribution functions, see [Gneiting, Tillman (2007)], for instance. We expect this attitude to be more informative than the classical mean squared error of prediction measure.

References

[Merz, Wuthrich (2008)] Merz, M., Wuthrich, M. V., 2008: Stochastic Claims Reserving Methods in Insurance, *Wiley*.

[England, Verrall (2002)] England, P. D., Verrall, R. J., 2002: Stochastic Claims Reserving in General Insurance (with discussion), *British Actuarial Journal*, **8**, 443-544.

[Gneiting, Tillman (2007)] Gneiting, T., Raftery, A. E. (2007): Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, **102**, 359-378.

POSTER
Poster

Nonparametric Mixture Models Based on Weight-Ordered Random Probability Measures

ASAEL FABIAN MARTÍNEZ*,†, RAMSÉS H. MENA*

*IIMAS–UNAM, Mexico

†email: fabian_023@yahoo.com.mx 301:AsaelFabianMartinez.tex,session:POSTER

In the nonparametric Bayesian approach, mixture models have been used in density estimation and classification problems. One construction for this class of models is given by

$$f(y) = \int_{\Theta} K(y;\theta) P(\mathrm{d}\theta),$$

where K is a density function for each θ and P is a discrete random probability measure, i.e. of the type

$$P(\mathrm{d}\theta) = \sum_{j\geq 1} w_j \delta_{\theta_j}(\mathrm{d}\theta),$$

where $\{\theta_j\}_{j\geq 1}$ are iid random variables with non atomic distribution P_0 and weights $\{w_j\}_{j\geq 1}$ are such that $0 < w_j < 1$, for all j, and constrained to sum one a.s.

First construction for weights $\{w_j\}_{j\geq 1}$ considered in this work is derived from the called stick–breaking representation. Here, weights take the form

$$w_1 = v_1, \qquad w_j = v_j \prod_{k<j} (1 - v_k) \quad j > 1,$$

where the $\{v_j\}_{j\geq 1}$ are independent Beta random variables with parameters (a_i, b_i).

In the second construction, weights are decreasingly ordered as follows

$$w_j = \sum_{k\geq j} \frac{p_k}{k} \quad j \geq 1,$$

where $\{p_k\}_{k\geq 1}$ are the probabilities for a certain distribution supported on the set of positive integers.

In this work we evaluate the performance of mixture models, obtained by the choice of specific cases within each construction, in density estimation problems.

A Survey of the Theory of the Petersburg Game

StPburgL St. Petersburg Mem. Lect.

ANDERS MARTIN-LÖF[*,†]

[*]Stockholm University, Stockholm, Sweden
[†]email: andersml@math.su.se

302:AndersMartin-Lof.tex,session:StPburgL

This year we can celebrate the 300th anniversary of the invention of the Petersburg game by Daniel Bernoulli, where the gain is 2, 4, 8,... with probability 1/2, 1/4, 1/8,... The fact that the expected gain is infinite has given rise to numerous discussions of the relevance of this quantity compared to some more reasonable utility. The recent studies of the game take the attitude that the gain is relevant and studies the total gain in a large number of games and tries to find interesting limit theorems for the distribution of this quantity. It turns out that an interesting class of limit distributions is obtained named semistable by Lévy. Luckily quite a simple asymptotic expression for the probability of a large gain can be found. Also variations of the game where interest on the capital is taken into account can be studied in a similar way.

Improved Adaptive Rejection Metropolis Sampling

POSTER Poster

LUCA MARTINO[*], JESSE READ[*], DAVID LUENGO[†]

[*]Universidad Carlos III de Madrid, Leganés, Spain,
[†]Universidad Politécnica de Madrid, Madrid, Spain
email: luca@tsc.uc3m.es, jesse@tsc.uc3m.es, david.luengo@upm.es

303:Luca_Martino_A2RMS.tex,session:POSTER

Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings (MH) algorithm, are widely used for Bayesian inference. One of the most important challenges for any MCMC method is speeding up the convergence of the Markov chain, which depends crucially on a suitable choice of the proposal density.

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks et al. (1995)] is a well-known MH scheme that generates samples from one-dimensional target densities by making use of adaptive piecewise linear proposals constructed using support points taken from rejected samples. The ARMS algorithm is often applied within a Gibbs sampler, where the reduction of the burn-in period is crucial.

In this work, we point out a critical drawback in the adaptive structure of ARMS and propose an alternative scheme (A2RMS) in order to speed up the convergence of the chain to the target distribution. With the A2RMS algorithm, the sequence of proposals densities converges to the true shape of the target, allowing us to perform virtually exact sampling, since the correlation among the samples vanishes quickly to zero. Moreover, at the same time, the computational cost is kept bounded.

Since the novel scheme also allows us to simplify the construction of the sequence of proposal distributions w.r.t. to the technique described in [Gilks et al. (1995)], then we also provide different simplified procedures to build the proposal. Numerical results show that the new algorithm outperforms the standard ARMS and other techniques in terms of estimation accuracy and reduced correlation among the generated samples.

Acknowledgment. This work has been partly financed by the Spanish government, through the ALCIT (TEC-2012-38800-C03-01), COMPREHENSION (TEC2012-38883-C02-01) and DISSECT (TEC2012-38058-C03-01) projects, as well as the CONSOLIDER-INGENIO 2010 Program (Project CSD2008-00010). The authors would

also like to thank Roberto Casarin (Universitá Ca' Foscari di Venezia), Fabrizio Leisen and Joaquín Míguez (Universidad Carlos III de Madrid) for many useful comments and discussions about the ARMS technique.

References

[Gilks et al. (1995)] Gilks, W. R., Best, N. G., and Tan, K. K. C., 1995: Adaptive Rejection Metropolis Sampling within GibbsSampling. *Applied Statistics*, **44**(4), pages 455-472.

CS19A Lim. Thms. Heavy Tails

Invariance Principles for a Multivariate Student Process in the Generalized Domain of Attraction of the Multivariate Normal Law

YULIYA V. MARTSYNYUK*,†

*University of Manitoba, Winnipeg, Canada

†email: yuliya_martsynyuk@umanitoba.ca 304:YuliyaMartsynyuk.tex,session:CS19A

For a d-variate Student process based on independent copies of a random vector X, with trajectories in the space of $\mathbb{R}^d$-valued cadlag functions on $[0,1]$, our main result establishes a uniform Euclidean norm approximation in probability with a sequence of appropriate d-variate processes with Wiener process components, assuming that X is in the generalized domain of attraction of the multivariate normal law (GDAN) and some additional conditions. As a consequence, a functional central limit theorem is also concluded for the Student process. The condition $X \in$ GDAN for these invariance principles is shown to be not only sufficient, but also necessary.

Acknowledgment. This research was supported by the NSERC Canada Individual Discovery Grant and University of Manitoba start-up funds of Yu.V. Martsynyuk.

CS30A Inf. on Distributions

Cramér-von Mises Test for Gauss Processes

GENNADY MARTYNOV*,†,‡

*Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow,

†National Research University Higher School of Economics, Moscow, Russia

‡email: martynov@iitp.ru 305:GennadyMartynov.tex,session:CS30A

One of the problems in the theory of the goodness-of-fit tests is the problem to test if an observed random process $S(t)$ on $[0,1]$ is the Gauss process with zero mean and a covariance function $K_S(t,\tau),\ t,\tau \in [0,1]$. This problem arises in particular in applications of financial mathematics, when the assumption that the process under study is a Gaussian process, is not deemed sufficient reasonable. The proposed test generalizes multidimensional goodness-of-fit tests in R^n. This test should be based on n realisations $S_1(t), S_2(t), ..., S_n(t),\ t \in [0,1]$, of $S(t)$. The process $S(t)$ and its realisations are considered here as the elements of the Hilbert space $L^2([0,1])$. We choose as a basis for $L^2([0,1])$ the orthonormal basis formed by eigenfunctions $g_1(t), g_2(t), ...$ of the covariance operator with the kernel $K_S(t,\tau)$. The processes $S(t)$ and $S_i(t)$ can be represented in the form of expansion in the mentioned basis as $\mathbf{s} = (s_1, s_2, s_3, ...)$ and $\mathbf{s}_i = (s_{i1}, s_{i2}, s_{i3}, ...)$, correspondingly. The vector $\mathbf{s}$ has independent components with normal distributions. It can be transformed to the random vector $\mathbf{T} = (T_1, T_2, T_3, ...),\ \ \mathbf{T} \in [0,1]^\infty$, with the independent components, having the uniform distribution on $[0,1]$. The observations $\mathbf{s}_i$ can be transformed similarly to the observations $\mathbf{T}_i = (T_{i1}, T_{i2}, T_{i3}, ...)$ of the "uniform" distribution on $[0,1]^\infty$. It can be introduced a "distribution function"

$$F(\mathbf{t}) = F(t_1, t_2, t_3...) = P\{T_1 \le t_1^{\alpha_1}, T_1 \le t_1^{\alpha_1}, T_1 \le t_1^{\alpha_1}...\} = t_1^{\alpha_1} t_2^{\alpha_2} t_3^{\alpha_3}... \ .$$

Here, α_i should tend sufficiently quickly toward zero. Correspondingly, the empirical distribution function can be introduced as

$$F_n(\mathbf{t}) = F_n(t_1, t_2, t_3...) = (1/n)\, \sharp\{\mathbf{T}_i : \ \ T_{i1} \le t_1^{\alpha_1},\ T_{i2}^{\alpha_2} \le t_2, ...\}.$$

The empirical process $\xi_n(\mathbf{t}) = \sqrt(n)(F_n(\mathbf{t}) - F(\mathbf{t})),\ \ \mathbf{t} \in [0,1]^\infty$, weakly converges to the Gaussian process in $L_2(L_2[0,1])$. This process has the zero mean and covariance function

$$K(\mathbf{t}, \mathbf{v}) = \prod_{i=1}^{\infty} \min(t_i^{\alpha_i}, v_i^{\alpha_i}) - \prod_{i=1}^{\infty} t_i^{\alpha_i} v_i^{\alpha_i}.$$

Limit distribution of the Cramér-von Mises statistic is calculated using the methods described in the papers listed in the bibliography.

References

[1] Deheuvels, P., Martynov, G. 2003: Karhunen-Loève expansions for weighted Wiener processes and Brownian bridges via Bessel functions. , *Progress in Probability, Birkhäuser, Basel/Switzerland*, **55**, 57-93.

[2] Martynov, G. V., 1979: The omega square tests, *Nauka, Moscow*, 80pp., (in Russian).

[3] Martynov, G. V., 1992: Statistical tests based on empirical processes and related questions, *J. Soviet. Math*, **61**, 2195 - 2271.

Kernel Estimators of the Tail Index

CsörgőS Csörgő Mem. Session

DAVID M. MASON*,†

*Department of Applied Economics and Statistics, University of Delaware, Newark, Delaware, USA

†email: davidm@udel.edu

306:DavidMason.tex,session:CsorgoS

The research that I describe in my talk began at an Oberwolfach Conference on Order Statistics, Quantile Processes, and Extreme Value theory in March of 1984. During this meeting Sándor Csörgő, Paul Deheuvels and I worked out how to apply a new weighted approximation of the uniform empirical process by a Brownian bridge to establish a central limit theorem for a class of kernel estimators of the tail index of a Pareto-type distribution. This is a class of estimators $t_{n,h}$ of the tail index, which can be motivated by the Nadaraya-Watson regression function estimator, and it generalizes and includes the classical Hill estimator $a_{n,k}$. Their central limit theorem for $t_{n,h}$ appeared in [1].

It is well-known that the Hill estimator $a_{n,k}$ is a consistent estimator of the tail index if and only if $k \to \infty$ and $k/n \to 0$. It is shown in [1] that under suitable assumptions its generalized kernel version $t_{n,h}$ is also consistent, whenever the bandwidth is taken to be a sequence of positive non–random numbers satisfying $h_n \to 0$ and $nh_n \to \infty$. While attending a conference in honor of Sándor Csörgő's sixtieth birthday in 2007 at the University of Szeged, Sándor suggested to Julia Dony and me that we look into the problem of establishing a uniform in bandwidth version of the consistency result in [1]. Following his suggestion, we were able to extend it to show that consistency holds uniformly over a certain range of bandwidths. This permits the treatment of estimators of the tail index based upon data–dependent bandwidths h, which are often used in practice. In the process, we established a uniform in bandwidth result for kernel–type estimators with a fixed design, which will likely be of separate interest. Our uniform in bandwidth consistency result for $t_{n,h}$ appeared in [2].

My talk is partially based on joint work with Sándor Csörgő, Paul Deheuvels and Julia Dony. It will present two interesting applications of the technologies of weighted approximations and modern empirical process theory, and will provide me with the opportunity to discuss the general lines of my collaboration with my longtime friend Sándor.

Acknowledgment. Part of my research was supported by an NSF Grant.

References

[1] Csörgő, S., Deheuvels, P., Mason, D.M., 1985: Kernel estimates of the tail index of a distribution, *Ann. Statist.* **13**, 1050–1077.

[2] Dony, J., Mason, D.M., 2010: Uniform in bandwidth consistency of kernel estimators of the tail index, *Extremes 13* 353-371.

CS36A Graphical Methods

Chain Graph Modelling of Cognitive Profiles

M. SOFIA MASSA[*,‡], GLYN HUMPREYS[†]

[*]Department of Statistics, University of Oxford (UK),
[†]Department of Experimental Psychology, University of Oxford (UK)
[‡]email: massa@stats.ox.ac.uk 307:SofiaMassa.tex,session:CS36A

Stroke is one of the most important health problems affecting senior people. Stroke patients usually suffer a wide range of cognitive problems in their attention, language, memory, motor functions etc. For the past seven years, a test instrument named BCoS (Birmingham Cognitive Screen) has been developed to screen individuals for cognitive problems with an approximately one hour testing session. By 2011, the BCoS database has entailed 773 stroke patients' performance in 23 cognitive tests across five cognitive domains, together with these patients' personal and clinical information. We apply chain graphical models techniques to investigate the association structure among stroke patients' personal traits, clinical conditions, and test performance recorded in the BCoS data set. Performing model selection via chain graphical models will permit to show the association between variables belonging to the five different domains. We will also show that, although each BCoS test is officially categorized into one single cognitive domain, some tests assess abilities across multiple domains.

IS25 Stat. SDE

Estimation of Stable-Like Stochastic Differential Equations

HIROKI MASUDA[*,†]

[*]Kyushu University, Kyushu, Japan
[†]email: hiroki@imi.kyushu-u.ac.jp 308:Masuda.tex,session:IS25

We consider the stochastic differential equation of the form

$$dX_t = a(X_t, \alpha)dt + c(X_{t-}, \gamma)dJ_t,$$

where the coefficients a and c are supposed to be known except for the finite-dimensional parameter $\theta = (\alpha, \gamma) \in \Theta \subset \mathbb{R}^p$, and the driving process J is a pure-jump Lévy process whose Blumenthal-Getoor index

$$\beta := \inf\left\{q > 0 : \int_{|z|\le 1} |z|^q \nu(dz) < \infty\right\} \in (0, 2),$$

with ν denoting the Lévy measure of J. We wish to estimate the true value $\theta_0 \in \Theta$ (supposed to exist) based on a discrete-time but high-frequency sample $X_{t_0}, X_{t_1}, \ldots, X_{t_n}$, where $t_j = t_j^n = jh_n$ with

$h_n \to 0$ as $n \to \infty$. A naive way would be to use the Gaussian quasi likelihood. However, although the Gaussian quasi likelihood is known to be well-suited for the case of diffusions, it leads to asymptotically suboptimal estimator in the pure-jump case; in particular, the Gaussian quasi-maximum likelihood estimation inevitably needs that $t_n \to \infty$.

In this talk, we will introduce another kind of quasi-maximum likelihood estimator $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\gamma}_n)$ based on the local-stable approximation of the one-step transition distribution $\mathcal{L}(X_{t_j}|X_{t_{j-1}})$; the proposed estimation procedure is a pure-jump counterpart to the Gaussian quasi-maximum likelihood estimation. Under some regularity conditions including that $\mathcal{L}(h^{-1/\beta}J_h)$ tends to the standard β-stable distribution as $h \to 0$, we will show that

$$\left\{\sqrt{n}h_n^{1-1/\beta}(\hat{\alpha}_n - \alpha_0),\ \sqrt{n}(\hat{\gamma}_n - \gamma_0)\right\}$$

are jointly asymptotically mixed-normal (reps. normal) when $\limsup_n t_n < \infty$ (resp. $t_n \to \infty$ and X is ergodic). As a result, in case of the stable-like J, the proposed estimator $\hat{\theta}_n$ can be asymptotically much more efficient than the Gaussian quasi-maximum likelihood estimator.

Acknowledgment. This research was partially supported by JSPS KAKENHI Grant Number 23740082.

Coordination of Conditional Poisson Samples

CS6D Dyn. Response Mod.

ANTON GRAFSTRÖM*, ALINA MATEI†,‡,§

*Swedish University of Agricultural Sciences, Umeå, Sweden,
†Institute of Statistics, University of Neuchâtel, Switzerland,
‡IRDP, Neuchâtel
§email: alina.matei@unine.ch

309:AlinaMatei.tex,session:CS6D

In survey sampling, consider samples drawn successively or simultaneously from overlapping finite populations. Sample coordination seeks to create a dependence between these samples. This dependence leads to a maximization or to a minimization of the sample overlap. The former is known as positive coordination, and the latter as negative coordination. Positive coordination is mainly employed for estimation reasons and data collection costs. Negative coordination is mainly performed to diminish the response burden of the sampled units.

Poisson sampling design with permanent random numbers provides an optimum coordination degree of two or more samples. The realized sample size of a Poisson sample is, however, random. Conditional Poisson (CP) sampling (Hajek, 1981) is a modification of the classical Poisson sampling that produces a fixed size πps sample, and possess the maximum entropy property subject to given inclusion probabilities. We introduce two methods to coordinate Conditional Poisson samples over time or simultaneously. The first one uses permanent random numbers and a sequential implementation of two CP-samples. The second one is based on a rejective implementation, uses a CP-sample in the first occasion and provides an approximate one in the second occasion. The methods are evaluated using the size of the expected sample overlap, and are compared with their competitors using Monte Carlo simulation. The new methods provide excellent coordination degree of two samples, close to the performance of Poisson sampling with permanent random numbers.

References

[Hajek (1981)] Hajek, J., 1981. *Sampling from a finite population.* Marcel Dekker, New York.

CS24A Branching Proc.

Maximum Likelihood Estimation for a Random Walk in a Parametric Random Environment

FRANCIS COMETS*, MIKAEL FALCONNET†, OLEG LOUKIANOV‡, DASHA LOUKIANOVA§, CATHERINE MATIAS†,¶

*Laboratoire Probabilités Modèles Aléatoires, Université Paris Diderot, UMR CNRS 7599, France
,
†Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071, USC INRA, France,
‡Département Informatique, IUT de Fontainebleau, Université Paris Est, France,
§Laboratoire Analyse et Probabilités, Université d'Évry Val d'Essonne, France.
¶email: cmatias@genopole.cnrs.fr 310:Catherine_Matias.tex,session:CS24A

We consider a one dimensional ballistic random walk evolving in a parametric independent and identically distributed random environment. We study the asymptotic properties of the maximum likelihood estimator of the parameter, based on a single observation of the path untill the time it reaches a distant site. This study relies on the link between the original process and a branching process with immigration in a random environment. We prove that the estimator is consistent, asymptotically normal and establish that it achieves the Cramér-Rao bound (thus being efficient), as the distant site tends to infinity. In a simulation setting, we also explore the numerical performances of this estimator as well as the behaviour of asymptotic confidence regions for the parameter value.

OCS23 Strong Limit Thm.

Some Covariance and Comparison Inequalities for Positively Dependent Random Variables and their Applications

PRZEMYSŁAW MATUŁA*,†, MACIEJ ZIEMBA*

*Institute of Mathematics, Marie Curie–Słodowska University, Lublin, Poland
†email: matula@hektor.umcs.lublin.pl 311:PrzemyslawMatula.tex,session:OCS23

Let X and Y be positively quadrant dependent (PQD) random variables (r.v.'s), i.e. such that

$$H(t,s) := P\left(X \leq t, Y \leq s\right) - P\left(X \leq t\right) P\left(Y \leq s\right) \geq 0,$$

for all $t, s \in \mathbb{R}$. Denote by X' and Y' two independent r.v.'s such that X' has the same distribution as X and Y' the same distribution as Y. We study the comparison inequalities, by finding the upper bounds for

$$H(K) := \left|P\left([X,Y] \in K\right) - P\left(\left[X',Y'\right] \in K\right)\right|$$

in terms of the covariance $\mathrm{Cov}(X,Y)$ of the original r.v.'s and some other their characteristics (e.g. norms of the densities in the absolutely continuous case). Here $K \subset \mathbb{R}^2$ is a set satisfying some regularity conditions.

In the case $K = (-\infty, t\rangle \times (-\infty, s\rangle$, obviously $H(K) = H(t,s)$ for PQD r.v.'s. The bounds for $H(t,s)$, called covariance inequalities, have been obtained by several authors listed in the references. In the presentation we shall show some new results including multivariate extensions of the aforementioned inequalities, as well as their applications.

References

[1] Bagai, I.; Prakasa Rao, B.L.S.: *Estimation of the survival function for stationary associated processes*, Statist. Probab. Lett. 12 (1991), no. 5, 385–391.

[2] Roussas, G. G.: *Kernel estimates under association: strong uniform consistency*, Statist. Probab. Lett. 12 (1991), no. 5, 393–403.

[3] Matuła, P.: *On some inequalities for positively and negatively dependent random variables with applications*, Publ. Math. Debrecen 63 (2003), no. 4, 511–522.

[4] Matuła, P.: *A note on some inequalities for certain classes of positively dependent random variables*, Probab. Math. Statist. 24 (2004), no. 1, Acta Univ. Wratislav. No. 2646, 17–26.

[5] Matuła, P.; Ziemba, M.: *Generalized covariance inequalities*, Cent. Eur. J. Math. 9 (2011), no. 2, 281–293.

The Probability Weighted Empirical Characteristic Function and Goodness-of-Fit Testing

CS9C Model Selection

SIMOS G. MEINTANIS*,†

*Department of Economics, National and Kapodistrian University of Athens, Athens, Greece
†email: simosmei@econ.uoa.gr 312:Meintanis.tex,session:CS9C

We introduce the notion of the probability weighted empirical characteristic function as a generalization of the empirical characteristic function. Then some of its properties are studied, and its potential use in goodness–of–fit testing is examined.

Dynamic Prediction for Multi-State Survival Data

OCS18 Lifetime Data Anal.

LUIS MEIRA-MACHADO*,†

*Centre of Mathematics and Department of Mathematics and Applications, University of Minho, Portugal
†email: lmachado@math.uminho.pt 313:Meira-Machado.tex,session:OCS18

The analysis of survival data may be described by the Markov process with two states, 'alive' and 'dead' and a single transition between them. This is known as the multi-state mortality model. Multi-state models may be considered a generalization of survival analysis where survival is the ultimate outcome of interest but where intermediate (transient) states are identified. For example, in cancer studies more than one endpoint may be defined such as 'local recurrence', 'distant metastasis' and 'dead'. The so-called "illness-death" model plays a central role in the theory and practice of multi-state models. In the irreversible version of this model, individuals start in the "healthy" state and subsequently move either to the "diseased" state or to the "dead" state. Individuals in the "diseased" state will eventually move to the "dead" state without any possibility of recovery. Many time-to-event data sets from medical studies with multiple end points can be reduced to this generic structure. Thus, methods developed for the three-state illness-death model have a wide range of applications. From a theoretical standpoint, this is the simplest multi-state generalization of the survival analysis model that incorporates both branching (as in a multiple decrement/competing risk model) and an intermediate state (as in a progressive tracking model). Thus, unlike the survival or the competing risk model, this model is not necessarily Markovian.

One important goal in multi-state modeling is the estimation of transition probabilities. In longitudinal medical studies these quantities are particularly of interest since they allow for long-term predictions of the process. In recent years significant contributions have been made regarding this topic. However, most of the approaches assume independent censoring and do not account for the influence of covariates. Other important targets include the state occupation probabilities (which can be seen as a particular case of the transition probabilities), the cumulative incidence function and

the waiting time distributions. This paper introduces feasible estimation methods for all these quantities in an illness-death model conditionally on current or past covariate measures. The proposed methods are illustrated using real data.

Acknowledgment. This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", in the form of grants PTDC/MAT/104879/2008 and Est-C/MAT/UI0013/2011.

References

[Meira-Machado et al. (2009)] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K., 2009: Multi-state models for the analysis of time to event data, *Statistical Methods in Medical Research*, **18**, 195 - 222.

[Meira-Machado et al. (2012)] Meira-Machado, L., de Uña-Álvarez, J., Datta, S., 2012: Conditional Transition Probabilities in a non-Markov Illness-death Model, *Discussion Papers in Statistics and Operation Research 11/03, 2011. Department of Statistics and Operations Research, University of Vigo*. http://webs.uvigo.es/depc05/reports/

OCS1 Longitudin Models

Joint Modelling of Longitudinal and Time-to-Event Outcome in Peritoneal Dialysis

DENISA MENDONÇA[*,§], LAETITIA TEIXEIRA[†], INÊS SOUSA[‡]

[*]Institute of Biomedical Sciences Abel Salazar and Public Health Institute, University of Porto, Portugal,
[†]Doctoral Program in Applied Mathematics, Institute of Biomedical Sciences Abel Salazar and Faculty of Sciences, University of Porto, Portugal,
[‡]Department of Mathematics and Applications, University of Minho, Portugal
[§]email: dvmendon@icbas.up.pt 314:DenisaMendonca.tex,session:OCS1

Survival and longitudinal analysis have been widely used in public health research from medical and epidemiological investigations to health economics and social and behavioral studies. End-stage renal disease patients starting peritoneal dialysis are submitted to a regular assessment since the entrance in the peritoneal dialysis program. This type of longitudinal studies produces two types of outcomes: (i) a set of repeated measures on covariates measured at baseline (e.g. gender, age and diabetes) and at each control visit (e.g., ultrafiltration and creatinine clearance) and (ii) the time to an event of interest, in the presence of competing risks. During this follow-up time period, peritoneal dialysis patients are exposed to several events (death, transfer to haemodialysis or renal transplantation) and the occurrence of one of them precludes the occurrence of the others. Therefore in the presence of these competing events, a competing risk survival analysis should be used. These two types of outcome are often separately analysed using two different models, one for longitudinal outcome (e.g. linear mixed model) and one for time to an event of interest outcome [Rizopoulos, 2010]. However, in these cases, joint modelling of longitudinal data and competing risks failure time data is required when the interest is the interrelationships between these two types of outcome [Williamson et al. 2008].

A typical example of the relevance of the joint modelling approach is the evaluation of peritoneal dialysis programs which has motivated this research, with the objectives: (1) to understand within-subject patterns of change in longitudinal outcomes and/or (2) to characterize the relationship between features of longitudinal outcomes and time to event of interest or other competing event [Tsiatis and Davidan, 2004]. For example, in this study, we could consider renal transplantation as the event of interest (with transfer to haemodialysis as competing risk) and several indicators of renal condition (such as ultrafiltration and creatinine clearance) as longitudinal outcomes. Other baseline

covariates, such as gender, age and diabetes, could also be considered. The `R` software was used, particularly the `JM` package for fitting shared parameter models for the joint modelling of normal longitudinal outcomes and event times under a maximum likelihood approach.

References

[1] Rizopoulos, D., 2010: JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data, *Journal of Statistical Software*, **35**(9), 1–33.

[2] Tsiatis, A. A., Davidian, M. 2004: Joint modeling of longitudinal and time-to-event data: An overview, *Statistica Sinica*, **14**(3), 809–834.

[3] Williamson, P. R., Kolamunnage-Dona, R., Philipson, R., Marson, A. G. 2008: Joint modelling of longitudinal and competing risks data, *Statistics in Medicine*, **27**(30), 6426–6438.

Asymptotic Representation for Presmoothed Kaplan-Meier Integrals with Covariates

CS19E Lim. Thms.

JORGE MENDONÇA[*,‡], JACOBO DE UÑA-ÁLVAREZ[†]

[*]Polytechnic Institute of Oporto, School of Engineering, Portugal,
[†]Department of Statistics and Operations Research, Faculty of Economics and Business, University of Vigo, Spain
[‡]email: jpm@isep.ipp.pt

315:JorgeMendonca.tex,session:CS19E

The Kaplan-Meier method is tipically used when lifetimes are observed under random censorship. The strong consistency of Kaplan-Meier integrals was proved in Stute and Wang (1993) and,in the presence of covariates in Stute(1993). An asymptotic representation of these integrals was introduced in Stute(1995) and in Stute(1996) when covariates are present. However, this method loses accuracy when many data are censored. In that case a semiparametric approach (Dikta 1998, 2000 and 2005) might be useful. The strong consistency of presmoothed Kaplan-Meier integrals in this scenario was proved in Dikta(2000) and in the presence of covariates in de Uña-Álvarez and Rodríguez-Campos (2004). In this work we obtain an asymptotic representation of presmoothed Kaplan-Meier integrals with covariates, under a semiparametric censorship model. As a Corollary, a CLT for the presmoothed integrals is established.

References

[de Uña-Álvarez J, Rodríguez-Campos C (2004)] de Uña-Álvarez J, Rodríguez-Campos C.,2004: Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present, *Statistics*, **38**., 483 - 496.

[Dikta G (1998)] Dikta G.,1998: On semiparametric random censorship models. *Journal of Statistical Planning and Inference*,**66**, 253-279.

[Dikta G (2000)] Dikta G.,2000: The strong law under semiparametric random censorship models. *Journal of Statistical Planning and Inference*,**83**, 1-10.

[Dikta et al. (2005)] Dikta G.,J. Ghorai, and C. Schmidt., 2005: The central limit theorem under semiparametric random censhorship models . *Journal of Statistical Planning and Inference* **127**, 23-51.

[Stute, W. (1993)] Stute, W.,1993: Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis***45**, 89 - 103.

[Stute, W. and Wang J-L(1993)] Stute, W. and Wang J-L,1993: The strong law under random censorship. *The Annals of Statistics***3**, 1591 - 1607.

[Stute, W. (1995)] Stute, W.,1995: The central limit theorem under random censorship. *The Annals of Statistics***23**, 422 - 439.

[Stute, W. (1996)] Stute, W.,1996: Distributional convergence under random censorship when covariates are present. *Scandinavian Journal of Statistics* **23**, 461 - 471.

SIL Spec. Invited Lecture

Being an Informed Bayesian: Assessing Prior Informativeness and Prior–Likelihood Conflict

XIAO LI MENG*,†

*Department of Statistics, Harvard University, Cambridge, MA, USA

†email: meng@stat.harvard.edu 316:XiaoLiMeng.tex,session:SIL

Dramatically expanded routine adoption of the Bayesian approach has substantially increased the need to assess both the confirmatory and contradictory information in our prior distribution, in reference to the information provided by our likelihood. Our diagnostic approach starts with the familiar posterior matching method; for a given likelihood model, we identify the difference in the sample sizes needed to form two likelihood functions that, when combined respectively with a given prior and a "baseline" prior, will lead to the same posterior summaries as chosen. This difference can be viewed as a "prior data size" M(k), relative to the likelihood based on k independent, identically distributed observations. The confirmatory information is captured by the M(k) function, which is roughly constant over k when no serious prior-likelihood conflict arises. The contradictory information is detectable in its derivative or finite difference as M(k) tends to decrease with k when contradictory prior specification detracts information from the likelihood. Intriguing findings include a universal low bound, -1, on the derivative of M(k) that represents the most extreme prior-likelihood conflict, and a super-informative phenomenon where the prior effectively gains an extra 50% prior data size relative to the baseline when the prior mean coincides with the truth. We demonstrate our method via several examples, including an application exploring the effects of immunoglobulin levels on lupus nephritis. We also establish theoretical results showing why the derivative of M(k) is a useful indicator for prior-likelihood conflict. (This is joint work with Matthew Reimherr and Dan Nicolae of The University of Chicago.)

OCS16 Interacting Particles

Universality Classes of Lozenge Tilings of a Polyhedron

ANTHONY METCALFE*,†

*KTH, Royal Institute of Technology, Stockholm, Sweden

†email: metcalf@kth.se 317:AnthonyMetcalfe.tex,session:OCS16

A regular hexagon can be tiled with lozenges of three different orientations. Letting the hexagon have sides of length n, and the lozenges have sides of length 1, we can consider the asymptotic behaviour of a typical tiling as n increases. Typically, near the corners of the hexagon there are regions of "frozen" tiles, and there is a "disordered" region in the center which is approximately circular.

More generally one can consider lozenge tilings of polyhedra with more complex boundary conditions. The local asymptotic behaviour of tiles near the boundary of the equivalent "frozen" and "disordered" regions is of particular interest. In this talk, we shall discuss work in progress in which we classify necessary conditions under which such tiles behave asymptotically like a determinantal random point field with the *Airy* kernel, and also with the *Pearcey* kernel. We do this by considering an equivalent interlaced discrete particle system.

Acknowledgment. Supported/Partially supported by the grant KAW 2010.0063 from the Knut and Alice Wallenberg Foundation.

Estimating Distribution of Age at Menarche Based on Recall Information

CS17A Causal Inference

SEDIGHEH MIRZAEI S.[*,†,‡], DEBASIS SENGUPTA[*]

[*]Indian Statistical Institute, Kolkata,
[†]Indian Statistical Institute, Kolkata
[‡]email: sedigheh_r@isical.ac.in, sdebasis@isical.ac.in
318:SedighehMirzaei.tex,session:CS17A

Average age at menarche finds application in a variety of contexts. It is a comparative indicator of population health and timing of maturation, and is also widely used as a demographic indicator of population fecundity. The most common approach of estimating age at menarche is the 'Status quo' approach, which makes use of dichotomous data and a logit/probit analysis. Dichotomous responses (whether menarche has occurred till the day of observation) are easy to obtain by asking respondent girls whether they have experience menarche, and statistical routines for logit and probit analysis of dichotomous data are available in most statistical packages. Recall data contain more information than 'status quo' data, and are expected to produce better estimates. Since recall data are generally interval censored, some scientists have used the non-parametric estimator proposed by Turnbull (1976). However, the nature of censoring involved in gathering retrospective menarchial data is informative. Alternative modelling has so far been limited to a parametric set-up. In this work we provide a non-parametric estimator, based on a likelihood that makes use of the special nature of the data at hand. Monte Carlo simulations produce encouraging results on the performance of the proposed estimator.

References

[Hediger M. L. et al. 1987] Hediger, M. L., and Stine R. A. (1987): Age at menarche based on recall data. *Ann. Hum. Biol.*, **14**, 133-142.

[Turnbull, B. W. 1976] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, **38**, 290-295.

[Lehman, E. L. 1999] Lehman, E. L. (1999). *Elements of large sample theory*. Springer, New York.

[Noceda, J. et al. 2006] Noceda, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York.

A New Method for Construction Spatio-Temporal Covariance with Copula Functions

CS7B Spatio-Temp. Stat II.

MOHSEN MOHAMMADZADEH[*,†], MEHDI OMIDI[*]

[*]Tarbiat Modares University, Tehran, Iran.
[†]email: mohsen_m@modares.ac.ir
319:Mohammadzadeh.tex,session:CS7B

Statistical analysis of natural phenomena with spatial and temporal correlation requires to specify the correlation structure of the data via a covariance function. Usually a separable spatio-temporal covariance function is used for the ease of application. But the separability of the spatio-temporal covariance function can be unrealistic. For these cases, a non-separable spatio-temporal covariance function is required.

Cressie and Huang (1999) introduced and developed a set of valid non-separable spatio-temporal covariance functions through Fourier transforms. Gneiting (2002) developed this approach to the

completely monotone and Bernstein functions. Fuentes *et al.* (2008) used spectral densities for introducing nonseparable covariance functions. Kent *et al.* (2011) showed that in certain circumstances Gneiting's model possesses a dimple, which detracts from its modeling appeal.

In this paper, we investigate the role of Stieltjes transformation and copula functions in construction of non-separable spatio-temporal covariances. Then, by using structural copula functions, a new family of non-separable spatio-temporal covariances is introduced. Next, we proved that this family of covariance functions does not possess the existed dimple in Gneiting's model. More over a Genetic algorithms is used to estimate the parameters of the constructed covariance model, in order to specify the spatio-temporal correlation of the Ozone data in Tehran city.

Acknowledgment. Partial support from Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad is acknowledged.

References

[Cressie and Huang (1999)] Cressie, N., Huang, H. C., 1999: Classes of non-separable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association*, **96**, 1330 - 1340.

[Fuents et al (2008)] Fuentes, M., Chen, L., Davis, J. M., 2008: A class of nonseparable and nonstationary spatial temporal covariance functions, *Environmetrics*, **19**, 487 - 507.

[Gneiting (2002)] Gneiting, T., 2002: Nonseparable, stationary covariance functions for space-time data, *Journal of the American Statistical Association*, **97**, 590 - 600.

[Kent et al (2011)] Kent, J., Mohammadzadeh, M., Mosammam, A., 2011: The Dimple in Gneiting's Spatio Temporal Covariance Model, *Biometrika*, **98**, 489 - 494.

Goodness-of-Fit Test for Linear Hypothesis in Nonparametric Regression Model

ZAHER MOHDEB[*,†]

[*]University Constantine 1, Department of Mathematics, Constantine, Algeria
[†]email: z.mohdeb@gmail.com 320:ZaherMohdeb.tex,session:CS31A

We consider the following regression model

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\,, \quad i = 1, \ldots, n\,,$$

where f is a unknown real function, defined on the interval $[0,1]$ and $t_{1,n} = 0 < t_{2,n} < \cdots < t_{n,n} = 1$, is a fixed sampling of the interval $[0,1]$. The errors $\varepsilon_{i,n}$ form a triangular array of random variables with expectation zero and finite variance σ^2, and for any n, the random variables $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ are independent.

Let $U_p = \text{span}\{g_1, \ldots, g_p\}$, where $g_1, \ldots, g_p$ denote p linearly independent functions defined on $[0,1]$. In order to test

$$\mathrm{H}_0: \quad f \in U_p \qquad \text{against} \qquad \mathrm{H}_1: \quad f \notin U_p\,,$$

we use a test statistic based on the mean of squared residuals and show that it has a parametric asymptotic behaviour. We show the asymptotic normality of the test statistic under the null hypothesis and the alternative. In order to investigate the small-sample properties of the level of significance and the power, we conduct a Monte Carlo study for sample size $n = 20$. The Monte Carlo results show that under H_0, the distribution of the proposed test statistic is close to the normal distribution and that the test has good power properties.

References

[1] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 778-800.

[2] Mohdeb, Z. and Mokkadem, A. (2004). On the use of nonparametric regression for testing linear hypotheses. *J. Nonparametr. Stat.*, **16**, no. 1-2, 3–12.

[3] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Stat.*, **12**, 1215-1230.

Determinantal Point Process Models and Statistical Inference

IS21 Spatial Point Proc.

FRÉDÉRIC LAVANCIER*, JESPER MØLLER†,‡, EGE RUBAK†

*University of Nantes, Nantes, France,
†Aalborg University, Aalborg, Denmark
‡email: jm@math.aau.dk 321:JesperMoller.tex,session:IS21

Determinantal point processes (DPPs) are largely unexplored in statistics, though they possess a number of appealing properties and have been studied in mathematical physics, combinatorics, and random matrix theory. In this talk we consider statistical models and inference for DPPs defined on $\mathbb{R}^d$, with a focus on $d = 2$.

DPPs are defined by a function C satisfying certain regularity conditions; usually C is a continuous covariance function where its spectrum is bounded by one. DPPs possess the following appealing properties:

(a) They are flexible models for repulsive interaction, except in cases with strong repulsiveness (as e.g. in a hard-core point process).

(b) All orders of moments of a DPP are described by certain determinants of matrices with entries given in terms of C.

(c) A DPP restricted to a compact set has a density (with respect to a Poisson process) which is expressible in closed form.

(d) A DPP can easily be simulated, since it is a mixture of 'determinantal projection processes'.

(e) A one-to-one smooth transformation or an independent thinning of a DPP is also a DPP.

In contrast, Gibbs point processes, which constitute the usual class of models for repulsive interaction, do not in general have moments that are expressible in closed form, the density involves an intractable normalizing constant, rather time consuming Markov chain Monte Carlo methods are needed for simulations and approximate likelihood inference, and an independent thinning of a Gibbs point process does not result in a tractable point process.

In the talk, we discuss the fundamental properties of DPPs, investigate how to construct parametric models, and study different inferential approaches based on moments or maximum likelihood.

Acknowledgment. Supported by the Danish Council for Independent Research | Natural Sciences, grant 09-072331, "Point Process Modelling and Statistical Inference", and grant 12-124675, "Mathematical and Statistical Analysis of Spatial Data". Supported by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation.

POSTER
Poster

Design Based Inference for a Continuous Spatial Population Mean

GIORGIO E. MONTANARI[*,†], GIUSEPPE CICCHITELLI[*]

[*]University of Perugia, Italy
[†]email: gem@unipg.it

322:GiorgioE.Montanari.tex,session:POSTER

In environmental studies, dealing with soil characterization, monitoring of natural resources or estimation of pollution concentration, it is often required making inference on the mean or total of a response variable defined continuously over a region. Classical sampling theory deals with finite populations, but sampling theory and continuous populations intersect in the statistical analysis of spatial data. In this regards, pioneering contributions have dealt with the comparison of various sampling strategies to estimate the average level of a response surface of a pollutant, and theory for the Horvitz-Thompson estimation of parameters of a continuous spatial population based on probability sampling has been developed. The key issue that allows a spatial continuous population to be the object of sampling theory is the possibility to label the population units (spatial points). In fact, this labelling process is the necessary condition for drawing random samples from the population. In a continuous spatial population the units are the points of the domain of study, and the labels are their geographical coordinates. The problem of the estimation of the mean or of the total of a continuous spatial population has been addressed, among others, by Stevens and Olsen (2004). The focus is on the improvement of the sampling strategy efficiency, using at the design stage the auxiliary information provided by the spatial coordinates of points in the domain. If we assume that the response variable has a spatial structure in the study domain, an efficient sample is evenly and regularly spread out across the domain as much as possible. There are various techniques to obtain spatially balanced samples. Here we refer, in particular, to the Random Tessellation Stratified design and to the Generalized Random Tessellation Stratified design. The Horvitz-Thompson estimator is coupled with those sampling designs. But another possibility is to construct a more efficient estimator making use of the auxiliary information provided by the spatial coordinates of units. This approach has been adopted in Cicchitelli and Montanari (2012), where a penalised spline regression model is employed to capture the spatial pattern in the data and a design consistent estimator is proposed within the framework of the model-assisted approach to inference for finite populations. In this paper we specialize on continuous spatial populations, and focus on two main aspects: the data-driven choice of the penalty factor of the penalised spline regression model and the estimation of the sampling variance of the proposed model-assisted estimator. A simulation study compares the proposed estimator with the block kriging predictor, currently used in geostatistics, the latter being the optimal predictor under a model dependent approach assuming a second order stationary spatial process.

References

[1] Cicchitelli, G. and Montanari, G. E., 2012: Design-based estimation of a spatial population mean, *International Statistical Review*, **80**, 111-126.

[2] Stevens, D. L., Jr. and Olsen A.R., 2004: Spatially balanced sampling of natural resources, *Journal of the American Statistical Association*, **99**, 262-278.

Nonparametric Estimation of a Distribution Function from Doubly Truncated Data under Dependence

CS6F Copulas

CARLA MOREIRA[*,§], JACOBO DE UÑA-ÁLVAREZ[†], ROEL BRAEKERS[‡]

[*]University of Vigo - Department of Statistics and O.R. Lagoas - Marcosende, 36 310, Vigo, Spain,
[†]Center of Mathematics and Department of Mathematics and Applications - Campus de Azurém, 4800-058 Guimarães, Portugal,
[‡]University of Hasselt, Campus Diepenbeek - Center for Statistics, 3590 - Diepenbeek, Belgium
[§]email: carla@uvigo.es 323:CarlaMoreira.tex,session:CS6F

The NPMLE of a distribution function from doubly truncated data was introduced in the seminal paper of Efron and Petrosian [Efron and Petrosian (1999)]. The consistency of the Efron-Petrosian estimator depends however on the assumption of independent truncation. In this work we introduce and extension of the Efron-Petrosian NPMLE when the lifetime and the truncation times may be dependent. The proposed estimator is constructed on the basis of a copula function which represents the dependence structure between the lifetime and the truncation times. Two different iterative algorithms to compute the estimator in practice are introduced, and their performance is explored through an intensive Monte Carlo simulation study. The asymptotic properties of the proposed estimator will be explored. Several applications to medical data are included for illustration purposes.

Acknowledgment. This research was partially supported by Grant MTM2011-23204 (FEDER support included) of the Spanish Ministerio de Ciencia e Innovación, SFRH/BPD/68328/2010 Grant of Portuguese Fundação Ciência e Tecnologia and IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

References

[Efron and Petrosian (1999)] Efron, B. and V. Petrosian (1999): Nonparametric methods for doubly truncated data, *Journal of the American Statistical Association*, **94**, 824 - 834.

Adaptive Estimation in Non-Regular Nonparametric Regression

CS6B Funct. Est., Regression

MORITZ JIRAK[*,‡], ALEXANDER MEISTER[†], MARKUS REISS[*]

[*]Humboldt-Universität zu Berlin, Berlin, Germany,
[†]Universität Rostock , Rostock, Germany
[‡]email: jirak@math.hu-berlin.de 324:MoritzJirak.tex,session:CS6B

We consider the model of non-regular nonparametric regression where smoothness constraints are imposed on the regression function and the regression errors are assumed to decay with some sharpness level at their endpoints. These conditions allow to improve the regular nonparametric convergence rates by using estimation procedures which are based on local extreme values rather than local averaging. We study this model under the realistic setting in which both the smoothness and the sharpness degree are unknown in advance. We construct adaptation procedures by Lepski's method and Pickands' estimator which show no loss in the convergence rates with respect to the integrated squared risk and a logarithmic loss with respect to the pointwise risk. Moreover we prove that this logarithmic deterioration of the rates cannot be avoided, hence our results are optimal in the minimax sense. The proofs are based on a concentration result, that has interest in itself. Some numerical simulations are provided.

IS19 Shape & Image

Bayesian Object Regression for Complex, High Dimensional Data

JEFFREY S. MORRIS*

*The University of Texas, MD Anderson Cancer Center

325:Morris.tex,session:IS19

A growing number of studies yield object data, which involve multiple measurements on some type of structured space, and include, for example, functions, images, shapes, graphs, and trees. The internal structure of the objects can be based on geometry or more complex scientific relationships, and should be accounted for in the modeling. In this talk, I will discuss very general and flexible Bayesian modeling frameworks that can be used to perform regression analyses on a broad array of such object data. Our strategy involves the use of various types of basis functions to capture different types of internal structure, using a modeling strategy that is conducive to parallel processing and scales up to very large data sets. The software to apply these methods will be general enough to handle a wide array of models and be used with nearly any type of object data sampled on a fine grid. Some methodological innovations I will touch upon include approaches for object-on-object regression, nonparametric additive models for object data, and functional spatial models. I will illustrate the flexibility of these methods in several application areas, including event-related potential neuroimaging data, functional MRI data, copy number genomics data, and opthalmological data involving measurements taken continuously on the surface of the eyeball.

OCS29 Stat. Branching Proc.

Extinction Probability in Two-Sex Branching Processes with Reproduction and Mating Depending on the Number of Females and Males in the Population

CHRISTINE JACOBS*, MANUEL MOLINA†, MANUEL MOTA†,‡

*Applied Mathematics and Informatics Unity, INRA, Jouy-en-Josas, France.,
†Department of Mathematics, University of Extremadura, Badajoz, Spain.
‡email: mota@unex.es

326:ManuelMota.tex,session:OCS29

Branching process theory deals with populations where an individual is followed up for some time and then may be replaced by other individuals. We focus our interest in branching processes for description of sexually reproducing populations. This situation was initially studied by Daley (1968) where the bisexual Galton-Watson process was introduced. From Daley's model, several classes of two-sex branching processes have been investigated, see Hull (2003) and Molina (2010).

In this work, we introduce a class of two-sex branching processes where, in each generation, both the reproduction and the mating depend on the numbers of females and males in the population. The mating between a female and a male is governed by a Bernoulli variable. These variables are correlated for the couples in a generation in such a way that both sexes do not play the same role: a female is assumed to form a real couple with a male at most. The variables governing the reproduction are supposed to be independent conditional on the knowledge of the couples formed in a generation.

For such a model, we provide several probabilistic contributions, including a sufficient condition for the extinction-explosion of the population. We present also some results about the probability of extinction, giving some conditions for both, the almost sure extinction of the process and the existence of a positive probability for the process to survive. On the basis of these probabilistic results, some inference on the main parameters of the model can be carried out.

Acknowledgment. This research has been supported by the Gobierno de Extremadura and the Ministerio de Economía y Competitividad of Spain and the FEDER, grants GR10118 and MTM2012-31235, respectively.

References

[1] Daley, D.J. (1968). Extinction conditions for certain bisexual Galton-Watson branching processes. Zeitschrift für Wahrscheinlichkeitstheorie, 9, 315-322.

[2] Hull, D.M. (2003). A survey of the literature associated with the bisexual Galton- Watson branching process. Extracta Mathematicae, 18, 321-343.

[3] Molina, M. (2010). Two-sex branching process literature. Lectures Notes in Statistics, 167, 279-293. Springer-Verlag.

Are Jumps in Price and Volatility Correlated?

IS11 Limit Thm. Appl.

JEAN JACOD*, CLAUDIA KLÜPPELBERG†, GERNOT MÜLLER‡,§

*Université Paris VI, Paris, France,
†Technische Universität München, Garching, Germany,
‡Carl von Ossietzky Universität, Oldenburg, Germany
§email: gernot.mueller@uni-oldenburg.de 327:GernotMueller.tex,session:IS11

Models for financial data involving a stochastic volatility and allowing for sample path discontinuities in the volatility as well as in the underlying asset price have become more and more popular in recent years. Moreover, there is an increasing number of publications showing empirical evidence of jumps in the asset prices as well as in the volatility process. The consequences of the existence of jumps for risk management, portfolio optimization and derivatives pricing have been extensively discussed already in the last decades.

After statistical evidence for the presence of jumps has been found, it is important to investigate in a next step a possible relation between price jumps and volatility jumps. For example, one might assume that price and volatility never jump together; however, this was ruled out by empirical evidence. Supposing now that price and volatility may jump at the same time, what are reasonable models for such "co-jumps"? Of course, there are many possibilities, ranging from an extreme one where there is a functional relationship between their sizes, to the other extreme where the two jump sizes are independent.

In this talk we develop an asymptotic test to answer the question whether common jumps in price and volatility are correlated or not. The test only investigates big jumps of the observable price process and uses local volatility estimates calculated from observation windows before and after the jumps under consideration. The performance of the test is checked in a simulation study. Finally, the test is applied to high-frequency data sets.

A Nonparametric Bayesian Model for a Clinical Trial Design for Targeted Agents

IS2 Bayesian Nonpar.

PETER MÜLLER*,†

*University of Texas at Austin, Austin, USA
†email: pmueller@math.utexas.edu 328:Peter_Mueller.tex,session:IS2

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 is based on 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 obejective of the trial. Patients are randomized to targeted therapy or standard chemotherapy.

OCS10 Dynamic Factor Models

Dynamic Factor Analysis of Environmental Systems I: Introduction and Initial lessons learned

RAFAEL MUÑOZ-CARPENA[*,‡], AXEL RITTER[†], DAVID KAPLAN[*]

[*]University of Florida, Gainesville, USA,
[†]Universidad de La Laguna, Tenerife, Spain
[‡]email: carpena@ufl.edu

329:Munoz-Carpena.tex,session:OCS10

Today's environmental problems are complex in nature and require a multidisciplinary approach where scientists and engineers collaborate in the solutions. In multidisciplinary research, success must be driven by the importance of the problem solved, rather than the elegance of the methods alone. Dynamic Factor Analysis (DFA) is an efficient dimension reduction technique that allows for the variance decomposition of multiple time series (response variables, RVs) into unexplained (common trends, CTs) and explained (explanatory variables, EVs) effects. Although non-stationarity may be overcome by statistical de-trending, the identification of non-stationary CTs may hold fundamental information about the system's temporal dynamics. This makes DFA particularly suited for environmental analysis. Márkus et al. (1999) introduced DFA in physical environmental sciences (hydrology) to study groundwater dynamics, with an initial analysis only considering CTs. The power of DFA was fully realized later by considering EVs in the analysis of groundwater quality dynamics (Muñoz-Carpena et al., 2005). Normalization of RVs and EVs and interpretation of the resulting EV regression coefficients allowed for separation and analysis of the influences of intrinsic/extrinsic factors on the environmental system. Using a well-defined physical system (regional drainage canals), Ritter and Muñoz-Carpena (2006) obtained a fully explanatory model by eliminating CTs and developing a spatial function of the EV regression coefficients to develop a dynamic groundwater model that matched the conceptual system and expensive numerical simulations (MODFLOW). Later, Ritter et al. (2007) included a model goodness-of-fit statistic (Nash and Sutcliffe efficiency) in the DFA model selection criteria to complement deficiencies found in the commonly used Akaike's Information Criteria (AIC). In this application, the t-test from standard errors was also used to assess the significance of EVs. Ritter et al. (2009) demonstrated that using DFA, the complex variability in multivariate environmental time series could be simplified without the need of a-priori detailed information about site-specific characteristics such as soil properties, vegetation cover, etc., and found DFA to be a useful scaling technique, whereby a (single) CT, together with spatially dependent regression parameters, reproduces time series of soil moisture at different locations within a watershed. Finally, Regalado and Ritter (2009) used DFA to identify CTs in a scattered dataset of soil water repellency (WR); these CTs served as a seed for a general WR-characterizing model, successfully applied to a large dataset of WR versus soil water content curves. The rich set of features of DFA has allowed our group's success in applying the method to important and varied environmental problems, with 18 scientific journal publications since our first application of the full method in 2005.

References

[Márkus el al.(1999)] L. Márkus, O. Berke, J. Kovács, W. Urfer. 1999. Spatial prediction of the intensity of latent effects governing hydrogeological phenomena. Environmetrics 10, 633-654.

[Muñoz-Carpena el al.(2005)] Muñoz-Carpena, R., A. Ritter, Y.C. Li. 2005. Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to Everglades National Park. J. Contam. Hydrol. 80: 49-70 .

[Regalado and Ritter (2009)] Regalado, C.M., A. Ritter. 2009. A bimodal four-parameter lognormal model of soil water repellency persistence. Hydrol. Process. 23: 881-892.

[Ritter and Muñoz-Carpena (2006)] Ritter, A., R. Muñoz-Carpena. 2006. Dynamic factor modeling of hydrological patterns in an agricultural area adjacent to Everglades National Park. J. Hydrol. 317: 340-354.

[Ritter el al.(2007)] Ritter, A., R. Muñoz-Carpena, D.D. Bosch, B. Schaffer, T.L. Potter. 2007. Agricultural land use and hydrology affect variability of shallow groundwater nitrate concentration in South Florida. Hydrol. Process. 21: 2464-2473.

[Ritter el al.(2009)] Ritter, A., C.M. Regalado, R. Muñoz-Carpena. 2009. CTs of topsoil water dynamics in a humid subtropical forest watershed. Vadose Zone J. 8: 437-449

Uniqueness and Non-Uniqueness for Stochastic Heat Equations with Hölder Continuous Coefficients

IS27
SPDE

LEONID MYTNIK*,†

*Technion - Israel Institute of Technology
†email: leonid@technion.ac.il 330:LEONID_MYTNIK.tex,session:IS27

We consider the question of uniqueness/non-uniqueness for stochastic partial differential equations (SPDEs). We focus on heat equations perturbed by a multiplicative noise, or the stochastic heat equations. Such equations with Hölder $1/2$ coefficients arise in population models. Does pathwise uniqueness hold for such equations? In 1971, the analogous question for SDE's was resolved in the affirmative by T. Yamada and S. Watanabe. As for stochastic heat equations we prove pathwise uniqueness in the case of Hölder coefficients of index $\gamma > 3/4$. We also consider the case of $\gamma < 3/4$, and discuss a number of open questions. These results are joint with Ed Perkins and Carl Mueller.

Consistency of Functional Data Depth

POSTER
Poster

STANISLAV NAGY*,†

*Charles University in Prague, Czech Republic
†email: nagy@karlin.mff.cuni.cz 331:StanislavNagy.tex,session:POSTER

Data depth is a modern tool for the analysis of multivariate (and infinite dimensional) data with a great perspective of usage in nonparametric statistics. An essential property that a reasonable depth function(al) must satisfy is the uniform consistency of its sample version. Without it, the key results such as the consistency of a multivariate analogy of a median (the point(s) with the highest depth value) or of the contours of depth cannot be derived.

In the contribution we focus on functional data depth and several different approaches of depth determination proposed in the literature. We show that a key result of authors López-Pintado and Romo (2009, Theorem 4) concerning the uniform consistency of the sample band depth *does not hold* (even under stronger assumptions), and we illustrate this proposition on a few easy counterexamples. We show the reason why the proof fails and propose two different methods of remedying for inconsistency. In the case of integral type depths we also generalize the known results of Fraiman and Muniz (2001, Theorem 3.1).

Acknowledgment. This work was supported by the grant P402/12/G097 from the Czech Science Foundation.

References

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

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

CS8A Bayesian Semipar.

Minimum Distance Estimators in Measurement Error Models

RADIM NAVRÁTIL[*,‡], HIRA L. KOUL[†]

[*]Charles University in Prague, Czech Republic,
[†]Michigan State University, East Lansing, USA
[‡]email: navratil@karlin.mff.cuni.cz 332:Navratil.tex,session:CS8A

Measurement error models (also called *errors-in-variables models*) are regression models that account for measurement errors in the independent variables (regressors). These models occur very commonly in practical data analysis, where some variables cannot be observed exactly, usually due to instrument or sampling error. Sometimes ignoring measurement error may lead to correct conclusions, however in some situations it may have dramatic consequences.

History of these models is very rich, it started at the end of the nineteenth century. Since then various methods for dealing with measurement errors were developed, such as least squares estimates, maximum likelihood estimates and most recently total least squares estimates. However, nonparametric methods are not very common, although they suit to this situation well due to the absence of knowledge of the measurement error's distribution.

We will introduce a class of minimum distance estimators into linear regression model with stochastic regressors that may be subject to measurement error. To do it we first consider estimates in model where regressors and model errors are not independent, but only uncorrelated. This result will be then extended into measurement error models. As a byproduct we also introduce a class of distribution-free rank tests for testing in measurement error models. All the theoretical results will be illustrated on examples and simulations. Suggested estimates and tests will be compared with the standard ones and will be shown their good performance, for some situation even better than classical approaches.

Acknowledgment. This research was supported by the Charles University Grant GAUK 105610 and by the Grant SVV 265 315.

OCS29 Stat. Branching Proc.

Online Change Detection in INAR(p) Models with General Offspring Distributions

FANNI NEDÉNYI[*,†]

[*]University of Szeged, Szeged, Hungary
[†]email: nedfanni@gmail.com 333:FanniNedenyi.tex,session:OCS29

Let us define $\{X_k,\ k \in \mathbb{N}\}$, an integer-valued autoregressive process of order p with fixed $p \in \mathbb{N}$, deterministic $X_{-p+1}, X_{-p+2}, \ldots, X_0 \in \mathbb{N}$ initial values and general offspring distribution. Therefore we consider the INAR(p) model given by

$$X_k = \sum_{i=1}^{X_{k-1}} \xi_1(k,i) + \cdots + \sum_{i=1}^{X_{k-p}} \xi_p(k,i) + \varepsilon(k), \quad k \in \mathbb{N}$$

where $\{\xi_\ell(k,i)\}_{i\in\mathbb{N}}$ is a sequence of independent, identically ditributed nonnegative random variables for all $1 \le \ell \le p$ and $\{\varepsilon(k)\}_{k\in\mathbb{N}}$ is the nonnegative sequence of innovations, such that all these sequences are independent of each other.

The change of the offspring or the innovation distribution of the INAR(p) process may cause disturbance in the model, so it is an important task to detect such changes. Our goal is to test the null hypothesis that the INAR(p) model does not change over time. In terms of applicability it is advantageous if the change detection is online, since collecting data is often costly. The basis of this sequential method is the one detecting parameter changes in linear regression models introduced by Horváth et al. (2004).

Based on a training sample we estimate the expected values by the conditional least squares method and using these estimations we construct a test to detect changes in the expected values. Although, as the offsprings have general distribution, changes can occur that do not affect the expected values. Therefore we introduce a test based on two-dimensional statistics, that also detects changes in the deviations. We define both tests as an open-end and also as a closed-end procedure, and we determine the limit distributions of the defined statistics in order to get the critical values for the tests. Also, in case of a simple alternative hypothesis we show that these tests are consistent.

Anomalous Shock Fluctuations in the Asymmetric Exclusion Process

OCS16 Interacting Particles

PATRIK FERRARI*, PETER NEJJAR*,†

*Bonn University, Germany
†email: nejjar@uni-bonn.de 334:PeterNejjar.tex,session:OCS16

We consider the totally asymmetric simple exclusion process (TASEP) with particles occupying every site of $2\mathbb{Z}$ at time $t = 0$. Particles starting from $2\mathbb{Z}_+$ have jump rate $\alpha < 1$ and particles starting from $2\mathbb{Z}_-$ have jump rate 1. This generates a macroscopic shock moving (to the left) with speed $-(1-\alpha)/2$. Since TASEP belongs to the Kardar-Parisi-Zhang (KPZ) universality class, one could expect that the shock fluctuations live on a scale of order $t^{1/3}$, where t is the time. We determine the large time fluctuations of particles around the shock region, which is given in terms of products of the GOE Tracy-Widom F_1 distribution functions from random matrices. In particular, by varying the particle number on the $t^{1/3}$ scale, one the distribution function interpolates between F_1 and F_1. This comes from the fact that away from the shock one has the Airy_1 process on both sides.

On Sociological Application of Discrete Marginal Graphical Models

CS17A Causal Inference

RENÁTA NÉMETH*,†, TAMÁS RUDAS*

*Eötvös Loránd University, Budapest, Hungary
†email: nemethr@tatk.elte.hu 335:RenataNemeth.tex,session:CS17A

Graphical models are defined by general and possibly complex conditional independence assumptions and are well suited to model direct and indirect associations and effects that are of central importance in many problems of sociology. The talk provides a unified view of many of the graphical models discussed in a largely scattered literature. The marginal modeling framework proposed here relies on parameters that capture aspects of associations among the variables that are relevant for the graph and, depending on the substantive problem at hand, may lead to a deeper insight than other approaches. In this context, model search, which uses a sequence of nested models, means the restriction of increasing subsets of parameters. As a special case, general path models for categorical

data are introduced. The method is applied to the social status attainment process in the USA, Hungary and Czechoslovakia at the end of the last century, and shows that policies in the latter socialist countries to prevent status inheritance had little success.

Acknowledgment. RN's work was supported by the János Bólyai Research Scholarship from the Hungarian Academy of Sciences. Both RN's and RT's work was partialy supported by the Hungarian Scientific Research Fund (OTKA K106154).

OCS32 Valuation in Stoch. Fin.

Liquidity Risk and Price Impact in Continuous Trading

GÁBOR LOI NGUYEN*,†, LÁSZLÓ MÁRKUS*

*Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest, Hungary
†email: gaborloi@gmail.com 336:Nguyen_Gabor.tex,session:OCS32

We provide a model for liquidity risk and price impacts within a limit order book setting. Following the insights of Kyle(1985), we extend a model of Cetin el al.(2004) with additional dimensions of liquidity. Combining their model with the price impact function introduced by Bouchaud el al.(2003) an opportunity to examine the necessary conditions to obtain an arbitrage free model arise. Using our approach the model of Roch(2011), which is closely related to our topic, could be seen from a different point of view. By making a clear distinction between the different types of liquidity, the model becomes flexible and the results remain valid under numerous different settings. Additionally we provide an example to show the main difficulties in the derivative asset pricing with price impacts. Even though achieving full scale solutions to these problems are beyond the limits of this work, we give an approximation to the price of the European Call option.

CS38A Appl. Multivariate Tech.

Multivariate Statistical Techniques and Analytical Hierarchical Procedure in Supplier Selection for Military Critical Items

SOCRATES J. MOSCHURIS*, CHRISTODOULOS NIKOU*,†

*Department of Industrial Management and Technology, University of Piraeus, Greece.
†email: cnikou@unipi.gr 337:CHRISTODOULOS.tex,session:CS38A

Logistics in a military sense include, among others, aspects of military operations which deal with the acquisition of parts, material and services acting as a force multiplier that attains the advantage from a given force configuration by increasing the timeliness and endurance of the force. In order to support effectively that kind of operations at the strategic level, special care should be given to procurement process, in order to ensure the availability of logistics resources in a timely manner. Military Critical Items (MCI) are items with a major impact on the safety and the accomplishment of a mission (ACSIMH, 2005), and, from a supply positioning model perspective, they may include critical and bottleneck items. Therefore, it is crucial to adopt a good supplier selection model/approach. The importance of creating such a model is also acknowledged by a report (available unrestricted in web) of the Chief of Hellenic Forces which in the 17 decisive points of his operational standard guidelines includes several factors/points that directly depend on a successful procedure of that kind. In this work, we evaluate real data collected through confidential questionnaires filled in by members of the Armed Forces. Initially, we use some summary statistical tools to examine the shape and spread of sample data. Then, by using Cluster Analysis (CA), a technique retrieved from the Multivariate Statistical Analysis area (Johnson and Wichern, 2007), we try to group and present in a simpler way 17 factors cited in the questionnaire that could complicate an MCI procurement process. After a review of the supplier selection criteria literature, the clusters of the abovementioned factors may be

linked with criteria that in our opinion could safeguard the procurement phases from their occurrence, thereof facilitating the supplier selection process. CA is also used to examine if a relationship exists between the 17 factors and the answers in 2 significant questions set also in the questionnaire, which relate to outsourcing procedures and the percentage of unexpected demand that a potential supplier should be able to provide in MCI cases. Pearson product moment correlation coefficient is also used to measure the degree of linear relationship between selected pairs of variables.

After the identification of the criteria to be used in the model, we created an expert team (Dagdeviren et al., 2009) and asked them to rate the selected criteria. Then, we applied Analytical Hierarchy Process (AHP), a robust technique, to an hypothetical number of 4 suppliers, in order to increase the efficiency of attributing weights to the criteria and get the final score that will determine the most appropriate supplier. "Expert Choice" software may be used as a tool to assess the consistency of the model.

References

[Dagdeviren et al. (2009)] Dagdeviren M., Yavuz S., Kilinc N., 2009 : Weapon Selection using the AHP and TOPSIS methods under Fuzzy environment *Expert Systems with Applications* , **36**, 8143–8151.

[Johnson and Wichern (2007)] Johnson R., Wichern D., 2007: Applied Multivariate Statistical Analysis, *Pearson Prentice Hall, New Jersey, USA* .

[JLCD (2005)] Joint Logistics Commanders/Department of the US Navy, 2005: ACSIMH-Aviation Critical Safety item Management Handbook, *Patuxent River, USA* .

Score-Based Methods for Causal Inference in Additive Noise Models

CS17A Causal Inference

CHRISTOPHER NOWZOHOUR*,†, PETER BÜHLMANN*

*ETH Zürich, Switzerland

†email: nowzohour@stat.math.ethz.ch 338:ChristopherNowzohour.tex,session:CS17A

Given data sampled from a number of variables $X_1, \ldots, X_p$, one is often interested in the causal structure connecting the variables, in the form of a directed acyclic graph. Real-world examples are gene expression data in biology or financial time series data in economics. In the general case, without interventions on some of the variables it is only possible to identify the graph up to its Markov equivalence class. However, in many situations one can find the true causal graph D just from observational data. This is possible e.g. in the case where the data is generated by a structural equation model with additive noise. As an example, consider $X_i = f_i(X_{\mathbf{pa}_i}) + \epsilon_i$ for some (nonlinear) edge functions f_i, with $\mathbf{pa}_i$ being the parents of node i in D and ϵ_i are i.i.d. However, most methods for achieving this rely on non-parametric independence tests, which are computationally expensive. A second problem with this testing approach is that the null hypothesis is independence, which is what one would like to get evidence for (while testing is suitable for finding evidence against the null hypothesis). We take a different approach in our work by using a penalized likelihood as a score for model selection. This is practically feasible in many settings and has the advantage of yielding a natural ranking of the candidate models. Provided the model class is identifiable and making mild smoothness assumptions on the density space, we can show that the method asymptotically identifies the correct model. We also provide illustrations by means of simulations with randomly generated edge functions. Applying our method to real-world 2-dimensional data sets (cause-effect pairs) yields similar results as other state-of-the-art methods.

NYA
Not Yet Arranged

Asymptotic Expansions with Monotonicity

HARUHIKO OGASAWARA[*,†]

[*]Otaru University of Commerce, Otaru, Japan
[†]email: hogasa@res.otaru-uc.ac.jp 339:HaruhikoOgasawara.tex,session:NYA

General formulas of the asymptotic cumulants of a studentized parameter estimator are given up to the fourth order with the added higher-order asymptotic variance. Using the sample counterparts of the asymptotic cumulants, formulas for the Cornish-Fisher expansions with the third-order accuracy are obtained. Although the results are asymptotically valid, it is known that the endpoints of the confidence intervals given by the Cornish-Fisher expansions are not monotonic functions of the value of the studentized estimator, which gives unpleasant results in practice.

Hall (1992) gave a monotonic cubic transformation of the studentized estimator with the second-order accuracy. The asymptotic cumulants of Hall's transformed studentized estimator up to the fourth-order with the added higher-order asymptotic variance are also given in this presentation. For the third-order accurate monotonic transformations, Fujioka and Maesono (2000) and Yanagihara and Yuan (2005) obtained monotonic quintic transformations of Hall's transformation. That is, their transformations are of order 15 in terms of the power of the original studentized estimator.

Some new methods of monotonic transformations of the studentized estimator are presented. Based on Fujioka and Maesono (2000) and Yanagihara and Yuan (2005), quintic transformation of the original studentized estimator with the third-order accuracy are given. In addition, similar transformations of a fixed normal deviate are proposed up to the same order with some asymptotic comparisons to the transformations of the studentized estimator. Further, a monotonic transformation using the studentized estimator or the normal deviate depending on a statistic is also proposed, which is called as a switching method. These transformations are cubic transformations of Hall's one and the corresponding transformation of the normal deviate. Applications to a mean and a binomial proportion are shown with a numerical illustration for estimation of the proportion. The full results corresponding to this presentation are given by Ogasawara (2012, JMVA, 103, 1-18).

OCS11
ENBIS

Discussion of the Integration of Info(Q) and PSE in a DMAIC Framework to Determine Six Sigma Training Needs

CHRISTOPHER MCCOLLIN[*], IRENA OGRAJENŠEK[†,‡]

[*]School of Science and Technology Nottingham Trent University, United Kingdom
[†]Faculty of Economics, University of Ljubljana, Slovenia
[‡]email: irena.ograjensek@ef.uni-lj.si 340:Irena_Ograjensek.tex,session:OCS11

Practical Statistical Efficiency and InfoQ are two tools to help determine the effectiveness of improvement programmes. These can be mapped into a Six Sigma Define-Measure-Analyse- Improve-Control framework which highlights four main issues of project success. These are project selection and control by senior management, project management and teamwork, statistical planning and analysis and the learning process. These four issues tie in with Deming's System of Profound Knowledge: understanding systems, variation, people and knowledge. From the perspective of team dynamics being a pre-requisite of project success, we use Belbin's team roles and the Margerison-McCann wheel (taken from the Briggs-Myers personality test) to identify the characteristics of project team members which will then enable identification of training needs of statisticians in a Six Sigma environment.

Approximations in the Susceptible-Infectious-Removed Epidemic Model

POSTER
Poster

CHANGHYUCK OH*,†

*Yeungnam University, Gyeongsan, Korea
†email: choh@yu.ac.kr 341:ChanghyuckOh.tex,session:POSTER

A simple approximation for the maximum likelihood estimates of infection and removal parameters used in the susceptible-infectious-removed (SIR) epidemic model is presented. This approximation can be applied when the numbers of susceptible and infected individuals are observable only at discrete points in time. Since, in such cases, a closed form of the likelihood function is generally too complicated to obtain, the proposed approximation method represents an important advance. Simulation results show that the method yields approximations quite close to the maximum likelihood estimates obtained under continuous observation.

References

[Bailey (1975)] Bailey, N.T.J., 1975: *The Mathematical Theory of Infec-tious Diseases and its Applications*, revised Ed. Griffin, London.

[Becker and Britton (1999)] Becker, N.G. and Britton, T., 1999: Statistical studies of infectious disease incidence. *J. R. Statist. Soc. B*, **61**, 287 - 307.

[Cauchemez and Ferguson (2008)] Cauchemez, S. and Ferguson, N., 2008: Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *J. R. Soc. Interface*, **5**, 885 - 897.

[Oh et al. (1991)] Oh, C., Severo, N.C., and Slivka, J., 1991: Approximation to the maximum likelihood estimate in some pure birth process. *Biometrika*, **78**, 295 - 299.

Convergence of Weighted Sums of Random Variables

CS19C
Lim. Thms. Sums of RVs

TONGUÇ ÇAĞIN*,†, PAULO EDUARDO OLIVEIRA*

*CMUC, Dep. Mathematics, University of Coimbra, Portugal
†email: tonguc@mat.uc.pt, paulo@mat.uc.pt 342:Oliveirape.tex,session:CS19C

We consider the convergence of $T_n = \sum_{i=1}^n a_{n,i} X_i$, normalized by $n^{1/p}$, where the X_i's are associated and identically distributed with finite absolute moments of order $p \in (1,2)$. For constant weights and independent variables it is well known that $n^{-1/p} T_n \longrightarrow 0$ almost surely if and only if $\mathrm{E}(|X_1|^p) < \infty$. The sufficiency result was extended to associated variables by Louhichi [3], who proved the same convergence assuming besides the existence of the moment of order p, an integrability condition on the covariances of truncated variables: with $g_M(u) = \max(\min(u, M), -M)$, $\overline{X}_n = g_M(X_n)$, the truncated variables, and $G_{i,j}(M) = \mathrm{Cov}(\overline{X}_i, \overline{X}_j)$, then one should assume that

$$\sum_{1 \le i < j < \infty} \int_{j^{1/p}}^{\infty} v^{-3}\, G_{i,j}(v)\, dv < \infty.$$

Weighted sums require an extra control on the weighting sequences. Following Cuzick [2] and Bai and Cheng [1] we assume that $\limsup_{n \to +\infty} n^{-1/\alpha} \left(\sum_{i=1}^n |a_{ni}|^\alpha\right)^{1/\alpha} < \infty$, with $\alpha > 0$. For these weighted sums Bai and Cheng [1] proved the convergence for independent variables assuming the existence of a moment of order β such that $1/p = 1/\alpha + 1/\beta$. We extend these characterizations

to weighted sums of associated random variables. Assuming a suitable behaviour on the weighting sequences, in order to keep association of the random variables, and $\alpha > \frac{2p}{2-p}$ we prove that $n^{-1/p} T_n \longrightarrow 0$ almost surely when the moment of order $p\frac{\alpha-2}{\alpha-2p}$ exists and

$$\sum_{1\leq i<j<\infty} \int_{j^{(\alpha-2p)/(\alpha p)}}^{\infty} v^{-3-2p/(\alpha-2p)}\, G_{i,j}(v)\, dv < \infty.$$

This integrability assumption can still be weakened, replacing the polynomial term by $v^{-\beta}$, with $\beta > 1$, if we assume the tail probabilities $\mathrm{P}(X_i \geq x, X_j \geq y) - \mathrm{P}(X_i \geq x)\mathrm{P}(X_j \geq y)$ decrease fast enough at infinity.

Acknowledgment. This research was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0324/2011.

References

[1] Bai, Z., Cheng, P., 2000: Marcinkiewicz strong laws for linear statistics, *Statist. Probab. Lett.* **46**), 105–112.

[2] Cuzick, J., 1995: A Strong of Large Numbers for Weighted Sums of I.I.D. Random Variables, *J. Theoret. Probab.* **8**, 625–641.

[3] Louhichi, S., 2000: Convergence rates in the strong law for associated random variables, *Probab. Math. Statist.* **20**, 203–214.

[4] Oliveira, P.E., 2012: Weighted sums of associated variables, *J. Kor. Statist. Soc.* **41**, 537–542.

POSTER
Poster

An R Package for Fitting Generalized Waring Regression Models

MARÍA JOSÉ OLMO-JIMÉNEZ*,‡, JOSÉ RODRÍGUEZ-AVI*, ANTONIO JOSÉ SÁEZ-CASTILLO*, SILVERIO VÍLCHEZ-LÓPEZ†

*Department of Statistics and Operations Research, University of Jaén, Spain,
†IES Las Fuentezuelas, Jaén, Spain
‡email: mjolmo@ujaen.es

343:OlmoJimenez.tex,session:POSTER

Rodríguez-Avi et al. (2009) present the generalized Waring regression model (*GWRM*) for overdispersed count data as an alternative to the negative binomial regression model. The former is based on the univariate generalized Waring distribution (*UGWD*) that arises when one of the parameters of the negative binomial distribution is considered a random variable. This fact allows the observed variability to be split into three components: randomness, internal differences between individuals and the presence of other external factors that have not been included as covariates in the model. So, the extra variability of data or overdispersion can be modelled with more flexibility and the origin of this overdispersion can even be explained.

The majority of the regression models for count data has already been implemented in the statistical computing environment R (R Development Core Team, 2012). Thus, for example, the function `glm()` in the stats package provides fits of Poisson model in the context of generalized linear models and the functions `glm.nb()` in the MASS package and `negbin()` in the aod package do the same for the negative binomial regression model.

In this vein and with the aim of extending the use of the *GWRM*, we implemented a package for R, also called `GWRM`. This package is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/GWRM.

In this paper, we discuss a new implementation of the `GWRM` package in the function `GWRM.fit()` and we introduce new functions for diagnostic and inference. The design of these functions as well as

the methods operating on the associated fitted model objects follows that of the base R functionality. Firstly, we give an introduction to the structure of the package including a brief description of the theoretical model that has been implemented and secondly, we demonstrate usage of the package using a data set from the health field.

References

[R Development Core Team (2012)] R Development Core Team, 2012: R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

[Rodríguez-Avi et al. (2009)] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J., Olmo-Jiménez, M.J., Martínez-Rodríguez, A.M., 2009: A generalized Waring regression model for count data, *Computational Statistics and Data Analysis*, **53**, 3717 - 3725.

A Logistic Regression Analysis of Malaria Control Data

POSTER
Poster

JAMIU SHEHU OLUMOH[*,†], OSHO OLUSHOLA AJAYI[*]

[*]American University Of Nigeria, Yola, Nigeria
[†]email: jamiu.olumoh@aun.edu.ng

344:Jamiuolumoh.tex,session:POSTER

This study investigates the effect of various methods of malaria control including mosquito net. A logistic regression fitted to the data suggested the level of risk involves for a respondent who failed to use mosquito net. The results show the odds of contracting malaria is about 2.788 times higher for a person who did not use mosquito net compared to a person who used it. Also, the odd of contracting malaria in urban area is almost twice that of the rural area. Age plays little or no effect on malaria status.

Acknowledgment. We are grateful to Tyndall and her team for affording us the opportunity to using the data.

A RKHS for Improper Complex Signals

POSTER
Poster

ANTONIA OYA[*,†], JESÚS NAVARRO-MORENO[*], JUAN C. RUIZ-MOLINA[*], ROSA M. FERNÁNDEZ-ALCALÁ[*]

[*]Department of Statistics and Operations Research, University of Jaén, Spain
[†]email: aoya@ujaen.es

345:Antonia_Oya.tex,session:POSTER

The theory of Reproducing Kernel Hilbert Space (RKHS) provides an unified framework for a wide range of problems in statistical signal processing. Specifically, kernel-based processing has shown to be an efficient tool for addressing non-linear problems in areas such as adaptive filtering, image processing, and modeling of motion. Although the theory of RKHS holds for complex spaces too, most of the kernel-based techniques have been only developed to process real-valued random signals. Recently, the kernel-based approach for treating complex-valued random signals has drawing increasing interest in the area of statistical signal processing [Bouboulis et al. (2012)].

On the other hand, complex-valued random signals appear in several fields such as electromagnetics, communications, optics and many others, in order to model physical effects where two random components are involved. In general, complex-valued random signals $s(t)$ are improper, that is, they are correlated with their complex conjugates $E[s(t)s(\tau)] \neq 0$ and thus the correlation function, $E[s(t)s^*(\tau)]$, does not provide a complete second-order description of the signal. In order to use the full statistical information of an improper complex-valued signal, we need to examine properties of both the correlation and complementary correlation functions i.e., to consider the so-called augmented statistics [Mandic and Su Lee Goh (2009), Schreier and Scharf (2010)]. The widely linear

(WL) processing approach, which is based on the augmented signal, $[s(t), s^*(t)]'$, has been shown to yield significant improvements in most areas of statistical signal processing such as estimation, modeling, detection, equalization, simulation, etc. [Mandic and Su Lee Goh (2009)].

In this paper, we focus our interest on the application of the WL perspective to the construction of a RKHS for complex-valued improper random signals. An explicit description of the structure of the complex RKHS and the expression of the associated inner product could allow us to obtain alternative solutions to different signal processing problems, for instance, detection problems. Every correlation function is also a reproducing kernel for some RKHS. As a consequence, there is a close connection (an isometric isomorphism, in fact) between the closed linear span of a random signal and the RKHS determined by its correlation function. Solutions to several statistical signal processing problems can then be expressed in terms of an appropriate RKHS inner product.

References

[Bouboulis et al. (2012)] Bouboulis, P., Theodoridis, S., Mavroforakis, M., 2012: The augmented complex kernel LMS, *IEEE Trans. Signal Proc.*, **60**(9), 4962 - 4967.

[Mandic and Su Lee Goh (2009)] Mandic, D. P., Su Lee Goh, V., 2009: *Complex valued nonlinear adaptive filters: noncircularity, widely linear, and neural models*, United Kingdom: Wiley.

[Schreier and Scharf (2010)] Schreier, P. J., Scharf, L. L., 2010: *Statistical signal processing of complex-valued data*, Cambridge University Press.

OCS4
3D Images

3D Projective Shapes of Leaves from Image Data

VICTOR PATRANGENARU*, ROBERT L. PAIGE†, MINGFEI QIU*

*Florida State University, Tallahassee, Florida, U.S.A.,
†Missouri University of Science and Technology, Rolla, Missouri, U.S.A.

346:RobPaige.tex,session:OCS4

We consider three dimensional image analysis of objects which may be more or less planar. We develop a nonparametric test of planarity for objects where regular camera pictures have been taken. Here the projective shapes are treated as points on projective shape manifolds. We also consider the three dimensional reconstruction for leaves which are not planar. We illustrate our methodology on a data set consisting pictures of leaves suspended in midair.

CS11A
SDE-s

Closed-Form Likelihood Approximation for Parameter Estimation of Non-Linear SDEs

PHILLIP PAINE*,†, SIMON PRESTON*, ANDREW WOOD*

*University of Nottingham, England
†email: pmxpp@nottingham.ac.uk

347:PhillipPaine.tex,session:CS11A

In fields such as finance, biology and engineering, a natural modelling approach involves using diffusion processes defined in terms of stochastic differential equations (SDEs). Our focus is on the case where the observations are sampled discretely. Our interest is in the question of how to estimate the parameters of an SDE based on discrete observations.

One approach is to use maximum likelihood estimation. One can express the likelihood function in terms of the transition density between consecutive observations and in theory this allows us to calculate the MLE. However, for all but a small number of SDE models there is no closed-form expression for the transition density.

Our approach is to derive an approximation to the transition density, which will be accurate when the interval between observations is relatively small. We can then form an approximate likelihood function for which we are able to compute the MLE. Our method involves using a suitable expansion to approximate the unknown transition density of a truncated Ito-Taylor expansion of the sample-path, and use this as an approximation to the true transition density. We have found in numerical simulations that this method performs favourably compared to existing methods, and possesses a number of other useful properties.

Acknowledgment. This research was partially supported by the UK Engineering and Physical Sciences Research Council.

CLT For Linear Spectral Statistics of Normalized Sample Covariance Matrices with Larger Dimension and Small Sample Size

OCS13 H-D Stat, R. Matrices

BIN BIN CHEN, GUANGMING PAN*

Nanyang Technological University, Singapore

*email: gmpan@ntu.edu.sg 348:Pan.tex,session:OCS13

Let $\mathbf{A} = \frac{1}{\sqrt{np}}(\mathbf{X}^T\mathbf{X} - p\mathbf{I}_n)$ where $\mathbf{X}$ is a $p \times n$ matrix, consisting of the independent and identically distributed $(i.i.d.)$ real random variables X_{ij} with mean zero and variance one. When $p/n \to \infty$, under fourth moment conditions the central limit theorem (CLT) for linear spectral statistics (LSS) of $\mathbf{A}$ defined by the eigenvalues is established. We also explore its applications in testing whether a population covariance matrix is an identity matrix.

Acknowledgment. This research was partially supported by the Ministry of Education, Singapore, grant No.: ARC14/11.

References

[1] Bai, Z. D. and Yao, J. F., 2005: On the convergence of the spectral empirical process of Wigner matrice, *Bernoulli*, **11**, 1059-1092.

[2] Bai, Z. D. and Yin, Y. Q. (1988) Convergence to the semicircle law. *Ann. Probab.* **16**, 863-875.

[3] Chen, B. B. and Pan, G. M. (2012) Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, 18, 1405-1420.

[4] Karoui, E. N. (2003) On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. arXiv: math/0309355.

[5] Ledoit, O. and Wolf, M. (2002) Some Hypothesis Tests for the Covariance Matrix When the Dimension Is Large Compare to the Sample Size. *The Annals of Statistics*, **30** 1081-1102.

[6] Marčenko, V. A. & Pastur, L. A. (1967) Distribution for some sets of random matrices. *Math. USSR-Sb.*, **1** , 457-483.

Doubly Spectral Analysis of Stationary Functional Time Series

IS9 Functional Time Ser.

VICTOR PANARETOS*,†

*Department of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL)

†email: victor.panaretos@epfl.ch 349:PanaretosVictor.tex,session:IS9

The spectral representation of random functions afforded by the celebrated Karhunen-Loève (KL) expansion has evolved into the canonical means of statistical analysis of independent functional data: allowing technology transfer from multivariate statistics, appearing as the natural means of regularization in inferential problems, and providing optimal finite dimensional reductions. With the aim

of obtaining a similarly canonical representation of dependent functional data, we develop a doubly spectral analysis of a stationary functional time series, decomposing it into an integral of uncorrelated functional frequency components (Cramér representation), each of which is in turn expanded into a KL series. This Cramér-Karhunen-Loève representation separates temporal from intrinsic curve variation, and it is seen to yield a harmonic principal component analysis when truncated: a finite dimensional proxy of the time series that optimally captures both within and between curve variation. The construction is based on the spectral density operator, the functional analogue of the spectral density matrix, whose eigenvalues and eigenfunctions at different frequencies provide the building blocks of the representation. Empirical versions are introduced, and a rigorous analysis of their large-sample behaviour is provided. (Based on joint work with S. Tavakoli, EPFL).

References

[1] Panaretos, V.M. & Tavakoli, S. (2013). Cramér-Karhunen-Loève Representation and Harmonic Principal Component Analysis of Functional Time Series. *Stochastic Processes and their Applications, 123 (7): 2779–2807*.

[2] Panaretos, V.M. & Tavakoli, S. (2013). Fourier Analysis of Stationary Time Series in Function Space. *Annals of Statistics, 41 (2): 568–603*.

OCS28 Stat. Affine Proc.

Asymptotic Behavior of Critical Multi-Type Continuous Time Branching Processes with Immigration

MÁTYÁS BARCZY*, ZENGHU LI†, GYULA PAP‡,§

*University of Debrecen, Debrecen, Hungary,
†Beijing Normal University, Beijing, People's Republic of China,
‡University of Szeged, Szeged, Hungary
§email: papgy@math.u-szeged.hu

350:GyulaPap.tex,session:OCS28

Under natural assumptions a Feller type diffusion approximation is derived for critical, positively regular multi-type continuous time branching processes with immigration (CBI processes). Namely, it is proved that a sequence of appropriately scaled random step functions formed from a critical, positively regular multi-type CBI process converges weakly towards a squared Bessel process supported by a ray determined by the Perron vector of a matrix related to the branching mechanism of the CBI process.

The result is a counterpart of a convergence theorem for critical, primitive multi-type Galton–Watson branching processes with immigration and for unstable integer valued autoregressive processes (INAR(p) processes), see [Ispány and Pap (2012)] and [Barczy et al. (2011)], respectively. The cases of not necessarily positively regular multi-type CBI processes will also be discussed.

Acknowledgment. The authors have been supported by the Hungarian Chinese Intergovernmental S & T Cooperation Programme for 2011-2013 under Grant No. 10-1-2011-0079. M. Barczy and G. Pap have been partially supported by the Hungarian Scientific Research Fund under Grant No. OTKA T-079128. Z. Li has been partially supported by NSFC under Grant No. 11131003 and 973 Program under Grant No. 2011CB808001.

References

[Ispány and Pap (2012)] Ispány, M. and Pap, G., 2012: Asymptotic behavior of critical primitive multi-type branching processes with immigration, arXiv: 1205.0388.

[Barczy et al. (2011)] Barczy, M., Ispány, M. and Pap, G., 2011: Asymptotic behavior of unstable INAR(p) processes, *Stochastic Process. Appl.*, **121(3)**, 583 - 608.

Multi-agent Statistical Relational Learning

CS3A Machine learning

DIMITRI PAPADIMITRIOU[*,‡], PIET DEMEESTER[†]

[*]Alcatel-Lucent Bell Labs, Antwerp, Belgium,
[†]Ghent University and iMinds, Gent, Belgium
[‡]email: dimitri.papadimitriou@alcatel-lucent.com, piet.demeester@intec.ugent.be
351:DimitriPapadimitriou.tex,session:CS3A

Several learning techniques have been proposed to enable adaptivity, predictivity and autonomy of multi-agent control processes. In this context, on-line methods for cooperative multi-agent learning from streams of data are considered to (proactively and/or reactively) adapt agents'decisions over time and to predict the emergence of new situations to be handled. However, commonly envisaged statistical learning techniques assumes that propositional data are identically and independently distributed and that random sample of homogeneous objects results from single relation whereas real world data sets, in particular, those characterizing real distributed settings are not identically distributed (heterogeneous) and not independent (showing complex multi-relational structures). On the other hand, most relational learning techniques neither assume noise nor uncertainty in data whilst real world data are characterized by distributions that show presence of uncertainty and noise. Finally, cooperative multi-agent learning raises multiple challenges depending on the degree of autonomy of the agents (from individual learning agents to single macro-agent), the spatial range of agents'interactions, and the degree of external supervision.

Statistical Relational Learning (SRL) also referred to as Probabilistic Logic Learning, combines relational logic learning to model complex relational structures and inter-dependencies properties in data with probabilistic graphical models (such as Bayesian networks or Markov networks) to model the uncertainty on data; the resulting process can perform robust and accurate learning about complex relational data. The goal of SRL is of particular for multi-agent systems learning from hidden dependencies between multi-relational, heterogeneous and semi-structured but also noisy and uncertain data. The ultimate objective of SRL is indeed to learn from non-independent and identically distributed data "as easily as" from independent and identically distributed data. SRL is particularly adapted to information extraction; it provides better predictive accuracy and understanding of domains. However, this technique induces a harder learning task and higher complexity. SRL is nowadays applied to social network analysis, hypertext and web-graph mining, etc. it is thus appropriate to exploit some of its potentials to multi-agent control processes as i) the models learned from both intrinsic (propositional) and relational information perform better than those learned from intrinsic information alone, ii) probabilistic relational models offer significant advantages over deterministic relational models (including better predictive accuracy and better understanding of relational structure in heterogeneous data set), iii) SRL algorithms can learn accurate models of (inter-)dependent data instances.

Recent results obtained by exploiting SRL techniques and extending them to time-changing relational data show significant advantages in multi-agent control performing distributed learning tasks associated to, e.g., self-diagnosis, self-optimization, self-stabilization, and self-healing. These benefits are of particular interest when adaptive, predictive and autonomous decisions enable in response to environmental changes (or changes in the interacting parts) to adjust and/or modify the configuration and operations (by sensing and analyzing new environmental conditions over time) so that further behaviour better suits the environment and is able to maintain or even improve performance over time. Comparative analysis of the performance obtained on representative use cases show that the additional model complexity is actually warranted compared to models learned from intrinsic information only.

CS25B Risk Mgment

A New Method for Dynamic Panel Models: Applied to Stress Testing

SAVAS PAPADOPOULOS*,†

*Bank of Greece,
†Democritus University of Thrace

352:SavasPapadopoulos.tex,session:CS25B

Among various risk management techniques, stress testing (ST) has become extremely popular and useful. In stress testing, vulnerability assessments are conducted to banks under extreme but feasible macroeconomic scenarios. A dynamic panel data (DPD) model for is built on data from EU banks appropriate for stress testing. DPD models were estimated using several methods including a novel one. The new method is based on a proved theorem and has various advantages in comparison to standard methods, especially to small samples for N and to unbalanced data sets. The models are simulated to assess the bias and the variability of the estimates. The new method excels the standard methods in many cases and it can be applied when the other methods cannot be executed or cannot provide satisfactory results due to small N. In the application to ST, standard methods do not give the desirable estimates due to the highly unbalanced data set, while the new method performs satisfactory.

Nonparametric Inference for Covariate-Adjusted Summary Indices of ROC Curves

JUAN CARLOS PARDO-FERNÁNDEZ*,§, ELISA M. MOLANES-LÓPEZ†, EMILIO LETÓN‡

*Universidade de Vigo, Spain,
†UC3M, Madrid, Spain,
‡UNED, Madrid, Spain
§email: juancp@uvigo.es

353:pardofernandez.tex,session:CS31A

In medical studies, the diagnostic of a patient is very often based on some characteristic of interest, which may lead to classification errors. These classification errors are calibrated on the basis of two indicators: sensitivity (probability of diagnosing an ill person as ill) and specificity (probability of diagnosing a healthy person as healthy).

When the diagnostic variable is continuous, the classification will necessarily be based on a cut-off value: if the variable exceeds the cut-off then the patient is classified as ill, otherwise the patient is classified as healthy. In this situation, it is of special interest the geometrical locus obtained when varying the cut-off values in the complement of the specificity versus the sensitivity. This geometrical locus is called the receiver operating characteristic curve (ROC curve), and it is of extensive use to analyse the discriminative power of the diagnostic variable. Some summary indicators, such as the area under the curve or Youden index, are used to describe the main features of the ROC curve.

In many studies, a covariate is available along with the diagnostic variable. The information contained in the covariate may modify the discriminatory capability of the ROC curve, and therefore it is interesting to study the impact of the covariate on the conditional ROC curve. This work will be devoted to the study of a nonparametric estimation procedure of the covariate-adjusted ROC curve and its associated summary indices, especifically, the covariate-adjusted AUC and the covariate-adjusted Youden index.

A data set concerning diagnosis of diabetes will be used as an illustration of the proposed methodology.

Frenet-Serret Framework for the Analysis of Multi-Dimensional Curves

OCS19 Multivar. funct. data

JUHYUN PARK[*,‡], NICOLAS BRUNEL[†]

[*]Lancaster University, Lancaster, U.K.,
[†]Laboratoire Statistique et Genome, Université d'Evry, France
[‡]email: juhyun.park@lancaster.ac.uk 354:JuhyunPark.tex,session:OCS19

It is increasingly common to observe multi-dimensional curves in biological and medical applications. They are often associated with positioning of moving objects in space but more generally they could be simultaneously observed curves that are related to each other. Understanding and characterising variability of the curves is still the main interest of the analysis, yet, extending the standard framework of analysing one dimensional curves through curve alignment and functional principal component analysis is not trivial. For example it may be possible to extend the covariance function for one dimensional curves to a cross-covariance function for multi-dimensional curves to extract main variability of the curves but interpretation of the results could be of an issue.

We are interested in simultaneous analysis of the multi-dimensional curves based on features of the curves contained in higher order derivatives. This could be done through a geometric framework known as Frenet-Serret equation. It is known that general smooth curves in space (i.e R^3, or higher dimension) can be represented with Frenet-Serret (moving) frames. These are based on features of the curve expressed through curvature and torsion of a curve, and are expressed through an ordinary differential equation.

We use this framework to estimate a mean curve that exploits geometric features of the curve, and then to measure variability with respect to the mean curve. Moreover we use this framework for defining shape and phase variability, in order to address curve alignment of multidimensional data that preserves curvature (and torsion).

Two Sample Tests for Mean 3D Projective Shapes of Surfaces from Digital Camera Images

OCS4 3D Images

VIC PATRANGENARU[*,†]

[*]FSU, Tallahassee, FL 32308, USA
[†]email: vic@stat.fsu.edu 355:VicPatrangenaru.tex,session:OCS4

It is known that for $k \geq 5$, a manifold of projective shapes of k-ads in 3D has a structure of a $3k-15$ dimensional Lie group that is equivariantly embedded in an Euclidean space. Therefore testing for mean projective shape difference amounts to a two sample test for extrinsic means on this Lie group. On the other hand, emulating human vision, in absence of occlusions, the 3D projective shape of a surface can be recovered from a stereo pair of images, thus allowing to test for mean 3D projective shape difference of two surfaces.

Acknowledgment. Research partially supported by award NSF-DMS-1106935.

CS5D
H-D
Inference

Feature Thresholding in High-Dimensional Supervised Classifiers

TATJANA PAVLENKO*,‡, ANNIKA TILLANDER†

*Dep. of Mathematics, KTH Royal Institute of Technology, Sweden,
†MEB, Karolinska Institute, Sweden
‡email: pavlenko@math.kth.se

356:Pavlenko.tex,session:CS5D

We focus on the two-class linear in a high-dimensional setting where the number of features can greatly exceed the number of observations. In such setting, *sparse and weak models* is extensively studied in the literature (see e.g. Donoho et al. (2008) and references therein) which assume that that there are very few informative features and each such feature contributes weakly towards class separation. *Higher criticism* (HC) was suggested for detecting informative features in such settings, in particular the threshold selection for a feature separation strength is derived assuming that feature variables are independent and normally distributed.

In this study, we suggest a HC feature selection which can accommodate nonindependent case and show how to incorporate this technique in the supervised classification. We first present our covariance model selection which learns the sparsity patterns of the inverse covariance matrix and specifies an equivalence class of block structure approximations, in the sense that all class members result in classifiers with asymptotically equal misclassification probability.

We further derive *asymptotically sparse and weak blocks* model, representing the case when only a small fraction of blocks is informative for classification. We formalize growing dimensions asymptotic framework for analyzing this model by considering a sequence of supervised classification problems with increasingly many blocks (under the constraint on the block size) and relatively lower sample sizes. We suggest a technique for quantifying the block separation strength by relating the Hellinger between class distance to the misclassification probability and derive the multiple testing procedure which yields an empirical HC threshold, bHC-threshold for feature selection.

The asymptotic properties of the suggested empirical thresholding, in particular the performance accuracy of the resulting classifier, are shown to be close to those achieved by the ideal threshold. We also show that our bHC test procedure is a member of the ϕ-divergence tests family [Jager et al (2007)] and investigate the attainment of the detection boundary defined in the *phase space* represented by the sparsity and weakness parameters. In the same way as [Jager et al (2007)], we show that there exist an *area undetectability*, i.e the area where the true block separation strength is so low and blocks are so rare that any feature selection procedure fails. Our suggested bHC-thresholding is shown to be optimally adaptive in a sense that it works without knowledge of the block sparsity and weakness parameters but is successful in the detectable area of the phase space.

Using both synthetic and real data, classification accuracy with bHC feature thresholding is shown to be better than a number of alternative procedures such as e.g. false discovery rate (FDR)-thresholding and the usual HC-thresholding indicating the advantages of taking into account the dependence structure underlying the data.

References

[Donoho et al (2008)] Donoho D. and Jin J. 2008: Higher Criticism thresholding. *Proc. Natn. Acad. Sci. USA*, **105**, 14790-14795.

[Jager et al (2007)] Jager L. and Wellner J. 2007: Goodness-of-fit test via ϕ-divergences.*Ann. Statist.*, **35**, 2018-2053.

Waterbird Habitat Classification via Tensor Principal Component Analysis

OCS12 Experiment Design

XIAO-LING PENG[*,‡], SHENGCHUN WU[†]

[*]BNU-HKBU United International College, Zhuhai, China,
[†]City University of Hong Kong, Hong Kong, China
[‡]email: xlpeng@uic.edu.hk 357:Xiao_Ling_Peng.tex,session:OCS12

As a natural extension of traditional Principal Component Analysis (PCA), Tensor Principal Component Analysis (TensorPCA) has received more and more attention in recent years. The main advantage of TensorPCA is that it can be used to analyze data represented in matrices (second order tensor) or higher order tensors, which is common in the real world. Started from 1997, a Waterbird Monitoring Program (WMP) has been developed to monitor the distribution and trend of waterbirds in the Deep Bay Area of Hong Kong. During the program, more than 80 waterbird species were counted at 16 monitoring stations every month. By using Tensor PCA, we investigated the waterbird habitats based on factors "species" and "month". Classification of waterbird habitats obtained from TensorPCA is much more explainable than the results got from traditional PCA.

Acknowledgment. This research was partially supported by College Research Grant of BNU-HKBU United International College with grant code: R201236.

A Comparison of Semiparametric Estimators for the Binary Choice Model

NYA Not Yet Arranged

ALICIA PÉREZ-ALONSO[*,‡], MAJA RYNKO[†]

[*]Research Group in Economic Analysis (RGEA), Universidade de Vigo, Spain,
[†]Educational Research Institute (IBE), Warszawa, Poland
[‡]email: apereza@uvigo.es 358:AliciaPrezAlonso.tex,session:NYA

This paper compares five semi-parametric estimators for binary outcome models: a 'semi-nonparametric' estimator, a maximum score estimator, a smooth version of the maximum score estimator, a kernel-based quasi-maximum likelihood estimator and a least squares estimator for the transformed dependent variable. Monte Carlo evidence comparing the performance of the estimators is presented. The empirical example models the propensity to smoke and illustrates the differences in parameter estimates across analysed frameworks.

3^k Fractional Factorials, Optimal Designs for Estimating Linear and Quadratic Contrasts for $N \equiv$ 0mod3

CS32A Nonparame

KATERINA PERICLEOUS[*,§], STAVROS A.CHATZOPOULOS[†], FOTINI KOLYVA-MACHERA[†], STRATIS KOUNIAS[‡]

[*]National Technical University of Athens, Greece,
[†]Department of Mathematics, Aristotle University of Thessaloniki, Greece,
[‡]Department of Mathematics, University of Athens, Greece
[§]email: katerina.pericleous@gmail.com 359:KaterinaPericleous.tex,session:CS32A

In studying factorial designs, with factors at 3 levels, the primary interest is the estimation of linear and quadratic contrasts of factor's effects, when the number of units is N. The notion of

majorization is utilized and E-, A- and D-optimal designs are derived for values of N which are multiples of 3. The estimators of the linear contrast in the optimal designs are uncorrelated with those of quadratic contrasts. In linear models, for N a multiple of 9, the orthogonal arrays (OA) are Φ-optimal in the class of all designs estimating linear and quadratic contrasts. For values of N multiples of 9 plus 3 or plus 6, the notion of balanced arrays and partially balanced arrays is introduced. For N a multiple of 9 plus 3 two competing designs for E-, A-, and D-optimality are given, while for N multiple of 9 plus 6 four competing designs for optimality are given. Construction methods for the optimal designs are given either using orthogonal arrays or algorithms are employed. Tables of optimal designs are presented for small values of N.

References

[Hedayat et al. (1999)] Hedayat, A. S., Sloane, N. J. A., Stufken, J., 1999: *Orthogonal arrays: theory and applications*, Springer Series in Statistics.

[Kiefer (1975)] Kiefer, J., 1975: Construction and optimality of generalized Youden designs, *In a Survey of Statistical Design and Linear Models*, Edited by J. N. Srivastava, 333-353, Amsterdam, North Holland.

[Kolyva-Machera (1989a)] Kolyva-Machera, F., 1989a: D-optimality in 3^k designs for $N \equiv 1$ mod9 observations, *J. Statist. Plann. Inference*, **22**, 95-103.

[Kolyva-Machera (1989b)] Kolyva-Machera, F., 1989b: Fractional factorial designs and G-optimality. *In Proceedings Forth Prague Symp. on Asymptotic Statistics*, 349-358, Prague, Charles University Press.

[Marshall and Olkin (1979)] Marshall, A., Olkin, I., 1979: *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

[Mukerjee (1999)] Mukerjee, R., 1999: On the optimality of orthogonal array plus one run plans, *Ann. Statist*, **27**, 1256-1271.

[Pukelsheim (1993)] Pukelsheim, F., 1993: *Optimal Design of Experiments*, John Wiley and Sons, Inc.

[Wu et al. (2000)] Wu, C. F. J., Hamada, M., 2000: *Experiments: planning, analysis, and parameter design optimization*, Wiley, New York.

OCS9 Design of Experiments

Model Uncertainty in Bayesian Experimental Design

ANTHONY PETTITT[*,†], CHRISTOPHER DROVANDI[*], JAMES MCGREE[*]

[*]Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia

[†]email: a.pettitt@qut.edu.au 360:Tony_Pettitt.tex,session:OCS9

Here we present a sequential Monte Carlo (SMC) algorithm that can be used for any oneat- a-time Bayesian sequential design problem in the presence of model uncertainty where discrete data are encountered. Our focus is on adaptive design for model discrimination but the methodology is applicable if one has a different design objective such as parameter estimation or prediction. An SMC algorithm is run in parallel for each model and the algorithm relies on a convenient estimator of the evidence of each model which is essentially a function of importance sampling weights. Methods that rely on quadrature for this task suffer from the curse of dimensionality. Approximating posterior model probabilities in this way allows us to use model discrimination utility functions derived from information theory that were previously difficult to compute except for conjugate models. A major benefit of the algorithm is that it requires very little problem specific tuning. We demonstrate the methodology on three applications, including discriminating between models for decline in motor neuron numbers in patients suffering from motor neuron disease.

The Dirichlet Curve of a Probability in $\mathbb{R}^n$

CS39A
Distributio
Theory

GÉRARD LETAC*, MAURO PICCIONI†,‡

*Institut de Mathématiques de Toulouse, Université Paul Sabatier, France,
†Dipartimento di Matematica, Sapienza Universitá di Roma, Italia
‡email: piccioni@mat.uniroma1.it 361:MauroPiccioni.tex,session:CS39A

For $a > 0$, denote by F_a the set of positive measures α on $\mathbb{R}^n$ with finite mass a and $E_a \subset F_a$ the ones such that $\int \log(1+|x|)\alpha(dx) < \infty$. The Dirichlet random probability governed by $\alpha \in F_a$ is $P = \sum_{n=1}^{\infty} \delta_{B_n} Y_n \prod_{k=1}^{n-1}(1-Y_k)$ where $B_1, \ldots, B_n, \ldots, Y_1, \ldots, Y_n, \ldots$ are independent random variables such that $B_n \sim \alpha/a$ and $Y_n \sim \beta(1,a)$. Feigin and Tweedie have shown that $X = \int xP(dx) = \sum_{n=1}^{\infty} B_n Y_n \prod_{k=1}^{n-1}(1-Y_k)$ exists if and only if $\alpha \in E_a$. We denote by $\mu(\alpha)$ the law of X. If $\alpha \in E_1$ the curve $t \mapsto \mu(t\alpha)$ valued in the probabilities of $\mathbb{R}^n$ is called the Dirichlet curve of the probability α. The talk is about the properties of that curve. For stating them we define the one dimensional Cauchy distribution $c_z(dx) = bdx/(b^2 + ((x-a)^2)$ where $b \geq 0$ and $z = a + ib$ (thus is Dirac δ_a if $b = 0$). A Cauchy distribution of X in $\mathbb{R}^n$ is defined as having all linear forms $f(X)$ being of the form $c_{z(f)}$ and one can prove that they are the stable laws of parameter 1 which are symmetric around a point. If $X \sim \alpha$ is a positive random variable independent of $Y \sim \nu$ valued in $\mathbb{R}^n$ we denote by $\nu \circ \alpha$ the distribution in $\mathbb{R}^n$ of XY. Here is a list of properties:

1. The Sethuraman characterization of $\mu(\alpha)$: let $\alpha \in E_a$. If X, Y, B are independent such that $Y \sim \beta(1,a)$ and $B \sim \alpha/a$ then $X \sim (1-Y)X + YB$ if and only if $X \sim \mu(\alpha)$.

2. The Yamato observation: If C is Cauchy the Dirichlet curve $t \mapsto \mu(Ct)$ is a point, namely $\mu(tC) = C$ for all $t \geq 0$. The Hjort-Ongaro invariance: $C \circ \mu(t\alpha) = \mu(tC \circ \alpha)$ if C is a Cauchy probability in $\mathbb{R}^n$.

3. The James invariance: If α is a probability in $\mathbb{R}^n$ then $\mu(a\alpha + b\delta_0) = \mu(a\alpha) \circ \beta(a,b)$.

4. $\lim_{t \searrow 0} \mu(t\alpha) = \alpha$. If α has a mean m then $\lim_{t \Rightarrow \infty} \mu(t\alpha) = \delta_m$.

5. The characterization of Cauchy: $\mu(t\alpha) = \alpha$ only if α is Cauchy. We give a short proof for $t = 1$ (already done by Lijoi and Regazzini) and $t = 2$. This is a conjecture for any other fixed t.

6. Double point of the Dirichlet curve. If $0 \leq t < s$ and if $\mu(t\alpha) = \mu(s\alpha)$ then α is Cauchy: we prove it when t and s are integers such that $s - t = 1$ or 2. If for an integer $N > 0$ we have $\mu(m\alpha) = \alpha$ for all integers $m \geq N$ this implies that α is Cauchy. If for $0 \leq b < c$ we have $\mu(t\alpha) = \alpha$ for all $t \in (b,c)$ this implies that α is Cauchy.

7. In general we do not have $\int \log(1+|x|)\mu(\alpha)(dx) < \infty$. For a given integer m suppose that the probability α is such that $\mu \circ \ldots \circ \mu(\alpha)$ (m times) is defined. Does $\mu \circ \ldots \circ \mu(\alpha) = \alpha$ implies that is Cauchy? Yes for $m = 2$, answer unknown for $m \geq 3$.

8. One can vaguely conjecture that if $t < s$ then $\mu(t\alpha)$ is less concentrated than $\mu(s\alpha)$. For instance, if α has bounded convex support S one can conjecture that $t \mapsto \int \psi(x)\mu(t\alpha)(dx)$ is decreasing when ψ is convex on S (this is the Cartier-Fell-Meyer-Loomis order) To support this we prove that if α is on $(0,\infty)$ then $t \mapsto \int_0^{\infty} x^n \mu(\alpha)(dx)$ is decreasing for $n = 2, 3, 4$.

9. Let $0 < t < s$. If X, Y, B are independent such that $Y \sim \beta(2t, t-s)$ and $B \sim \beta(t,t)$ then $X \sim (1-Y)X + YB$ if and only if $X \sim \beta(s,s)$. This proves the last conjecture when α is the uniform distribution on the circle of $\mathbb{R}^2$ or is $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ in $\mathbb{R}$.

TL-Moments and L-Moments Estimation Based on Regression Quantiles

JAN PICEK*,†

*Technical University of Liberec, Czech Republic
†email: jan.picek@tul.cz 362:Picek.tex,session:CS31A

The conventional L-moments are analogues of the univariate mean and central moments and have similar interpretations but remain well-defined for all orders under merely a first moment assumption and possess other appealing properties as well. They are calculated using linear combinations of the expectation of ordered data. In practice, L-moments must usually be estimated from a random sample drawn from an unknown distribution as a linear combination of ordered statistics. Trimmed L-moments (TL-moments) assign zero weight to extreme observations and such that they are robust generalisations of population and sample L-moments.

The aim of present contribution is to extend the L and TL-moments to the linear regression model. We propose regression L and TL-moments based on the two-step regression quantiles as a suitable tools of that extension. The two-step regression quantiles are one of possible modifications of the regression quantiles which can be seen as a generalization of the quantile idea in linear regression model. The properties of regression L and TL-moments are illustrated on simulated and climatological data.

Acknowledgment. This research was partially supported by the Czech Science Foundation under project P209/ 10/ 2045

References

[Elamira] Elamira, E.A.H., Seheultb, A.H., 2003: Trimmed L-moments, *Computational Statistics & Data Analysis*, **43**, 3, 299–314.

[Hosking (1990)] Hosking, J.R.M., 1990: L-momentsanalysis and estimation of distributions using linear combinations of statistics, *J. Roy. Statist. Soc. B*, **52**, 105–124.

[JureckovaPicek] Jurečková, J., Picek, J., 2005: Two-step regression quantiles. *Sankhya*, **67**,2, 227–252.

POSTER
Poster

Bootstrap and the Moment Estimator of Tail Index

VÁCLAV KOHOUT*,†, JAN PICEK*,‡

*Technical University of Liberec, Czech Republic,
†University of West Bohemia, Czech Republic
‡email: jan.picek@seznam.cz 363:VclavKohout.tex,session:POSTER

In this contribution, we discuss the estimation of an extreme value index, the primary parameter in Statistics of Extremes. The estimation of the extreme value index usually performed on the basis of the largest k order statistics in the sample on the excesses over a high level u. The question that has been often addressed in practical applications is the choice of the sample fraction k. We shall mainly focus on the bootstrap metodology to choose the optimal sample fraction. We shall be also interested in the use of resampling-based computer-intensive methods for an choice of the thresholds The used methods will be demonstrated by numerical illustrations.

Acknowledgment. This research was partially supported by the Student Grant Competition of Technical University of Liberec under project 58006.

References

[De Haan] De Haan, L., Ferreira, A., 2006: Extreme Value Theory An Introduction, Springer , New York.

[Draisma)] Draisma, G., de Haan, L. , Peng., L. , Pereira, T.T., 1999: A bootstrap - based method to achieve optimality in estimating the extreme–value index, *Extremes*, **2**, 367–404.

Structure Learning using a Focused Information Criterion in Graphical Models—'Large p, small n' considerations

CS9B Model Sel, Info Crit

EUGEN PIRCALABELU[*,‡], GERDA CLAESKENS[*], LOURENS WALDORP[†]

[*]ORSTAT and Leuven Statistics Research Center, KU Leuven, Leuven, Belgium,
[†]Department of Psychological Methods University of Amsterdam, Amsterdam, the Netherlands
[‡]email: eugen.pircalabelu@kuleuven.be 364:EugenPircalabelu.tex,session:CS9B

A new method for model selection for Gaussian graphical models (GGMs), in cases where the number of nodes exceeds the number of cases is constructed to have good mean squared error properties. The method is based on the Focused information criterion (FIC) and unlike the traditional AIC or BIC, the FIC allows for selecting individual models, tailored to a specific purpose (the focus), as opposed to attempting an identification of a single model that should be used for all purposes.

Under the assumption of multivariate normality, there is a one-to-one correspondence between a 0-element in the inverse of the covariance matrix and the presence of an edge between two nodes. Given data, the goal is to learn plausible positions of edges in the graph, or equivalently conditional independencies between variables.

Working under a local misspecification framework which assumes that the true model is in a 'neighborhood' of the least complex model one is willing to assume, we define a *focus parameter*, i.e. $\mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ as a function of the parameters of the model density, and potentially of a user-specified vector of covariate values, for any particular node in the graph $G(E, V)$. Subsequent steps involve specifying a collection of models and a penalized objective function which is optimized for parameters corresponding to the prespecified models. The FIC estimates MSE($\hat{\mu}_S$) for each of a collection of models S, and selects the model with the smallest value. The focus of the research (i.e. the purpose of the model), directs the selection and different focuses may lead to different selected models. In this way, one can obtain better selected models in terms of MSE, than obtained from a global model search. For this application of FIC for structure learning in graphs, the focus is the expected value of a variable, reflecting interest in discovering a topology of the graph that produces a low MSE for this focus.

By the application of FIC on fMRI data, it is shown that using FIC one has at disposal a powerful method to study evolutions over time of networks constructed to study the functional connectivity between brain regions. Moreover, the method identifies more clearly important brain regions which seem to be highly connected with others, acting as 'hubs' or 'informational gateways'.

A small simulated data example suggests that the FIC scoring approach is able to identify in some settings graphical models with better MSE values than obtained from existing methods, such as the graphical Lasso or Chow-Liu algorithms.

Acknowledgment. This research was partially supported by KU Leuven grant GOA FlexStatRob.

References

[Claeskens, Hjort (2003)] Claeskens, G., Hjort, N.L., 2003: The focused information criterion (with discussion and a rejoinder by the authors), *JASA*, **98**, 900 - 916.

[Lauritzen (1996)] Lauritzen, S.L., 1996: Graphical Models, Oxford University Press.

IS29 Stoch. in Finance

Benchmarked Risk Minimization

KE DU*, ECKHARD PLATEN*,†

*University of Technology, Sydney, Australia
†email: eckhard.platen@uts.edu.au

365:EckhardPlaten.tex,session:IS29

The paper discusses the problem of hedging not perfectly replicable contingent claims by using a benchmark, the numeraire portfolio, as reference unit. The proposed concept of benchmarked risk minimization generalizes classical risk minimization, pioneered by Foellmer, Sondermann and Schweizer. The latter relies on a quadratic criterion, requesting the square integrability of contingent claims and the existence of an equivalent risk neutral probability measure. The proposed concept of benchmarked risk minimization avoids these restrictive assumptions. It employs the real world probability measure as pricing measure and identifies the minimal possible price for the hedgeable part of a contingent claim. Furthermore, the resulting benchmarked profit and loss is orthogonal to traded uncertainty and forms a local martingale that starts at zero. Benchmarked profit and losses, when pooled and sufficiently independent, become in total negligible. This property is highly desirable from a risk management point of view. It is making asymptotically benchmarked risk minimization the least expensive method for pricing and hedging of an increasing number of not fully replicable benchmarked contingent claims.

OCS6 Asympt. for Stoch Proc.

A Test for the Rank of the Volatility Process

JEAN JACOD*, MARK PODOLSKIJ†

*Université Pierre et Marie Curie, Paris, France,
†University Heidelberg, Heidelberg, Germany

366:MarkPodolskij.tex,session:OCS6

In this paper we present a test for the maximal rank of the matrix-valued volatility process in the continuous Itô semimartingale framework. Our idea is based upon a random perturbation of the original high frequency observations of an Itô semimartingale, which opens the way for rank testing. We develop the complete limit theory for the test statistic and apply it to various null and alternative hypotheses. Finally, we demonstrate a homoscedasticity test for the rank process.

CS9A Model Sel, Lin Reg

Delete or Merge Regressors for Linear Model Selection

PIOTR POKAROWSKI*,‡, ALEKSANDRA MAJ†, AGNIESZKA PROCHENKA†

*University of Warsaw, Warsaw, Poland,
†Polish Academy of Sciences, Warsaw, Poland
‡email: pokar@mimuw.edu.pl

367:PiotrPokarowski.tex,session:CS9A

Model selection is usually understood as selection of explanatory variables. However, when a factor (categorical predictor) is considered, in order to reduce model's complexity, we can either exclude the whole factor or merge its levels. A method introduced by Bondell and Reich (Biometrics 2009) called CAS-ANOVA executes merging levels with the use of the LASSO. Gertheiss and Tutz (Ann. Appl. Stat. 2010) proposed a modification of the CAS-ANOVA, which is more computationally efficient because of using the LAR algorithm. We propose a backward selection procedure "Delete or Merge Regressors", which combines deleting the continuous variables with merging levels of factors. We describe two variants of the method: the first is similar to the backward stepwise regression

and the second, faster implementation combines the agglomerative clustering for k-sample problem introduced by Ciampi et al (Pattern Anal. Appl. 2008) with ranking regressors by squared t-statistics proposed Zheng and Loh (JASA 1995). We proved that our selectors are consistent and post-selection estimators are asymptotically efficient. We describe also a simulation study and discuss a pertaining R package. The simulations show an advantage in the percent of times of choosing the true model in comparison to the other known algorithms.

Supervised Learning and Prediction of Spatial Epidemics

CS13A Epidem. Models

GYANENDRA POKHAREL[*,†], ROB DEARDON[*]

[*]University of Guelph, Ontario, Canada
[†]email: gpokhare@uoguelph.ca 368:GyanendraPokharel.tex,session:CS13A

Mechanistic models of infectious disease spread are commonly used to model space-time data. However, parameter estimation for such models can be highly computationally intensive. Nsoesie et al. (2011) introduced an approach for inference on infectious disease data based around the idea of supervised learning or clustering. Broadly, this method involved simulating epidemics from various infectious disease models, and then using a classifier built from the epidemic curve data to predict which model was most likely to have generated some other observed epidemic curves. Here, we extend this work to the case where the underlying infectious disease model is inherently spatial, and the nature of the spatial mechanism is unknown.

A major goal of this study is to compare the use of global epidemic curves for building the classifier, with the use of sets of spatially stratified epidemic curves. Two stratification methods (Rectangular and Circular) and various resolutions of stratification were used. The prediction error rate was observed using a confusion matrix. Both stratification methods gave significantly better result up to a certain degree of stratification. Comparing the two methods of stratification, the circular method performed better in a sense of reducing the prediction error. The prediction error rate was optimized in different degree of stratification depending on the parameter set.

References

[1] Nsoesie, E., Beckman, M. and Lewis, B. (2011). Prediction of an Epidemic Curve: A Supervised Classification Approach. *Statistical Communication in Infectious Diseases*, 3(1)

A Fractional Diffusion-Telegraph Equation and its Stochastic Solution

CS11A SDE-s

MIRKO D'OVIDIO[*], FEDERICO POLITO[†,‡]

[*]Dipartimento di Scienze di Base e Applicate per l'Ingegneria, "Sapienza" Università di Roma, Italy,
[†]Dipartimento di Matematica, Università degli Studi di Torino, Italy
[‡]email: federico.polito@unito.it 369:FedericoPolito.tex,session:CS11A

We present the explicit stochastic solution to a fractional diffusion-telegraph equation in which the time-operator is related to the so-called Prabhakar operator, i.e. an integral operator with a generalized Mittag–Leffler function in the kernel. The stochastic solution is given as a time changed process and is connected to the inverse process to a linear combination of stable subordinators with different indices. Note that the considered framework interpolates and thus generalizes directly fractional diffusion equations and fractional telegraph equations.

CS28A Random Graphs

Scale-Free Property in a Random Graph Model Based on N-Interactions

ISTVÁN FAZEKAS*, BETTINA PORVÁZSNYIK*,†

*University of Debrecen, Debrecen, Hungary
†email: porvazsnyik.bettina@inf.unideb.hu 370:BettinaPorvazsnyik.tex,session:CS28A

During the last 15 years the behaviour of many type of real-world networks was investigated such as the WWW, the Internet, social and biological networks. The main common characteristic of such networks is their scale-free property, in other word the power law degree distribution, i.e. $p_k \sim Ck^{-\gamma}$, as $k \to \infty$. To describe the phenomenon, in [1] the preferential attachment model was introduced. There are versions of the preferential attachment model. In [2] a model based on the interaction of three vertices was introduced. The power law degree distribution in that model was obtained in [3].

In this paper we consider a generalization of the model given in [2]. More precisely, we introduce a model based on the interaction of N vertices $(N \geq 3)$. We extend certain results of [3]. That is we prove the scale-free property both for the weights and the degrees for the N-interaction model. We follow the lines of [3] therefore the main tool of the proof is the Doob-Meyer decomposition of submartingales.

Acknowledgment. The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[1] Barabási, A.L., Albert, R., Emergence of scaling in random networks. *Science* **286** (1999), 509-512.

[2] Backhausz, Á., Móri, T. F., A random graph model based on 3-interactions. *Annales Univ. Sci. Budapest., Sect. Comp.* **36** (2012), 41-52.

[3] Backhausz, Á., Móri, T. F., Weights and degrees in a random graph model based on 3-interactions. arXiv: 1206.0633v1 [math.PR] 4 Jun 2012

CS19C Lim. Thms. Sums of RVs

On the Order of Magnitude of Sums of Negative Powers of Integrated Processes

BENEDIKT M. PÖTSCHER*,†

*University of Vienna, Vienna, Austria
†email: benedikt.poetscher@univie.ac.at 371:poetscher.tex,session:CS19C

Upper and lower bounds on the order of magnitude of $\sum_{t=1}^{n} |x_t|^{-\alpha}$, where x_t is an integrated process, are obtained. Furthermore, upper bounds for the order of magnitude of the related quantity $\sum_{t=1}^{n} v_t |x_t|^{-\alpha}$, where v_t are random variables satisfying certain conditions, are also derived.

NYA Not Yet Arranged

On Robust Procedures in Change-Point Problem

ZUZANA PRÁŠKOVÁ*,†, ONDŘEJ CHOCHOLA*

*Charles University in Prague, Czech Republic
†email: praskova@karlin.mff.cuni.cz 372:Praskova.tex,session:NYA

Detecting possible changes in the stochastic structure of observed data is one of the most important statistical problems. We will focus on robust procedures for detecting changes in parameters

of linear models that were developed to reduce some sensitivity of statistical decision procedures against outlying observations and heavy-tailed distributions. We will consider a class of CUSUM-type test statistics based on M-estimators and weighted M-residuals assuming that both the regressors and the errors are sequences of weakly dependent random variables or vectors, and study limit properties of the proposed test statistics both under the null hypothesis of no change and under alternatives. We concentrate on off-line procedures but on-line methods will be also mentioned. Computational aspects of the proposed procedures will be discussed.

On Size and Power of Heteroscedasticity and Autocorrelation Robust Tests

CS4B Time Series I.

DAVID PREINERSTORFER*,†, BENEDIKT M. PÖTSCHER*

*Department of Statistics and Operations Research, University of Vienna, Austria
†email: david.preinerstorfer@univie.ac.at 373:preinerstorfer.tex,session:CS4B

Testing restrictions on regression coefficients in linear models often requires correcting the conventional F-test for potential heteroscedasticity or autocorrelation amongst the disturbances, leading to so-called heteroskedasticity and autocorrelation robust test procedures. These procedures have been developed with the purpose of attenuating size distortions and power deficiencies present for the uncorrected F-test. We develop a general theory to establish positive as well as negative finite-sample results concerning the size and power properties of a large class of heteroskedasticity and autocorrelation robust tests. Using these results we show that nonparametrically as well as parametrically corrected F-type tests in time series regression models with stationary disturbances have either size equal to one or nuisance-minimal power equal to zero under very weak assumptions on the covariance model and under generic conditions on the design matrix. In addition we suggest an adjustment procedure based on artificial regressors. This adjustment resolves the problem in certain cases in that the so-adjusted tests do not suffer from size distortions. At the same time their power function is bounded away from zero. As a second application we discuss the case of heteroscedastic disturbances.

Numerical Wiener Chaos and Applications to the Stochastic Korteweg–de Vries Equation

OCS22 Numeric SPDE

JACQUES PRINTEMS*

*Université Paris–Est, Créteil, France
†email: printems@u-pec.fr 374:JacquesPrintems.tex,session:OCS22

Various phenomena of localized waves in dispersive media can be modelized by weakly nonlinear dispersive partial differential equations. This is the case with the Korteweg-de Vries (KdV) equation and the soliton phenomenum which can occur in water waves, plasma physics, beam propagation. In order to take into account the heterogeneity of the physical medium or the partial knowledge of the material property, we are often lead to introduce some random coefficients or random forcing terms. This gives the stochastic KdV equation:

$$\frac{\partial u}{\partial t}(x,t) + \varepsilon \frac{\partial^3 u}{\partial x^3}(x,t) + \frac{1}{2}\frac{\partial}{\partial x}(u^2(x,t)) = \dot{\xi}(x,t), \quad x \in \mathbb{R}, \quad t > 0, \tag{1}$$

with the initial condition $u(x,0) = u_0(x)$, $x \in \mathbb{R}$. Here $\xi = \xi(x,t)$ can be a centered Gaussian stochastic process.

In various applications, we are interested in the approximation of some averaged quantities depending on the solution (e.g. its first moments). A traditional approach is to use Monte Carlo simulations (or Galerkin Monte Carlo, see [2]). The main drawback of this approach is that the way the noise is discretized depends on the deterministic method used in order to solve the deterministic part. This yields also to a numerical spatial correlation.

A different approach is possible where the stochastic and deterministic parts of the equations are independently discretized owing to a polynomial chaos decomposition of the solution. Let us note that this approach has already been introduced in [3] in the context of Zakaï equations, or in [1] where the noise does not depend on the spatial variable. It relies on the introduction of an orthonormal basis of $L^2(\Omega;\mathbb{R})$, let say $\{\xi^\alpha\}$ where $\alpha = [\alpha_{i,j}]$ denotes a multi-index, built after Hermite polynomials (in the Gaussian case). The next step consists in the projection of the solution on this basis, i.e. $u_\alpha(x,t) = \mathbb{E}(u(x,t)\xi^\alpha)$. Then the initial equation (1) is replaced by a family of deterministic PDE coupled by the nonlinear term:

$$u(x,t) = \sum_\alpha u_\alpha(x,t)\xi^\alpha, \ \frac{\partial u_\alpha}{\partial t} + \varepsilon\frac{\partial^3 u_\alpha}{\partial x^3} + \mathbb{E}(u\frac{\partial u}{\partial x}\xi^\alpha) = \mathbb{E}(\dot{\xi}(x,t)\xi^\alpha), \quad x \in \mathbb{R}, \quad t > 0. \tag{2}$$

Numerically speaking, the nonlinear term in the case of a random forcing is the main source of problems in the algorithm, especially because $\xi^\alpha\xi^\beta \neq \xi^{\alpha+\beta}$. Typically, the way the different chaoses have to be multiplied is crucial for numerical concerns.

To that extent, we introduce here a way to take into account the nonlinear term for numerics and compare the complexity of our method with respect to Monte Carlo simulations. We will show also how to use Wiener chaos expansion in order to reduce the variance of a Monte Carlo estimator.

References

[1] G. Lin, L. Grinberg and G. E. Karniadakis, *Numerical studies of the stochastic Korteweg-de Vries equation*, J. of Comp. Phys. **213**, 676–703 (2006).

[2] I. Babuska, R. Tempone and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal. **42**(2), 800–825 (2004).

[3] S. Lototsky, R. Mikulevicius and B. Rozovskii, *Nonlinear filtering revisited: a spectral approach*, SIAM J. Control Optim. **35**(2), 435–461 (1997).

OCS3 Spectral Analysis

The Exponential Model for the Spectrum of a Time Series: Extensions and Applications

TOMMASO PROIETTI[*,‡], ALESSANDRA LUATI[†]

[*]University of Rome Tor Vergata, Italy,
[†]University of Bologna, Italy
[‡]email: tommaso.proietti@uniroma2.it 375:TommasoProietti.tex,session:OCS3

The exponential model for the spectrum of a time series and its fractional extensions are based on the Fourier series expansion of the logarithm of the spectral density. The coefficients of the expansion form the cepstrum of the time series. After deriving the cepstrum of important classes of time series processes, also featuring long memory, we discuss likelihood inferences based on the periodogram, for which the estimation of the cepstrum yields a generalized linear model for exponential data with logarithmic link, focusing on the issue of separating the contribution of the long memory component to the log-spectrum. We then propose two extensions. The first deals with replacing the logarithmic link with a more general Box-Cox link, which encompasses also the identity and the inverse links: this enables nesting alternative spectral estimation methods (autoregressive, exponential, etc.) under

the same likelihood-based framework. Secondly, we propose a gradient boosting algorithm for the estimation of the log-spectrum and illustrate its potential for distilling the long memory component of the log-spectrum.

On the Ergodicity of the Lévy Transformation

CS23A Diffusions & Diff. Eq.

VILMOS PROKAJ[*,†]

[*]Eötvös Loránd University
[†]email: prokaj@cs.elte.hu 376:ProkajVilmos.tex,session:CS23A

In the talk we present a new approach to the old standing open problem of the ergodicity of the so called Lévy transformation of the Wiener space. This transformation T sends a Brownian motion B into another one by the formula

$$(TB)_t = \int_0^t \operatorname{sign}(B_t) dB_t = |B_t| - L_t,$$

where L is the local time of B at the level zero. The main result is that the question could be answered in the affirmative by investigating the sequence $\{(T^n B)_1 : n \geq 0\}$. Roughly speaking, if the expected hitting time to small neighborhoods of zero is inversely proportional to the size of the neighborhood then even strong mixing of the Lévy transformation follows.

A Mean-Field Limiting Method for Reliability Computation in Repairable Stochastic Networks

OCS18 Lifetime Data Anal.

QUAN-LIN LI[*,‡], MENG WANG[*], JUAN ELOY RUIZ-CASTRO[†]

[*]School of Economics and Management Sciences Yanshan University, Qinhuangdao 066004, China,
[†]Department of Statistics and Operational Research, University of Granada, Spain
[‡]email: liquanlin@tsinghua.edu.cn 377:Quan-LinLi.tex,session:OCS18

Reliability analysis of complex stochastic networks is always an interesting and challenging topic in many practical areas such as computer networks, information security, manufacturing systems and transposition networks. The main purpose of this paper is to develop a product-form solution by means of the asymptotic independence, which is supported by the mean-field limit when the scale of stochastic networks (for example, nodes, routines and repairmen) goes to infinity. As an illustrating example, we consider a repairable supermarket model under continuous improvement of the dynamic randomized load balancing schemes, including the allocation of arrival customers, the priority of working servers joined by any arriving customer, and the mass scheduling of repairmen.

In this paper, we set up an infinite-dimensional system of differential equations satisfied by the expected fraction vector in terms of a technique of tailed equations. As N goes to infinity, we apply the operator semigroup to provide the mean-field limit for the sequence of Markov processes which asymptotically approach a single trajectory identified by the unique and global solution to the infinite-dimensional system of limiting differential equations. We provide an effective and efficient algorithm for computing the fixed point, which leads to the product-form solution for reliability analysis of this supermarket model. Furthermore, some numerical examples show how the reliability measures of this repairable supermarket model depend on some crucial parameters under four classes of dynamic randomized load balancing schemes.

NYA
Not Yet
Arranged

Delay-Dependent Optimal Guaranteed Cost Control of Stochastic Neural Networks with Interval Nondifferentiable Time-Varying Delays

GRIENGGRAI RAJCHAKIT*,†

*Division of Mathematics, Faculty of Science, Maejo University, Chiangmai, 50290, Thailand
†email: griengkrai@yahoo.com 378:GrienggraiRajchakit.tex,session:NYA

Stability and control of neural networks with time delay has been attracted considarable attention in recent years [1-8]. In many practical systems, it is desirable to design neural networks which are not only asymptotically or exponentially stable but can also guarantee an adequate level of system performance. In the area of control, signal processing, pattern recognition and image processing, delayed neural networks have many useful applications. Some of these applications require that the equilibrium points of the designed network be stable. In both biological and artificial neural systems, time delays due to integration and communication are ubiquitous and often become a source of instability. The time delays in electronic neural networks are usually time-varying, and sometimes vary violently with respect to time due to the finite switching speed of amplifiers and faults in the electrical circuitry. Guaranteed cost control problem [9-12] has the advantage of providing an upper bound on a given system performance index and thus the system performance degradation incurred by the uncertainties or time delays is guaranteed to be less than this bound. The Lyapunov-Krasovskii functional technique has been among the popular and effective tool in the design of guaranteed cost controls for neural networks with time delay. Nevertheless, despite such diversity of results available, most existing work either assumed that the time delays are constant or differentiable [13-16]. Although, in some cases, delay-dependent guaranteed cost control for systems with time-varying delays were considered in [12, 13, 15], the approach used there can not be applied to systems with interval, non-differentiable time-varying delays. To the best of our knowledge, the guaranteed cost control and state feedback stabilization for stochastic neural networks with interval, non-differentiable time-varying delays have not been fully studied yet (see, e.g., [4–16] and the references therein), which are important in both theories and applications. This motivates our research.

In this paper, we investigate the guaranteed cost control for stochastic delayed neural networks problem. The novel features here are that the delayed neural network under consideration is with various globally Lipschitz continuous activation functions, and the time-varying delay function is interval, non-differentiable. A nonlinear cost function is considered as a performance measure for the closed-loop system. The stabilizing controllers to be designed must satisfy some mean square exponential stability constraints on the closed-loop poles. Based on constructing a set of augmented Lyapunov-Krasovskii functionals combined with Newton-Leibniz formula, new delay-dependent criteria for guaraneed cost control via memoryless feedback control is established in terms of LMIs, which allow simultaneous computation of two bounds that characterize the mean square exponential stability rate of the solution and can be easily determined by utilizing MATLABs LMI Control Toolbox.

Acknowledgment. This research was partially supported by the National Research Council of Thailand, the Thailand Research Fund Grant, the Higher Education Commission and Faculty of Science, Maejo University, Thailand.

References

[1] Hopfield J.J., "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. USA*, **79**(1982), 2554-2558.

[2] Kevin G., *An Introduction to Neural Networks*, CRC Press, 1997.

[3] Wu M., He Y., She J.H., *Stability Analysis and Robust Control of Time-Delay Systems,* Springer, 2010.

[4] Arik S., An improved global stability result for delayed cellular neural networks, IEEE Trans. Circ. Syst. **499**(2002), 1211-1218.

[5] He Y., Wang Q.G., M. Wu, LMI-based stability criteria for neural networks with multiple time-varying delays, Physica D, **112**(2005), 126-131.

[6] Kwon O.M., J.H. Park, Exponential stability analysis for uncertain neural networks with interval time-varying delays. *Appl. Math. Comput.,* **212**(2009), 530-541.

[7] Phat V.N., Trinh H. , Exponential stabilization of neural networks with various activation functions and mixed time-varying delays, *IEEE Trans. Neural Networks,* **21**(2010), 1180-1185.

[8] Botmart T.and P. Niamsup, Robust exponential stability and stabilizability of linear parameter dependent systems with delays, *Appl. Math. Comput.,* **217**(2010), 2551-2566.

[9] W.H. Chen W.H., Guan Z.H., Lua X., Delay-dependent output feedback guaranteed cost control for uncertain time-delay systems, *Automatica,* **40**(2004), 1263 - 1268.

[10] Palarkci M.N., Robust delay-dependent guaranteed cost controller design for uncertain neutral systems,*Appl. Math. Computations,* **215**(2009), 2939-2946.

[11] Park J.H., Kwon O.M., On guaranteed cost control of neutral systems by retarded integral state feedback, *Applied Mathematics and Computation*, **165**(2005), 393-404.

[12] Park J.H., Choi K., Guaranteed cost control of nonlinear neutral systems via memory state feedback, *Chaos, Fractals and Solitons,* **24**(2005), 183-190.

[13] Fridman E., Orlov Y., Exponential stability of linear distributed parameter systems with time-varying delays. *Automatica,* **45**(2009), 194-201.

[14] Xu S., Lam J., A survey of linear matrix inequality techniques in stability analysis of delay systems. *Int. J. Syst. Sci.,* **39**(2008), 12, 1095-1113.

[15] Xie J.S., Fan B.Q., Young S.L., Yang J., Guaranteed cost controller design of networked control systems with state delay, *Acta Automatica Sinica,* **33**(2007), 170-174.

[16] Yu L., Gao F., Optimal guaranteed cost control of discrete-time uncertain systems with both state and input delays, *Journal of the Franklin Institute,* **338**(2001), 101-110.

[17] Gu K., Kharitonov V., Chen J., Stability of Time-delay Systems, Birkhauser, Berlin, 2003.

Variability Measures of Neural Spike Trains and Their Estimation

POSTER
Poster

KAMIL RAJDL[*,†,‡], PETR LÁNSKÝ[*,†]

[*]Institute of Physiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic,
[†]Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno, Czech Republic
[‡]email: xrajdl@math.muni.cz

379:KamilRajdl_poster.tex,session:POSTER

One of the most important questions in neuroscience is how neurons code transferred information into sequence of spikes, which are usually described as a stochastic point process. The basic and frequently considered concept is rate coding, which assumes that the information is coded by spike rate, mostly calculated as a spike count average. The spike rate does not fully depend on exact spike times and different spike trains can yield the same spike rate. So there is a possibility that also spiking variability carries some information. The question whether variability coding is used in real neurons or the variability is caused just by a noise without useful information is still unanswered. However, this issue increases the importance of suitable variance quantification.

There are various ways how to measure spike train variability. Probably, the most often applied are coefficient of variation (CV, the ratio of standard deviation of lengths of interspike intervals to their mean) and Fano factor (FF, the ratio of variance of numbers of spikes in an observation window to mean of the numbers). Unfortunately, their reliable estimation is often difficult. The main problem

is strong bias for short observation window, but there are also the questions how to segment a spike train for estimation of FF and which measure is more suitable in which situation.

Standard estimator of FF assuming that the spike trains follow a general renewal process is studied. We investigate its bias in dependence on the length of the observation window and on the refractory period. Both, relative and absolute refractory period, cause an initial decrease of the mean of the estimator, which increases the bias. Next, we derive an approximate asymptotic formula for mean square error of the estimator. This formula can be used, among others, to find the optimal segmentation of a single spike train. Finally, we compare the accuracy of the FF and CV estimators in various situations. The results are illustrated for gamma and inverse Gaussian probability distributions of the interspike intervals.

CS26A Extremes

Spatial Modeling by Generalized Pareto Process

PÁL RAKONCZAI[*,‡], FERIDUN TURKMAN[†]

[*]Eötvös Loránd University, Budapest, Hungary,
[†]CEAUL, University of Lisbon, Portugal
[‡]email: rakonczai.p@gmail.com

380:PalRakonczai.tex,session:CS26A

The multivariate extreme value distribution (MEVD) and generalized Pareto distribution (MGPD) are successfully applicable in practice for modeling extreme values of observations of finite sites, although the well-known methods are getting cumbersome and often untractable with increasing number of dimensions. In order to describe the behavior of extreme observations occurring at an uncountable set, e.g. on a surface, it is more appealing to fit stochastic processes which are extended from known finite dimensional extreme value models. For modeling maxima over a fixed period of time we can either apply max-stable process or for threshold exceedances the so-called generalized Pareto (GP) process. Here we compare which of the above processes are more effective in practice and also apply them for environmental data in Portugal.

Acknowledgment. P. Rakonczai's research was supported by the Hungarian National Research Fund OTKA, mobility grant No.: MB08A-84576.

References

[1] de Haan, L. (1984) A spectral representation for max-stable processes. *Ann. Probab.*, **12**, p.1194-1204

[2] Ferreira, A. and de Haan, L. (2012) The Generalized Pareto process; with application. arXiv: 1203.2551

[3] Rakonczai, P. and Zempléni, A.: Bivariate generalized Pareto distribution in practice: models and estimation. *Environmetrics* **23** (2012), 219-227.

[4] Smith, R. L. (1990) Extreme value theory. *In Handbook of Applicable Mathematics* (ed. W. Ledermann), **7**, p.437-471.

CS19B Lim. Thms. Point Proc.

Modelling Multi-site Rainfall Time Series using Stochastic Point Process Models

NADARAJAH I RAMESH[*,†], RASIAH THAYAKARAN[*]

[*]School of Computing and Mathematical Sciences, University of Greenwich, UK
[†]email: N.I.Ramesh@greenwich.ac.uk

381:NadarajahRamesh.tex,session:CS19B

We consider stochastic point process models, based on doubly stochastic Poisson process, to analyse rainfall data collected in the form of raingauge bucket tip-times over a network of stations in a catchment. Multi-site doubly stochastic point process models are constructed whereby the arrival

rate of the process varies according to a finite-state Markov chain thought to be representing the underlying environmental weather conditions. The tip-time series at a station is viewed as a univarite stochastic point process evolving in time and its multi-variate generalisation is studied to analyse data from multiple sites across the network. The likelihood function of this class of multi-site models, which is not usually available for most point process models, is derived by conditioning on the underlying Markov chain of the process. This allows us to make use of the likelihood approach for parameter estimation and inference.

The application of the proposed class of multi-site models, a useful alternative to the well known Poisson cluster models based on either Bartlett-Lewis or Neyman-Scott processes, in rainfall modelling is explored. We use data from the Hydrological Radar Experiment (HYREX) project, supplied by the British Atmospheric Data Centre (BADC), over a dense raingauge network in Brue experimental catchment in Somerset, South-West England. The models are used to make inferences about the properties of accumulated rainfall in discrete time intervals of equal length with the focus on fine time-scale. The proposed models and their variant that incorporate local covariate information such as elevation, temperature, sea-level pressure and relative humidity are utilised to study properties of rainfall time series from multiple sites. Results of the models that incorporate covariates are compared with the results of the model that does not take account of any covariates. The analysis shows the potential of this class of models in reproducing temporal and spatial variability of rainfall characteristics over the catchment area.

References

[Ramesh, N.I. et al. (2013)] Ramesh, N.I., Thayakaran, R. and Onof, C. 2013: Multi-site doubly stochastic Poisson process models for fine-scale rainfall, *Stochastic Environmental Research and Risk Assessment*, 1-14. DOI: 10.1007/s00477-012-0674-x

Sample Variability and Causal Inference with Instrumental Variables

IS14 Causal Inference

ROLAND RAMSAHAI[*,†]

[*]University of Cambridge, Cambridge, United Kingdom
[†]email: ramsahai@statslab.cam.ac.uk 382:RolandRamsahai.tex,session:IS14

Instrumental variables allow the computation of causal bounds from the observational distribution. However the only falsifiable constraints imposed by the model are inequalities. Causal inference accounts for sampling uncertainty in data by a likelihood analysis of the instrumental variable model. This involves a non-standard likelihood ratio test of the inequalities and maximum likelihood estimation of the causal bounds. This approach is described for the instrumental variable model and its variants, which relax or impose the exclusion restriction, randomization and monotonicity assumptions. The occurrence of sampling zeros in the data affects the validity of the results. The adjustment of the likelihood analysis for sparse contingency tables is discussed.

CS34A Clinical Studies

On a Better Identification of Survival Prognostic Models

PAOLA M.V. RANCOITA[*,†,‡,¶], CASSIO P. DE CAMPOS[†,‡], FRANCESCO BERTONI[‡,§]

[*]University Centre of Statistics for Biomedical Sciences (CUSSB), Vita Salute San Raffaele University, Milan, Italy,
[†]Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno, Switzerland,
[‡]Lymphoma & Genomics Research Program, Institute of Oncology Research (IOR), Bellinzona, Switzerland,
[§]Oncology Institute of Southern Switzerland (IOSI), Bellinzona, Switzerland
[¶]email: rancoita.paolamaria@unisr.it 383:PaolaMVRancoita.tex,session:CS34A

A prognostic model/index is a classification procedure that partitions patients into groups that have different event-free survival curves, on the basis of a subset of clinical and/or biologic variables. The identification of a prognostic model is a key task in many clinical studies. In fact, each class of patients defined by the model is associated with a different grade of prognosis, which may help clinicians to decide for the most suitable treatment for each patient. For instance, patients with good prognosis according to the model might be treated with less intensive regimens, while those with very poor outcome might receive additional and/or experimental therapies.

Survival trees are well-known methods that stratify patients with respect to survival outcomes and can be used to predict survival of new patients. The resulting classification procedure is represented by a decision tree: recursively, at each node, the corresponding patients are divided into subgroups based on their values of a given covariate. Finally, the terminal nodes (also called leaves) represent the actual groups in which the model has partitioned the patients. Two of the main advantages of these methods are: 1) they automatically select the variables to include in the model, even accounting for the relationships among them, 2) they decide what is the best cut-point of each variable for the purpose of partitioning the patients. Nevertheless, a straightforward use of survival trees to develop a prognostic index is not necessarily the best approach, since they do not account for different clinical causes (corresponding to the leaves) that lead to similar survival outcome. Only few procedures have been presented to merge the groups defined by the leaves in a survival tree, but they show limitations. To overcome those limitations, we propose to apply a clustering algorithm on the survival curves corresponding to the leaves' groups, using a proper dissimilarity measure. We investigate the performance of several choices both in terms of clustering algorithms and dissimilarity measures. Using simulated and public real data, we compare our method against other procedures (especially against those that reduce or combine leaves) and we show that it usually outperforms all of them.

Moreover, the usually employed measures to assess the performance of prognostic models only evaluate one or two characteristics that an ideal prognostic model should have. Therefore, with the purpose of enhancing the comparison between methods and improving the selection of best prognostic models, we derive a new index of separation which accounts for three important characteristics of prognostic groups: 1) they must keep the same order (i.e. grade of prognosis) when applied to distinct cohorts, 2) they must be reliable/robust in terms of the size of each subgroup, 3) all corresponding survival curves must be "well separated". Our separation index is intended to be used together with an error measure of survival prediction (such as the Brier score), which together induce a complete evaluation of a prognostic model. Our experimental results, obtained in several scenarios of simulated and real data, support that the new separation index always performs equally or better than other commonly used measures.

Acknowledgment. This work was partially supported by grants from Nelia et Amadeo Barletta Foundation (Lausanne, Switzerland), Ente Ospedaliero Cantonale (EOC) (Bellinzona, Switzerland) and Hasler Foundation grant no. 10030.

Bayesian Predictive Risk Modeling of Microbial Criterion for Campylobacter in Broilers

POSTER
Poster

JUKKA RANTA[*,‡], ANTTI MIKKELÄ[*], PIRKKO TUOMINEN[*], MAARTEN NAUTA[†]

[*]Finnish Food Safety Authority, Evira, Helsinki, Finland,
[†]Technical University of Denmark, Søborg, Denmark
[‡]email: jukka.ranta@evira.fi

384:JukkaRanta.tex,session:POSTER

Microbial Criteria define the acceptability of food production, based on the presence or detected number of microorganisms in samples from the production. The criterion is applied at the level of defined food lots or batches of production, such that batches that are not complying with the criterion are rejected or receive some additional treatment to reduce the contamination. As highly contaminated batches are likely to be rejected, a risk reduction for consumers is expected. However, a quantitative estimate of the implied risk reduction is non-trivial, because it depends on many unknown parameters. Variable quantity and quality of data leads to uncertainties which can be assessed by computing posterior distribution of the parameters. The available data may not even be informative about all the parameters, so that information from different supporting sources needs to be drawn by Bayesian evidence synthesis. This is implemented as a hierarchical model describing both prevalence and bacterial concentrations and their variation between and within batches. As a result of this, the posterior distribution of the parameters describes the remaining uncertainty, conditional to the stated combined evidence. Finally, the outcome of the microbial criterion (rejection/acceptance/not applied) for a batch can then be treated as additional evidence concerning the particular batch. Posterior predictive risk estimates can then be computed concerning the food safety risks that would result from such batch with the given observed outcome. An example study on Campylobacter in broilers from a Nordic project is presented. Computations are implemented in OpenBUGS.

Superhedging under Liquidity Constraints

IS29
Stoch. in Finance

PAOLO GUASONI[*], MIKLÓS RÁSONYI[†,‡]

[*]Boston University, US, Dublin City University, Ireland,
[†]University of Edinburgh, UK,,
[‡]Rényi Institute, Budapest, Hungary

385:MiklosRasonyi.tex,session:IS29

A most natural question in mathematical finance is the following: An investor is facing a payment obligation at time $T > 0$. Which initial endowments (at time 0) allow him/her to meet this obligation with probability one ?

The answer depends on the trading mechanism of the given market. Classical results apply if we assume that there are no market frictions. However, the picture changes when more realistic models are considered (e.g. when transaction costs are taken into account).

We review the available results and then present a new theorem for a general, continuous-time model with liquidity constraints. This result has applications to optimal investment problems as well.

NYA
Not Yet
Arranged

Dynamics of the Many Particle Jaynes-Cummings Model

NIKOLAY (JR.) BOGOLUBOV[*,§], MUKHAYO RASULOVA[†,‡,¶], ILKHOM TISHABAEV[‡,‖]

[*]V.A.Steklov Institute of Mathematics of the Russian Academy of Sciences Moscow 119991, Russia,
[†]MIMOS BHD, Technology Park Malaysia, Kuala-Lumpur 57000, Malaysia,
[‡]Institute of Nuclear Physics Academy of Sciences Republic of Uzbekistan Ulughbek, Tashkent 100214, Uzbekistan
email: [§]bogolubv@mi.ras.ru; [¶]rasulova@live.com; [‖]tishabaev@inp.uz
386:MukhayoRasulova.tex,session:NYA

We consider the dynamics of a system consisting of N two-level atoms interacting with a multimode cavity field, as an example of the generalized Jaynes-Cummings model. Based on formulation of the collective atom variables the Jaynes-Cummings model is generalized to a system of N two-level atoms. For the given system, the generalized kinetic equation is obtained and a condition is given under which solution of the generalized kinetic equation is reduced to solution of the linear equation.

CS39A
Distributio
Theory

Comparison of Certain Differential Geometrical Properties of the Manifolds of The Original Distributions and Their Weighted Versions Arising in Data Analysis

MAKARAND RATNAPARKHI[*,†]

[*]Wright State University,Dayton,Ohio,USA
[†]email: makarand.ratnaparkgi@wright.edu
387:MakarandRatnaparkhi.tex,session:CS39A

During the last twenty-five years many researchers in the medical studies, for example, in the cancer studies, have observed that the collected data are length-biased (more generally, referred to as the size-biased data). The other fields of applications where the length-biased data arises are survival analysis, reliability, environmental studies and many others. Such data, obviously, do not represent the original distribution that is of interest to the experimenter. Therefore, out of necessity, for modeling the length-biased data the weighted (length-biased) versions of the original distributions are used in all such studies. The underlying assumption for the use of such weighted version is that the parameter space of the weighted distribution will carry forward the desired structure of the parameter space of the original distribution and hence the inference based on the weighted distribution should be adequate/appropriate for all practical purposes. However, such assumption may not hold good for all distributions. The weighted distributions are, basically, the mathematical constructs (and not the original statistical distributions) obtained by using appropriate (mathematical) operator. Such operator, for some original distributions, may change the structure of the parameter space and hence complicate the parameter estimation. To demonstrate the existence of the above mentioned situation, as clearly as possible, the differential geometrical properties of some commonly used original distributions such as normal, lognormal, gamma, and Weibull and their weighted versions are compared. In particular, the alpha-curvature and the geodesics of the statistical manifolds are considered for this presentation. In view of the role of differential geometry in statistical inference ,in general, and the above concepts in particular, the results obtained here should be useful in practice.

Using Scan Statistics on Multiple Processes with Dependent Variables, with Application to Genomic Sequence Search

CS7B Spatio-Temp. Stat II.

ANAT REINER-BENAIM[*,†]

[*]University of Haifa, Haifa, Israel
[†]email: areiner@stat.haifa.ac.il 388:AnatReinerBenaim.tex,session:CS7B

The problem of locating sequences of interest along the genome is frequently confronted by genome researchers. The challenge here is to identify short intervals within noisy and much longer sequences, which exhibit certain behavior. One example is the search for introns, which are DNA intervals that are spliced out on the path to synthesize proteins. Inference on the presence of intronic intervals can be made using genome-wide expression data produced by the tiling array technology. A scan statistic is suggested to test whether an interval, within a specified gene, is exhibiting the behavior expected to occur in an intronic interval. The statistic integrates several important considerations related to the dependence between adjacent measures of expression along the genomic sequence. An analytical assessment of the scan statistics distribution considering this dependence is presented, along with its effect on FDR and power when testing simultaneously many random processes (genes).

Computational Lower Bounds for Sparse PCA

IS12 Machine Learning

PHILIPPE RIGOLLET[*], QUENTIN BERTHET[*]

[*]Princeton University

389:RigolletPhillipe.tex,session:IS12

In the context of sparse principal component detection, we bring evidence towards the existence of a statistical price to pay for computational efficiency. We measure the performance of a test by the smallest signal strength that it can detect and we propose a computationally efficient method based on semidefinite programming. We also prove that the statistical performance of this test cannot be strictly improved by any computationally efficient method. Our results can be viewed as complexity theoretic lower bounds conditionally on the assumptions that some instances of the planted clique problem cannot be solved in randomized polynomial time.

Bayes' Theorem Then and Now

Bayes Bayes Mem. Lecture

CHRISTIAN P. ROBERT[*,†,‡]

[*]Université Paris-Dauphine,
[†]CREST, INSEE
[‡]email: bayesianstatistics@gmail.com 390:Christian_P_Robert.tex,session:Bayes

What is now called Bayes' Theorem was published and maybe mostly written by Richard Price in 1763, 250 ago. It was re-discovered independently (?) in 1773 by Pierre Laplace, who put it to good use for solving statistical problems, launching what was then called inverse probability and now goes under the name of Bayesian statistics. The talk will cover some historical developments of Bayesian statistics, focussing on the controversies and disputes that marked and stil mark its evolution over those 250 years, up to now. It will in particular address some arguments about prior distributions made by John Maynard Keynes and Harold Jeffreys, as well as divergences about the nature of testing by Dennis Lindley, James Berger, and current science philosophers like Deborah

Mayo and Aris Spanos, and misunderstandings on Bayesian computational issues, including those about approximate Bayesian computations (ABC).

CS7A Spatio-Temp. Stat I.

Considering High-Order Interdependencies in Spatial Statistics: A Cumulant Generating Function Approach

JHAN RODRÍGUEZ[*,†], ANDRÁS BÁRDOSSY[*]

[*]Universität Stuttgart, Stuttgart, Germany

[†]email: Jhan.Rodriguez@iws.uni-stuttgart.de 391:JhanRodriguez.tex,session:CS7A

In the field of geostatistics and spatial statistics, variogram based models have proved a very flexible and useful tool. However, such spatial models take into account only interdependencies between pairs of variables, mostly in the form of covariances. In the present work, we point out to the necessity to extend the interdependence models beyond covariance modeling; we summarize some of the difficulties arising when attempting such extensions; and propose an approach to address these difficulties.

The necessity for extending covariance models, apart from the common sense notion that there can be more structure in a dataset than that expressed in terms of pairwise relations, has been suggested recently in one of the authors' research. For example, two multivariate datasets/models with identical correlation matrices can exhibit systematically different congregation patterns, as expressed by entropy based measures applied to multivariate ($d \geq 3$) marginals; this has been observed for daily precipitation values in a 32-dimensional dataset of Southwest Germany.

An initial difficulty in trying to consider interdependence measures which go beyond pair-wise measures, is to conceptualize what, say, a three-wise correlation coefficient might mean, or how is it to be interpreted. We suggest that joint cumulants are legitimate extensions of the covariance coefficients, since both represent the integral of a well known interaction measure (the Lancaster Interaction Measure); the covariance being the special case for $d = 2$. Then, from a more practical point of view, we suggest to address the issue of higher order interdependence via subject-matter relevant *manifestations* of such interdependence. Three example manifestations are provided, and their connection with multivariate joint cumulants is exhibited, namely: the distribution of the sum, the joint survival function, and the differential entropy of subsets S of the random vector representing the random field under study, where $\|S\| > 2$. The importance of the first of these for rainfall modeling is illustrated.

An important difficulty in trying to consider extensions to covariance models is the high dimensionality incurred. This high dimensionality is palliated by the use of low dimensional variogram models in traditional spatial statistics. By considering a cumulant generating function (c.g.f.) as a dependence structure, and introducing an archetypal c.g.f., we show that much of this low-dimensional approach can be kept, while allowing the consideration of higher order interdependencies.

Finally, it is indicated how we can use this archetypal dependence structure (i.e., c.g.f.) together with marginal transformations, in order to give more flexibility to the method, while retaining its low-dimensional desirable properties. The specific case of marginal transformations of the form $Y_j = \sum_{r=0}^{R} a_r X_j^r$ is presented, where $(X_1, \ldots, X_J)$ is the random vector possessing the archetypal cumulant generating function.

Acknowledgment. This research has been supported by the German Academic Exchange Service (DAAD).

A New Regression Model for Overdispersed Count Data

POSTER
Poster

JOSÉ RODRIGUEZ-AVI*,†, MARÍA JOSÉ OLMO-JIMÉNEZ*, ANTONIO CONDE-SÁNCHEZ*, ANA MARÍA MARTÍNEZ-RODRÍGUEZ*

*Department of Statistics and Operations Research. University of Jaén, Spain
†email: jravi@ujaen.es 392:RodriguezAvi.tex,session:POSTER

The Complex Biparametric Pearson distribution (CBPD) is a biparametric count-data distribution of infinite range that belongs to the family of Gaussian distributions generated by the hypergeometric function ${}_2F_1(\alpha, \beta; \gamma; 1)$ where the parameters of the numerator are complex conjugates without real part, that is to say, it is a biparametric distribution generated by the ${}_2F_1(bi, -bi; \gamma; 1)$ function, where i is the imaginary unit and b and γ are real non-negative numbers. The functional expression of the probability mass function of the CBPD allows us to obtain explicit expressions for the mean and the variance of the distribution.
We propose the use of the CBPD in order to build a regression model for overdispersed count data analysis. Specifically, we present several models: considering that one of the parameters depends on the covariates through the mean of the model and the other is free, and considering that both parameters depend on the covariates through the mean and the variance. We study the properties of these models and we write several functions in R in order to estimate them and to show results. As illustration, we present an example in the Sports field. We model the variables "number of goals scored" and "number of goals received" by the German football team along all its history and with several explanatory variables, such as the result of the match, the type of match or the location of the match -in Germany or out of Germany- among others. We show the fits obtained by the CBPD regression model and we compare them with those obtained using other useful regression models for count data, such as Poisson, Negative Binomial or Generalized Waring regression models.

References

[Rodríguez-Avi et al (2003)] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J.: A new class of discrete distributions with complex parameters. Statistical Papers 44 (1), 67-88

Delay Equations with Non-negativity Constraints Driven by a Hölder Continuous Function of Order $\beta \in (\frac{1}{3}, \frac{1}{2})$

CS23A
Diffusions & Diff. Eq.

MIREIA BESALÚ*, DAVID MÁRQUEZ-CARRERAS*, CARLES ROVIRA*,†

*Facultat de Matemàtiques, Universitat de Barcelona, Catalunya
†email: carles.rovira@ub.edu 393:carlesrovira.tex,session:CS23A

The purpose of this paper is to study a differential delay equation with non-negativity constraints driven by a Hölder continuous function y of order $\beta \in \left(\frac{1}{3}, \frac{1}{2}\right)$. We will consider the problems of existence, uniqueness and boundedness of the solutions. As an application we will study a stochastic delay differential equations with non-negativity constraints driven by a fractional Brownian motion with Hurst parameter $H \in \left(\frac{1}{3}, \frac{1}{2}\right)$.

More precisely, we consider a delay differential equation with positivity constraints on $\mathbb{R}^d$ of the form:

$$\begin{aligned} x(t) &= \eta(0) + \int_0^t b(s,x)ds + \int_0^t \sigma(x(s-r))dy_s + z(t), \quad t \in (0,T], \\ x(t) &= \eta(t), \qquad t \in [-r, 0], \end{aligned}$$

where r denotes a strictly positive time delay, y is a m-dimensional β-Hölder continuous function with $\frac{1}{3} < \beta < \frac{1}{2}$, $b(s,x)$ the hereditary term, depends on the path $\{x(u), -r \leq u \leq s\}$, while $\eta : [-r,0] \to \mathbb{R}^d_+$ is a non negative smooth function, with $\mathbb{R}^d_+ = \left\{u \in \mathbb{R}^d;\, u_i \geq 0 \text{ for } i = 1,\ldots,d\right\}$ and z is a vector-valued non-decreasing process which ensures that the non-negativity constraints on x are enforced.

CS5D
H-D
Inference

Classification of Higher-Order High-Dimensional Data

ANURADHA ROY*,‡, RICARDO LEIVA†

*Dep. of Management Science and Statistics, The University of Texas at San Antonio, USA,
†Departamento de Matemática, F.C.E., Universidad Nacional de Cuyo, Mendoza, Argentina
‡email: Anuradha.Roy@utsa.edu 394:Roy.tex,session:CS5D

Although devised in 1936 by Fisher (Fisher, 1936), discriminant analysis is still rapidly evolving, as the complexity of contemporary data sets grows exponentially. Our classification rules explore these complexities by modeling various correlation structures in higher-order high-dimensional data (Kroonenberg, 2008). Moreover, our classification rules are suitable to data sets where the number of response variables is comparable or larger than the number of observations. We assume that the higher-order high-dimensional observations have a doubly exchangeable covariance structure and different Kronecker product structures on the mean vectors. Appropriate classification rule needs to be developed that is suitable for a particular data set. The main idea of this talk is to employ the information of a double exchangeability of a variance-covariance matrix for third-order data, which allows partitioning a covariance structure into three unstructured covariance matrices, corresponding to each of the three orders. As a consequence, the number of estimated covariance parameters is substantially reduced, comparing to the classical approach, which enables us to apply the proposed procedures even to a very small number of observations. This is of critical importance to a variety of applied problems with multivariate repeated measures in medicine, biostatistics and social sciences. The new discriminant functions (Leiva and Roy, 2011, 2012) are very efficient in discriminating individuals in small sample scenarios. Iterative algorithms are proposed to calculate the maximum likelihood estimates of the unknown population parameters as closed form solutions do not exist for these unknown parameters. The proposed classification rules are demonstrated on a real medical data set which illustrates the benefits of these new methods over the traditional ones.

References

[Fisher R.A. (1936)] Fisher, R.A., 1936: The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 179–188.

[Kroonenberg, P.M., (2008)] Kroonenberg, P.M., 2008: *Applied Multiway Data Analysis*, John Wiley & Sons, Inc., New Jersey.

[Leiva et al. (2011)] Leiva, R., Roy, A., 2011: Linear discrimination for multi-level multivariate data with separable means and jointly equicorrelated covariance structure, *Journal of Statistical Planning and Inference*, **141(5)**, 1910–1924.

[Leiva et al. (2012)] Leiva, R., Roy, A., 2012: Linear discrimination for three-level multivariate data with separable additive mean vector and doubly exchangeable covariance structure, *Computational Statistics and Data Analysis*, **56(6)**, 1644-1661

Log-linear Models on Non-Product Spaces

CS35A Discretre Response M.

TAMÁS RUDAS[*,‡], ANNA KLIMOVA[†]

[*]Eötvös Loránd University, Budapest, Hungary,
[†]Institue of Science and Technology, Klosterneuburg, Austria
[‡]email: rudas@tarki.hu 395:TamasRudas.tex,session:CS35A

Log-linear models on non-product spaces arise naturally in machine learning (feature selection), in many problems of official statistics (e.g., the analysis of congenital abnormalities) or in market research (market basket analysis). In these cases, not all combinations of the 'Yes' – 'No' categories of the variables are observed: newborns with no congenital abnormalities are not recorded, each purchase consists of, at least, one item.

A classical variant of independence, applicable when the sample space is not a Cartesian product, is the Aitchison – Silvey independence, which assumes that there is no overall effect, that is there is no common effect that would apply to all category combinations or cells. Relational models are common generalizations of these and of standard log-linear models. In a relational model, arbitrary subsets of the cells may feature the same effect and the probability of a cell is the product of the parameters that appear in it. Some of the properties of the families of distributions in this model class are quite surprising, both from the algebraic and the stochastic points of view. Algebraically, the models can be expressed using a set of generalized odds ratios, and if there is no overall effect present, there is exactly one out of these generalized odds ratios, that is non-homogeneous. Poisson and multinomial likelihoods under models without the overall effect are not equivalent, sufficient statistics are not preserved in the maximum likelihood estimates, and the existence of a factorization of a model does not necessarily imply likelihood independence of the components.

The talk will present these properties and give simple illustrations of the applications of relational models. The presentation is based on [Klimova, Rudas & Dobra, 2012].

References

[Klimova, Rudas & Dobra, 2012] Klimova, A., Rudas, T., Dobra, A., 2012: Relational models for contingency tables, *J. Mult. Anal.*, **104**, 159 - 173.

Bayesian Estimation of Thermal Conductivity in Polymethyl Methacrylate

OCS11 ENBIS

ETTORE LANZARONE[*], VALERIO MUSSI[†], SARA PASQUALI[*], FABRIZIO RUGGERI[*,‡]

[*]CNR IMATI, Milano, Italy,
[†]Politecnico di Milano, Piacenza, Italy
[‡]email: fabrizio@mi.imati.cnr.it 396:FabrizioRuggeri.tex,session:OCS11

A Bayesian approach is developed for estimating the thermal conductivity of a homogeneous material from the temperature evolution acquired in few internal points.

Temperature evolution is described by the classical one-dimensional heat equation, in which the thermal conductivity is one of the coefficients. Noisy measurements lead to a partial differential equation with stochastic coefficients and, after discretisation in time and space, to a stochastic differential equation. Euler approximation at sampled points leads to a likelihood function, used in the Bayesian estimation of the thermal conductivity under different prior densities.

An approach for generating latent observations over time in points where the temperature is not acquired is also included.

Finally, the methodology is experimentally validated, considering a heated piece of polymethyl methacrylate (PMMA) with temperature measurements available in few points of the material and acquired at high frequency.

OCS18 Lifetime Data Anal.

Improving the Performance of a Complex Multi-State System through Random Inspections

JUAN ELOY RUIZ-CASTRO*,†

*University of Granada, Granada, Spain
†email: jeloy@ugr.es

397:Ruiz-Castro.tex,session:OCS18

Serious damage and considerable financial losses are caused when a system failure occurs due to poor reliability. Thus, preventive maintenance is useful especially for improving the reliability of a system. But, preventive maintenance produces a cost whose profitability must be analyzed. This work models a complex system that evolves in discrete time with a preventive maintenance policy. The internal performance of the device can pass through several states. These states are grouped according several internal degradation levels. The device is subject to internal failures, repairable and non-repairable and external shocks. The external shocks produce degradation independent of the internal performance. When a threshold is reached then the device undergoes a non-repairable failure. Each time that a non-repairable failure takes place, then the device is replaced by a new and identical one. On the other hand, if the device undergoes an internal repairable failure the device goes to corrective repair.

Random inspections are introduced for improving the reliability measures of the device. When an inspection occurs, then the internal and external degradation level is observed. There are several internal and external degradation levels and, in both cases, only the first degradation level is non-significant. Thus, several preventive repairs are carried out according to the different internal and external degradation levels observed by inspection. If inspection observes internal significant (non-significant) damage and external non-significant (significant) damage then the external (internal) damage state is saved in memory. The significant internal (external) damage is repaired and the device begins working with the same external (internal) non-significant damage. If inspection observes significant internal and external damage, both of them are repaired according to degradation levels. Phase-type distributions, introduced in [1], are assumed in this work. These ones enable us to express the results through well structured matrix blocks. An analysis of block-structured stochastic models is provided in [2]. Phase type distributions have been considered in the modelling of reliability systems that evolves in discrete time. The behaviour of a warm discrete standby system is described in [3].

The system is modelled in an algorithmic form and the transient and stationary distribution is worked out through matrix blocks in a well structured form. Some measures such as the reliability, availability and conditional probability of failure are calculated. Rewards and costs are introduced in the model showing the effectiveness and the profitability of the preventive maintenance policy.

Acknowledgment. This paper is partially supported by the Junta de Andalucía, Spain, under the grant FQM-307 and by the Ministerio de Ciencia e Innovación, España, under Grant MTM2010-20502.

References

[1] Neuts, M.F., 1981: *Matrix geometric solutions in stochastic models. An algorithmic approach*. Baltimore, John Hopkins, University Press.

[2] Quan-Lin, Li, 2010: *Constructive Computation in Stochastic Models with Applications: The RG-Factorization*. Springer.

[3] Ruiz-Castro, J.E. and Fernández-Villodre, G., 2012: A complex discrete warm standby system with loss of units. *European Journal of Operational Research*, **218**, 2, 456-469.

Simulation-Based High Dimensional Experimental Design for Nonlinear Models

OCS9 Design of Experiments

ELIZABETH RYAN*,†, CHRISTOPHER DROVANDI*, HELEN THOMPSON*, ANTHONY PETTITT*

*Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia
†email: eg.ryan@qut.edu.au 398:Elizabeth_Ryan.tex,session:OCS9

The use of Bayesian methodologies for solving optimal experimental design problems has increased in the last few years. Many of these methods have been found to be computationally intensive for high dimensional design problems. Here we present a simulation-based approach that can be used to solve high dimensional optimal design problems. Our approach involves the use of lower dimensional parameterisations that consist of two design variables, which generate multiple design points. Using this approach, one simply has to search over two design variables, rather than searching for a large number of optimal design points, thus providing substantial computational savings. We demonstrate our methodologies on applications that come from pharmacokinetic studies and chemistry, and involve nonlinear models. We also compare and contrast several Bayesian and pseudo-Bayesian design criteria, as well as several different lower dimensional parameterisation schemes for generating the high dimensional designs.

Simulation of Diffusion Bridges with Application to Statistical Inference for Stochastic Differential Equations

IS25 Stat. SDE

MICHAEL SØRENSEN*,†

*University of Copenhagen, Denmark
†email: michael@math.ku.dk 399:MichaelSorensen.tex,session:IS25

A simple method for simulating diffusion bridges is presented. A diffusion bridge is a solution to a stochastic differential equation in an interval, where the starting point as well as the end point are fixed. Diffusion bridges play a crucial role in simulation-based likelihood and Bayesian inference for stochastic differential equations. The new method consists in constructing a good approximation to a diffusion bridge from two diffusions, one moving forward in time, the other backward. These two basic diffusions can be simulated by a simple method like the Milstein scheme. The approximate diffusion bridge is then used as a proposal for a pseudo-marginal MCMC algorithm with exact diffusion bridges as the target distribution. The algorithm works particularly well for long time intervals, where other algorithms for simulation of diffusion bridges tend not to work. In fact, the computer time increases linearly with the length of the interval.

The usefulness of the new simulation method to inference for stochastic differential equations is demonstrated by an application to estimation for data that are integrals of a diffusion process observed with measurement error. This is a potential model for ice-core data on paleo-temperatures. The data can be viewed as incomplete observations from a model with a tractable likelihood function. Therefore a simulated EM-algorithm is used to obtain maximum likelihood estimates of the model parameters. The new algorithm for simulating diffusion bridges forms an essential part of the estimation method, where it is used to simulate the full hidden data given the observations. In a simulation study the proposed method works well.

Acknowledgment. The lecture is based on joint work with Mogens Bladt and Fernando Baltazar-Larios.

References

[1] Andrieu, C. and Roberts, G. (2009): A pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

[2] Bladt, M. and Sørensen, M. (2012): Simple simulation of diffusion bridges with application to likelihood inference for diffusions. To appear in *Bernoulli*.

[3] Baltazar-Larios, F. and Sørensen, M. (2010): Maximum likelihood estimation for integrated diffusion processes. In Chiarella, C. and Novikov, A. (eds.): *Contemporary Quantitative Finance: Essays in Honour of Eckhard Platen*, Springer, Heidelberg, pp. 407–423.

CS12A Hierarchica Bayesian

Bayesian Hierarchical Model for Genetic Association with Multiple Correlated Phenotypes

HABIB SAADI[*,‡], ALEXANDRA LEWIN[*], LEONARDO BOTTOLO[*], SYLVIA RICHARDSON[†]

[*]Imperial College, London,
[†]MRC Biostatistics Unit, Cambridge
[‡]email: h.saadi@imperial.ac.uk 400:HabibSaadi.tex,session:CS12A

Genetic association studies are used to find regions of the genome associated with phenotypes of interest. Most association studies to date have been carried out for a single phenotype, or for small combination of phenotypes, with a single regression being performed for each marker. Here we study a data set in which the phenotypes are 127 metabolites measured by NMR in a birth cohort of 4000 people. The genetic associations with these phenotypes have been analysed previously one at a time, ie separate regressions have been done for each marker x phenotype combination. See for example Tukiainen et al. 2012 for a recent study.

Rather than carrying out thousands of separate univariate analyses, we model all the data simultaneously. We use the Bayesian model HESS (Hierarchical Evolutionary Stochastic Search) developed by Bottolo et al. 2011, for detecting genetic regulation of gene expression, in particular for finding genetic markers which are associated with multiple expression phenotypes. The hierarchical formulation incorporates a regression model for each phenotype against all the genetic markers, with a sparsity prior to induce variable selection amongst the markers. The Bayesian hierarchical model treats the multiple phenotypes in parallel, enabling information sharing across phenotypes, whilst allowing for different markers to be associated with different phenotypes.

The data set used here exhibits a strong correlation structure between the phenotypes, linked to their related biological function. In this work we investigate the performance of the hierarchical model for analysing correlated outcomes, discuss a range of analysis strategy where the correlation of the phenotypes is exploited, and develop methods for assessing statistical importance of selected markers in the presence of correlated data.

Given the size of the data, the estimation of the model is performed thanks to a C++ code running on a GPU.

References

[Bottolo et al. (2011)] Bottolo L., Petretto E., Blankenberg S., Cambien F., Cook S., Tiret L. and Richardson S., 2011: Bayesian Detection of Expression Quantitative Trait Loci Hot Spots, *Genetics*, **189**, 1449 - 1459.

[Tukiainen et al. (2012)] Tukiainen T, Kettunen J, Kangas AJ, Lyytikäinen LP, Soininen P, Sarin AP, Tikkanen E, O'Reilly PF, Savolainen MJ, Kaski K, Pouta A, Jula A, Lehtimäki T, Kähönen M, Viikari J, Taskinen

MR, Jauhiainen M, Eriksson JG, Raitakari O, Salomaa V, Järvelin MR, Perola M, Palotie A, Ala-Korpela M, Ripatti S., 2012: Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci, *Hum Mol Genet*, **21(6)**, 1444 - 1455.

A New Automatic Set Estimation Method for the Support

CS30A Inf. on Distributions

ALBERTO RODRÍGUEZ-CASAL*, PAULA SAAVEDRA-NIEVES*,†

*University of Santiago de Compostela, Spain

†email: paula.saavedra@usc.es 401:PaulaSaavedra-Nieves.tex,session:CS30A

This work deals with the problem of estimating the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X from independent and identically distributed observations, $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, taken in it.

Several proposals for reconstructing S have been considered in the literature. For instance, Devroye-Wise (1980) proposed

$$S_n = \bigcup_{i=1}^{n} B_{\epsilon_n}[X_i]$$

as an estimator of S, where $B_{\epsilon_n}[X_i]$ denotes the closed ball with center X_i and radius ϵ_n depending only on n. More sophisticated estimators can be used if we have some additional information on the set. For example, if we know that S is convex then the convex hull of the sample is a natural support estimator. But convexity assumption may be too restrictive in practice. For instance, if S is not connected. Then, it is necessary to consider a more flexible shape restrictions such as $r-$convexity. A closed set $A \subset \mathbb{R}^d$ is said $r-$convex, for $r > 0$, if $A = C_r(A)$, where

$$C_r(A) = \bigcap_{\{B_r(x): B_r(x) \cap A = \emptyset\}} (B_r(x))^c$$

denotes the $r-$convex hull of A and $B_r(x)$, the open ball with center x and radius r. Furthermore, if A is $r-$convex then it is $\overline{r}$-convex for all $\overline{r} \leq r$. So, if we assume that S is r-convex then a natural support estimator will be the $r-$convex hull of the sample points, $C_r(\mathcal{X}_n)$. In Rodríguez-Casal (2007), it is proved that if r is correctly chosen, the r-convex hull of the sample achieves the same convergence rates (in hausdorff and distance in measure) as the convex hull. But in practice, S is unknown and, consequently, the real value of the smoothing parameter r too. In this work, we propose an almost sure consistent estimator of the largest value of r such that S is $r-$convex, under the assumption that the distribution of X is uniform on S. The estimator of r is based on the uniformity test proposed by Berrendero et al. (2012). The support estimator obtained from this smoothing parameter, which is fully automatic given the random sample, is able to achieve the same convergence rates as the convex hull for estimating convex sets.

Acknowledgment. This work has been supported by Project MTM2008-03010 from the Spanish Ministry of Science and Innovation and the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences) from Belgian Science Policy.

References

[Berrendero et al. (2012)] Berrendero, J. R., Cuevas, A. and Pateiro-López, B., 2012: A multivariate uniformity test for the case of unknown support, *Can. J. Stat.*, **40**, 378 - 395.

[Devroye et al. (1980)] Devroye, L. and Wise, G. L., 1980: Detection of abnormal behavior via nonparametric estimation of the support, *SIAM J. Appl. Math.*, **38**, 480 - 488.

[Rodríguez-Casal (2007)] Rodríguez-Casal, A., 2007: Set estimation under convexity type assumptions, *l'I.H.P.- Probabilités & Statistiques*, **43**, 763 - 774.

POSTER
Poster

Saddlepoint Approximation for the Density of Regression Quantiles

RADKA SABOLOVÁ*

*Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
†email: sabolova@karlin.mff.cuni.cz 402:Sabolova.tex,session:POSTER

Let $\mathbf{Y_n} = (Y_1, \ldots, Y_n)$ be observations satisfying the linear regression model

$$\mathbf{Y_n} = \mathbf{X_n}\boldsymbol{\beta} + \mathbf{e_n}$$

where $\mathbf{e_n}$ are i.i.d. errors with unknown distribution function F, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is unknown parameter, $\mathbf{X_n}$ is known $n \times (p+1)$ matrix with $x_{i0} = 1$ for $i = 1, \ldots, n$. α-regression quantile $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ is a solution of minimization problem

$$\text{argmin}_{\mathbf{b}} \sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{X_{ni}}\mathbf{b}),$$

where

$$\rho_\alpha(z) = |z|\{\alpha \mathrm{I}[z > 0] + (1-\alpha)\mathrm{I}[z < 0]\}, \quad z \in \mathbb{R}$$

and $\mathbf{X_{ni}}$ denotes i-th row of matrix $\mathbf{X_n}$.

Averaged regression quantile $\overline{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$ is equal to

$$\overline{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = \overline{\mathbf{X}}_{\mathbf{n}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$$

where

$$\overline{\mathbf{X}}_{\mathbf{n}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X_{ni}}.$$

Quantile regression enables us to study the conditional quantiles of a dependent random variable and therefore offers a more complex view at the behavior of random variables. By appropriate choice of quantile we can study not only conditional median but also behavior in tails and get an idea about variability of studied random variable.

Saddlepoint approximations for the density were firstly introduced in statistics approximately 60 years ago. As they provide us very precise approximations even for small sample sizes, it is of interest to use them for approximating densities of various estimators.

In this contribution, saddlepoint approximation of the density for regression quantiles will be derived, as well as the density of averaged regression quantiles. The theoretical results will be accompanied by numerical simulations.

POSTER
Poster

Superprocesses and Related Lattice Approximations as Models of Information Dissemination Between Mobile Devices

LAURA SACERDOTE*,‡, FEDERICO POLITO*, MATTEO SERENO†, MICHELE GARETTO†

*Dipartimento di Matematica, Università di Torino, Italy,
†Dipartimento di Informatica, Università di Torino, Italy.
‡email: laura.sacerdote@unito.it 403:Sacerdote.tex,session:POSTER

New technologies, services and applications of internet come into play at a high rate and there is an increasing need to update existing quantitative methods and performance evaluation tools to

manage these new trends. Use of mobile devices to substitute or to support servers is a possible future direction to disseminate information. Performance evaluations request answers on a set of topics such as their ability to achieve city level coverage, the estimation of the time delay to reach far away users, the presence of zones not attained by the signal. However the analysis implies a substantial change of perspective with respect to previous approaches, which are often based on Markov chain simulations. The number of involved devices prevents the use of Markov chains for the analysis of interest and suitable continuous limits become necessary Here we present some preliminary ideas to deal with these problems. We propose to model the information dissemination between mobile devices through superprocesses and we discuss a possible approach to their simulation.

On the Relascope: A New Graphical Tool Leading to Formal Tests

CS36A Graphical Methods

JEAN-BAPTISTE AUBIN[*,†], SAMUELA LEONI-AUBIN[*,†,‡]

[*]Institut National des Sciences Appliquées de Lyon, Villeurbanne, France,
[†]Institut Camille Jordan, Lyon, France
[‡]email: samuela.leoni@insa-lyon.fr 404:SamuelaLeoni-Aubini.tex,session:CS36A

A new graphical tool, the *relascope*, is presented. The relascope possesses several applications: for example, independence test or homoscedasticity test in univariate regression model can be developped. The *relascope* is constructed the following way. Let's consider n realisations of (X, Y): $(x_1, y_1), \ldots, (x_n, y_n)$ where X and Y are univariate continuous random variables. In the scatterplot, for each real δ, the length of the longest run of observations consecutively (with respect to X) on the same side of the horizontal line $y = \delta$, noted $L_n(\delta)$, is computed. Then, the plot obtained by putting δ in x-axis and $L_n(\delta)$ in y-axis is called the *relascope*.

Some formal tests can be derived from the relascope. The key idea of one of them is based on the fact that, if Y depends on X, it will exist a level δ for which the number of consecutive observations (with respect to X) that are over/under δ, $L_n(\delta)$, is significatively different from the analogous quantity under the null hypothesis of independence $\mathcal{H}_0$. When $\mathcal{H}_0$ holds, the exact probability distribution of $L_n(\delta)$ and a $p-$value $p_n(\delta)$ is deduced for each δ (Aubin and Leoni-Aubin (2011)). Then, the minimum of the obtained p-values: $p_{min} = \min_\delta p_n(\delta)$ is considered. To test whether this dependence is significant, p_{min} is compared to the empirical distribution of p_{min}'s obtained by considering various samples of general form $(x_1, y_{p(1)}), \ldots, (x_n, y_{p(n)})$ where $(p(1), \ldots, p(n))$ is a permutation of $(1, \ldots, n)$. If the probability to be more extreme than p_{min} is smaller than the nominal level α, then the null hypothesis of independence is rejected.

Another kind of test can be performed when studying homoscedasticity of a time series putting X as time and Y as the squared residuals and following the same process. Some applications of these tests are presented.

References

[Aubin and Leoni-Aubin (2011)] Aubin, J.-B., Leoni-Aubin, S., 2011: A nonparametric lack-of-fit test for heteroscedastic regression models, *Comptes Rendus Mathématique*, **349**, 215 - 217.

IS10 High-Dim. Inference

Independent Component Analysis via Nonparametric Maximum Likelihood

RICHARD J. SAMWORTH[*,‡], MING YUAN[†]

[*]University of Cambridge, Cambridge, United Kingdom,
[†]University of Wisconsin–Madison, Madison, United States of America
[‡]email: r.samworth@statslab.cam.ac.uk 405:RichardSamworth.tex,session:IS10

Independent Component Analysis (ICA) models are very popular semiparametric models in which we observe independent copies of a random vector $X = AS$, where A is a non-singular matrix and S has independent components. We propose a new way of estimating the unmixing matrix $W = A^{-1}$ and the marginal distributions of the components of S using nonparametric maximum likelihood. Specifically, we study the projection of the empirical distribution onto the subset of ICA distributions having log-concave marginals. We show that, from the point of view of estimating the unmixing matrix, it makes no difference whether or not the log-concavity is correctly specified. The approach is further justified by both theoretical results and a simulation study.

OCS19 Multivar. funct. data

Spatial Functional Data Analysis

LAURA M. SANGALLI[*]

[*]MOX - Dipartimento di Matematica, Politecnico di Milano, Italy
[†]email: laura.sangalli@polimi.it 406:SangalliLauraM.tex,session:OCS19

In this talk I will describe a novel class of models for the accurate estimation of surfaces and spatial fields. The proposed models have a generalized additive framework with a differential penalization, and merge advanced statistical methodology with numerical analysis techniques. Thanks to the combination of potentialities from these two scientific areas, this new class of models has important advantages with respect to classical techniques used in spatial data analysis. The models are able to efficiently deal with data distributed over irregularly shaped domains, with complex boundaries, strong concavities and interior holes. Moreover, they can comply with specific conditions at the boundaries of the domain, which is fundamental in many applications to obtain meaningful estimates. The proposed models can also deal with data scattered over non-planar domains. Moreover, they have the capacity to incorporate problem-specific priori information about the spatial structure of the phenomenon under study. Space-varying covariate information is also included via a semiparametric framework. The estimators have a penalized regression form, they are linear in the observed data values and the usual inferential tools are available. A generalized linear approach allows also for link functions other than linear, as for instance the logit, further enhancing the very broad potential use of these class of models. Many important extensions can be envisioned, as for instance to space-time data and to volume data. The use of numerical analysis techniques, and specifically of finite elements, makes the models computationally very efficient.

The seminar is based on joint work with James Ramsay (McGill University), and with Laura Azzimonti, Bree Ettinger, Fabio Nobile, Simona Perotto and Piercesare Secchi (Politecnico di Milano). Funding by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca 2008* research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" (http://mox.polimi.it/users/sangalli/firb/), and by the program Dote Ricercatore Politecnico di Milano - Regione Lombardia, research project "Functional data analysis for life sciences".

Estimation of the Transition Density of a Markov Chain

CS6C Func. Est., Smooth-ing

MATHIEU SART*,†

*Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, France

†email: msart@unice.fr 407:MathieuSart.tex,session:CS6C

We are interested in estimating the transition density of an homogeneous Markov chain $(X_i)_{i\geq 0}$. This statistical setting has been introduced in 1969 by Roussas who considered a quotient estimator, that is, an estimator based on the division of an estimator of the joint density of (X_i, X_{i+1}) by an estimator of the density of X_i. This kind of estimator suffers however from the fact that their rates of convergence depend on the properties of the density of X_i. They may thus converge slowly even when the transition density is smooth. Since [Clémencon (2000)], other kind of estimators have been studied to overcome this issue, but, as far as we know, they require all the knowledge (or at least a suitable estimation) of various quantities depending on the unknown transition density. Besides, these quantities not only influence the way the estimators are built but also their performances since they are involved in the risk bounds.

The aim of this talk is to present an adaptive procedure introduced in [Sart (2013)] that is fully data-driven and almost assumption-free. Our estimation strategy is a mixture between an approach based on the minimization of a contrast and an approach based on robust tests. We propose a new way of using a test to select among a family of piecewise constant estimators when the partitions ensue from an adaptive approximation algorithm. This new procedure can be interpreted as being an implementable version of the test procedures of [Baraud and Birgé (2009), Baraud (2011)]. We shall prove an oracle-type inequality from which we shall deduce uniform rates of convergence over ball of (possibly) inhomogeneous Besov spaces with possibly small regularity index. These rates coincide, up to a possible logarithmic factor to the usual ones over such classes. Finally, we carry out numerical simulations and compare our estimator with the one of [Akakpo and Lacour (2011)].

References

[Akakpo and Lacour (2011)] Akakpo, N., Lacour, C., 2011: Inhomogeneous and anisotropic conditional density estimation from dependent data, *Electronic Journal of Statistics*, **5**, 1618–1653.

[Baraud (2011)] Baraud, Y., 2011: Estimator selection with respect to Hellinger-type risks, *Probability Theory and Related Fields*, **151**, 353–401.

[Baraud and Birgé (2009)] Baraud, Y., Birgé, L., 2009: Estimating the intensity of a random measure by histogram type estimators, *Probability Theory and Related Fields*, **143**, 239–284.

[Clémencon (2000)] Clémencon, S., 2000: Adaptive estimation of the transition density of a regular Markov chain, *Mathematical Methods of Statistics*, **9**, 323–357.

[Sart (2013)] Sart, M., 2013: Estimation of the transition density of a Markov chain, *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, To appear.

Averaging across Asset Allocation Models

NYA Not Yet Arranged

PETER SCHANBACHER*,†

*University of Konstanz, Germany

†email: peter.schanbacher@uni-konstanz.de 408:PeterScahnbacher.tex,session:NYA

Combination of asset allocation models is rewarding if (i) the applied risk function is concave and (ii) there is no dominating model. We show that most common risk functions are either concave or at least concave in common applications. In a large empirical study using standard asset allocation

models we find that there is no constantly dominating model. The ranking of the models depends on the data set, the risk function and even changes over time. We find that a simple average of all asset allocation models can outperform each individual model. Our contribution is twofold. We state an explanation why the combination is expected to dominate most individual models. In a large empirical study we show that model combinations perform exceptionally well in asset allocation.

CS25A Stoch. Finance I.

A Representation Theorem for Smooth Brownian Martingales

HENRY SCHELLHORN*,†

*Claremont Graduate University, USA

†email: Henry.Schellhorn@cgu.edu 409:SchellhornHenry.tex,session:CS25A

We show that, under certain smoothness conditions, a Brownian martingale, when evaluated at a fixed time, can be represented as an exponential of its value at a later time. The time-dependent generator of this exponential operator is equal to one half times the Malliavin derivative. This result can also be seen as a generalization of the semi-group theory of parabolic partial differential equations to the parabolic path-dependent partial differential equations introduced by Dupire (2009) and Cont and Fournié (2011). The exponential operator can be calculated explicitly in a series expansion, which resembles the Dyson series of quantum mechanics. Our continuous-time martingale representation result is proved by a passage to the limit of a special case of a backward Taylor expansion of an approximating discrete-time martingale. The latter expansion can also be used for numerical calculations. This result can be extended to martingales defined as the conditional expectation of a functional of fractional Brownian motion. We are currently investigating whether it can be extended to functionals of multifractional Brownian motion. We present an application to bond pricing in finance.

References

[Cont et al. (2010)] Cont, R., Fournié, D.-A., 2010: Change of variable formulas for non-anticipative functionals on path space. *Journal of Functional Analysis* **259**, 1043-1072.

[Dupire, B. (2009)] Dupire, B., 2009: Functional Ito Calculus. Portfolio Research Paper 2009-04, Bloomberg.

IS17 Random Matrices

Optimal Estimates on the Stieltjes Transform of Wigner Matrices

CLAUDIO CACCIAPUOTI*, ANNA MALTSEV†, BENJAMIN SCHLEIN*,‡

*Institute of Applied Mathematics, University of Bonn, Germany,

†School of Mathematics, University of Bristol, UK

‡email: benjamin.schlein@hcm.uni-bonn.de 410:Schlein.tex,session:IS17

We consider the Stieltjes transform $m(z)$ of a Wigner matrix and compare it with the Stieltjes transform m_{sc} of the semicircle law. We show that $|m(E+i\eta) - m_{sc}(E+i\eta)| \geq K/(N\eta)$ with probability smaller than any power of K^{-1}. Similar results have already been established by Erdös, Knowles, Yau and Yin. With respect to their work, we eliminate logarithmic corrections, proving the convergence with the optimal rate up to the optimal scale $\eta \simeq N^{-1}$. We will discuss the implications of these bounds.

Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression Models

CS5B H-D Distribution

BENEDIKT M. PÖTSCHER*, ULRIKE SCHNEIDER†,‡

*University of Vienna, †Vienna University of Technology
‡email: ulrike.schneider@tuwien.ac.at 411:UlrikeSchneider.tex,session:CS5B

We study the distribution of hard-, soft-, and adaptive soft-thresholding estimators within a linear regression model where the number of parameters k can depend on sample size n and may diverge with n. In addition to the case of known error-variance, we define and study versions of the estimators when the error-variance is unknown. We derive the finite-sample distribution of each estimator and study its behavior in the large-sample limit, also investigating the effects of having to estimate the variance when the degrees of freedom $n - k$ does not tend to infinity or tends to infinity very slowly. Our analysis encompasses both the case where the estimators are tuned to perform consistent model selection and the case where the estimators are tuned to perform conservative model selection. Furthermore, we discuss consistency, uniform consistency and derive the minimax rate under either type of tuning.

Acknowledgment. This research was partially supported by the Deutsche Forschungsgemeinschaft project FOR916.

Bootstrap Confidence Intervals of Hedonic Price Indices: An Empirical Study with Housing Data

CS25C Stoch. Finance II.

MICHAEL BEER*, OLIVIER SCHÖNI*,†

*University of Fribourg, Department of Quantitative Economics, Switzerland
†email: olivier.schoeni@unifr.ch 412:OlivierSchoeni.tex,session:CS25C

Hedonic price indices are designed to accurately measure price changes by holding the quality of the considered goods constant. This goal is achieved by means of regression techniques relying on the so-called hedonic hypothesis: Each good is considered as a bundle of characteristics and its price, therefore, solely depends on these characteristics. In each time period the observed prices are thus regressed on the goods' characteristics and the period specific hedonic price functions are estimated. The estimated hedonic price functions are subsequently used to compute price changes according to classic price index formulae. Unfortunately, besides the obvious advantages deriving from the use of a price index axiomatic, hedonic price indices inherit one important drawback, namely the price index problem: None of the suggested price index formulae seem to be superior to others, which complicates the choice of a specific index formula.

The present paper aims to exploit the stochastic nature of hedonic regressions to gauge the statistical properties of commonly used hedonic price indices, thus providing a complementary instrument to choose among different price index formulae. This aim is attained by comparing the mean confidence interval width of a given price index over a considered time period to the mean confidence interval width of other hedonic price indices. Since hedonic price indices are nonlinear functions of the hedonic price functions' parameters, usual confidence interval formulas do not apply. Moreover, hedonic regressions are often plagued by heteroskedasticity, making the confidence intervals' estimation difficult. We therefore suggest using the wild bootstrap technique to compute confidence intervals of hedonic price indices.

For the present paper we use transaction prices and characteristics of single-family houses observed in the Swiss canton of Zurich from the first quarter of 2001 to the fourth quarter of 2011. The data

were kindly provided by Wüest & Partner, an international consultancy firm for real estate. For each quarter, the data are collected from insurances, banks, and other real estate agencies, providing information for more than 50% of the transactions occurred in the canton of Zurich.
It is shown that the obtained confidence interval widths, in conjunction with the axiomatic approach, allow to considerably reduce the number of hedonic price index formulae the price statistician has to consider.

IS5 Envtl. Epidem. Stat.

Healthy Environment, Healthy People—Making the Statistical Connections

MARIAN SCOTT[*,‡], DANIELA COCCHI[†]

[*]School of Mathematics and Statistics, University of Glasgow, UK,
[†]Department of Statistical Sciences, University of Bologna, Italy
[‡]email: Marian.Scott@glasgow.ac.uk 413:MarianScott.tex,session:IS5

Over the past decade or so, there has been an increasing growth in the use of indicators and indices for policy, management and communication purposes around the state of the environment. "Environmental indicators have a crucial role to play in the simplification, quantification, standardisation and communication of environmental conditions to regulators and policy- makers" (Johnson, ICES, 2008). Environmental indicators such as the European 2010 Biodiversity indicators, which are 26 indicators, grouped into ecosystem integrity, goods and services, sustainable use, status and trends of components of biodiversity, threats to biodiversity provides a good example of the diverse nature of the many indicators that exist.

At the same time, there has been a growth in 'softer' indicators for wellbeing and happiness including Sustainable Economic Welfare, Genuine Progress Indicator, Living Planet Index, Human Development Index, Happy Planet Index, and the Ecological Footprint.

The OECD (2008) defines a composite environmental index as a set of aggregated or weighted parameters or indicators. A composite index is created by combining individual indicators - there may or may not be an underlying model. Composite indices are usually created to measure multidimensional or latent concepts, e.g. sustainability, wellbeing or ecosystem health. Examples include an index of multiple deprivation (SIMD), which combines 37 indicators in 7 domains- employment, income, health, education, housing widely used in the United Kingdom and the environmental performance indicator (EPI) first produced by Yale University. A composite index helps provide insight into the overall state of the environment (an ecosystem approach), recognising that ecological systems are connected, involving-physical, chemical and biological factors, therefore spatial and temporal scale will be important in the aggregation rule. The idea of latent concepts leads naturally to latent variable modelling such as factor analysis.

Naturally, the wish is to link together a 'healthy' environment and human health and wellbeing. In this paper we will consider the statistical models that underpin some at least of the composite indices of environment and wellbeing and consider their statistical properties.

NYA Not Yet Arranged

Collinearity and Micronumerosity: A New Ridge Regression Approach

PEDRO MACEDO[*], MANUEL SCOTTO[*,†]

[*]Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, Portugal
[†]email: mscotto@ua.pt 414:ManuelScotto.tex,session:NYA

Collinearity and micronumerosity are two major concerns in statistics. They are responsible for

inflating the variance associated with the regression coefficients estimates and, in general, may affect the signs of the estimates, as well as statistical inference. If there are cases where it is possible to collect additional information (normally requiring more time and cost), there are other cases where such additional information simply does not exist. In either case, it is necessary to make the best possible predictions with such limited information.

Despite more recent approaches, ridge regression still plays a key role in regression models affected by collinearity and micronumerosity, and outperforms other competitors in many problems. However, the challenge in ridge regression remains the selection of the ridge parameter. This choice is usually made by the inspection of the ridge trace or by a formal method. Recently, the Ridge-GME parameter estimator appears in the literature as one of the best ridge parameter estimators, although requiring some subjective information from visual inspection of the ridge trace.

In this talk, the Ridge-GME estimator will be developed so that no subjective information is needed to define the ridge interval or the supports for the parameters. A simulation study and an empirical application will be used to illustrate the performance of this new version of the Ridge-GME estimator. A complete MATLAB code for the Ridge-GME estimator will be also available and discussed.

Acknowledgment. This work was supported by *FEDER* funds through *COMPETE*–Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

References

[Macedo et al. (2010)] Macedo, P., Scotto, M., Silva, E., 2010: On the choice of the ridge parameter: a maximum entropy approach, *Commun. Stat. - Simul. C.*, **39**, 1628 - 1638.

Adaptive Bayesian Density Estimation Using General Kernel Mixtures

CS8B Bayesian Nonpar.

CATIA SCRICCIOLO*

*Bocconi University, Milan, Italy

†email: catia.scricciolo@unibocconi.it 415:CatiaScricciolo.tex,session:CS8B

We consider Bayesian nonparametric estimation of the probability density generating a sample of observations, under the assumption that the density is smooth. A prior distribution is constructed on a collection of location mixtures of Gaussians or more general scale-family of kernels, endowing the mixing distribution with a Pitman-Yor or a normalized inverse-Gaussian process prior and the scale parameter with an inverse-Gamma distribution, the mixing distribution and the scale being independent. The main results provide rates of convergence for estimating, in all L_p-norms, densities in a scale of regularity spaces characterized by tail bounds on the Fourier transform. The obtained rates are minimax-optimal, up to a logarithmic factor, and are achieved under a prior that does not depend on the unknown parameter quantifying the regularity level of the sampling density. The resulting hierarchical Bayes procedure, with a fixed prior, is thus rate adaptive. Results rely on a novel approximation based on the expansion of an integral transform of convolution-type with the sinc-kernel, which is fundamental in understanding the behaviour of the procedure.

Acknowledgment. This research was mainly supported by grants from Bocconi University and the Italian Ministry of Education, University and Research, MIUR Grant No. 2008MK3AFZ003.

References

[de Jonge, van Zanten (2010)] de Jonge, R., van Zanten, J. H., 2010: Adaptive nonparametric Bayesian inference using location-scale mixture priors, *Ann. Statist.*, **38**, 3300 - 3320.

[Kruijer et al. (2010)] Kruijer, W., Rousseau, J., van der Vaart, A., 2010: Adaptive Bayesian density estimation with location-scale mixtures, *Electron. J. Statist.*, **4**, 1225 - 1257.

[van der Vaart, van Zanten (2009)] van der Vaart, A. W., van Zanten, J. H., 2009: Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth, *Ann. Statist.*, **37**, 2655 - 2675.

NYA
Not Yet Arranged

Alternative Agreement Coefficients Between Two Continuous Measurements

MUSTAFA SEMIZ*, NESLIHAN IYIT*

*Department of Statistics, Faculty of Science, Selcuk University, 42031, Campus, Konya, Turkey

`416:mustafasemiz.tex,session:NYA`

Accurate and precise measurement is an important isssue in any study and in any scientific area. In this study, the proposed similarity (MS) coefficients are based on the contribution of each observational discrepancies as alternative agreement coefficients. Both proposed MS coefficients are as a function of agreement level of observations for continuous data. Also, the calculation and the interpretation of these proposed agreement coefficients are very simple. Finally, the coefficients are presented in some special cases and all coefficients are applied to a real data example.

CS16A
Empirical processes

Testing for Positive Cure-Rate under Random and Case-1 Interval Censoring

ARUSHARKA SEN*,†

*Concordia University, Montreal, Canada

†`email: arusharka.sen@concordia.ca`

`417:ArusharkaSen.tex,session:CS16A`

Consider a sample of individuals on each of whom some sort of *time-to-event* data is being collected, for instance, onset time of a disease following exposure to infection, time to death under a terminal disease etc. In most such cases, there may be a possibility that the individual may be *immune* (e.g., not catch a disease) or get *cured* (e.g., cured of a disease). Cure is usually quantified by the probability of cure, or the *cure-rate*: $p = P\{X = \infty\}$, where X is the time-to-event of interest. In this paper we study the problem of testing for positive cure-rate, i.e., testing $H_0 : p = 0$ vs. $H_1 : p > 0$, under *random* as well as *Case-1 interval censoring*. The choice of appropriate test-statistics becomes tricky here, as both of the known nonparametric estimators under these two models (Maller and Zhou (1996) and Sen and Tan (2008), respectively) have *degenerate* limiting distributions under $H_0 : p = 0$. We obtain test-statistics under both models using the same idea as in Sen and Tan (2008), namely that of Poisson convergence of the censored empirical processes when the data are in the max domain of attraction of some extreme-value distribution. We also assume the Koziol-Green model of censoring in both cases. Performance of the above tests will be illustrated using simulated data.

Acknowledgment. This research was supported by a Discovery grant of NSERC, Canada.

References

[1] Maller, R.A. & Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley, Chichester.

[2] Sen, A. & Tan, F. (2008). Cure-rate estimation under Case-1 interval censoring. *Statist.Methodology* **5**, 106–118.

Model Selection as a Decision Problem

CS9C Model Selection

RAFFAELLO SERI*,‡, CHRISTINE CHOIRAT†

*Università degli Studi dell'Insubria, Varese, Italy,
†Universidad de Navarra, Pamplona, Spain
‡email: raffaello.seri@uninsubria.it 418:RaffaelloSeri.tex,session:CS9C

Model selection is the task of selecting a statistical model from a set of candidate models, on the basis of data. In a frequentist setting, several approaches have been proposed to perform this task. However, most treatments of the model selection problem are either restricted to special situations (lag selection in AR, MA or ARMA models, regression selection, selection of a model out of a nested sequence) or to special selection methods (selection through testing or penalization). Our aim is to provide some basic tools for the analysis of model selection as a statistical decision problem, independently of the situation and of the method used. In order to achieve this objective, we embed model selection in the theoretical framework offered by decision theory.

First of all, we analyze what a "best" model should be. In order to do so, we introduce a preference relation on the collection of models under scrutiny; this relation is defined in terms of the asymptotic goodness-of-fit (as measured by a function $Q_{\infty,j}\left(\theta_j^*\right)$ for $j = 1, \ldots, J$, that can be a likelihood, a generic objective function or a measure of forecasting performance) and of their parsimony (as measured by the number of parameters p_j). The preference relation, written as $\blacktriangleright$, is a lexicographic order in which model j is preferred to model i if it has a higher value of the objective function or if the two have the same value of the objective function but j is more parsimonious than i. The "best" model is defined as the set of majorants of the relation $\blacktriangleright$.

We then pass to analyze the finite-sample situation. We define a property of a model selection procedure, that we call "consistency" and corresponds to the fact that the best model is asymptotically selected with probability 1. Then we derive a condition that allows for reducing the asymptotic analysis of a model selection procedure to the case in which only two models are in competition. As concerns the situation in a finite sample, we show that model selection through penalization of the objective function $Q_{n,j}\left(\hat{\theta}_j\right)$ arises in a natural way from some properties of the preference relation $\blacktriangleright$. Model selection through pairwise penalization and through testing are also briefly reviewed.

As a major application of our framework, we derive necessary and sufficient conditions that an information criterion has to satisfy in the case of independent and identically distributed realizations in order to deliver asymptotically with probability 1 the "best" model out of a class of J models. It turns out that the bounds on the model selection procedures arise as strong limit theorems (Laws of Large Numbers and Laws of the Iterated Logarithm, for sums and $V-$statistics) associated with the weak limit theorems (Central Limit Theorems) that constitute the basis of Vuong's model selection procedures based on likelihood ratio tests. "In probability" versions of the previous results are also discussed.

Nonparametric Estimation of Regression Level Sets Using Kernel Plug-in Estimator

CS6A Funct. Est., Kernel Meth.

THOMAS LALOË*, RÉMI SERVIEN†,‡

*Université de Nice, France, †UMR Toxalim, INRA, France
‡email: remi.servien@toulouse.inra.fr 419:Servien.tex,session:CS6A

We consider the problem of estimating the level sets of a regression function. More precisely,

consider a random pair (X, Y) taking values in $\mathbf{R}^d \times J$, where $J \subset \mathbf{R}$ is supposed to be bounded. The goal of this paper is then to build an estimator of the level sets of the regression function r of Y on X, defined for all $x \in \mathbf{R}^d$ by $r(x) = \mathbb{E}[Y|X = x]$. For $t > 0$, a level set for r is defined by

$$\mathcal{L}(t) = \{x \in \mathbf{R}^d : r(x) > t\}.$$

Assume that we have an independent and identically distributed sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ with the same distribution as (X, Y). We then consider a plug-in estimator of $\mathcal{L}(t)$. More precisely, we use the consistent kernel estimator r_n of r. Assume that we can write $r(x) = \varphi(x)/f(x)$, where f is the density function of X, and φ is defined by $\varphi(x) = r(x)f(x)$. Let K be a kernel on $\mathbf{R}^d$. We define, for all $x \in \mathbf{R}^d$,

$$\varphi_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_n}\right) \text{ and } f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right).$$

For all $x \in \mathbf{R}^d$, the kernel estimator of r is then

$$r_n(x) = \begin{cases} \varphi_n(x)/f_n(x) & \text{if } f_n(x) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the plug-in estimator of $\mathcal{L}(t)$ is $\mathcal{L}_n(t) = \{x \in \mathbf{R}^d : r_n(x) > t\}$. The main advantage of this estimator is the simplicity of his calculation, inherited from the plug-in approach and the use of the well-known kernel estimate. Moreover, our estimator does not require strong assumptions on the shape of level sets. Despites the regularity assumptions for the regression function inherited from the kernel estimator, our consistency results are obtained for general shapes of level sets.

Theorem 6. *Under regularity assumptions on r, and if $h_n \to 0$ and $nh_n^d / \log n \to \infty$, then*

$$\mathbb{E}\,\lambda\left(\mathcal{L}_n(t) \Delta \mathcal{L}(t)\right) \underset{n\to\infty}{\to} 0,$$

where λ is the Lebesgue measure and $A\Delta B$ the symmetrical difference between sets A and B.

Furthermore, we establish a rate of convergence for our estimator. From now on, $\Theta \subset (0, \sup_{\mathbf{R}^d} r)$ is an open interval.

Theorem 7. *Under regularity assumptions on r, if $nh_n^d/(\log n) \to \infty$ and $nh_n^{d+4} \log n \to 0$, then for almost all $t \in \Theta$*

$$\mathbb{E}\,\lambda\left(\mathcal{L}_n(t) \Delta \mathcal{L}(t)\right) = O(1/\sqrt{nh_n^d}).$$

Finally, a simple simulation study is provided. It confirms the theoretical rate obtained in Theorem 7. The choice of the bandwidth is also investigated in this simulation study. We compare a choice made by the R package "np' to a naive cross-validation approach. The first method seems to be more efficient in this simulation study.

This work has been accepted for publication in the *Journal of the Korean Statistical Society* (DOI: 10.1016/j.jkss.2012.10.001).

Integrals of Random Functions over Level Sets of Gaussian Random Fields

CS20A R. Fields & Geom.

ALEXEY SHASHKIN*,†

*Moscow State University, Russia
†email: ashashkin@hotmail.com 420:Shashkin.tex,session:CS20A

Geometrical functionals of Gaussian random fields have been extensively studied in recent decades. In particular large attention was drawn to excursion and level sets of continuous random fields, due to their applications in tomography and astrophysics. A natural geometrical characteristics of a level set is its Hausdorff measure, or the surface area of the corresponding excursion set. If the realizations of random fields are sufficiently smooth, then the level set measure is a.s. finite and an analogue of Rice formula for its expectation holds. Expressions for higher moments and central limit theorems for this measure were given by Malevich, Adler, Wschebor, Kratz etc. In a series of recent papers we studied the asymptotic behaviour of random processes obtained by considering all possible real levels and their Hausdorff measures. A group of functional central limit theorems was established. However, a substantial drawback there was the fact that the continuity properties of the processes under study were not provided. This note is aimed to closing this gap, as well as generalization from measures to integrals over these measures.

Let $d \geq 3$ and let $X = \{X(t), t \in \mathbb{R}^d\}$ be a centered isotropic Gaussian random field with C^1 realizations. Assume that its covariance function R_X is integrable over $\mathbb{R}^d$ together with its partial derivatives of order 2. Denote by $\mathcal{H}_{d-1}$ the $(d-1)$-dimensional Hausdorff measure. We require two conditions to hold:

$$\mathsf{P}(X(s) = u, \nabla X(s) = 0 \text{ for some } s \in \mathbb{R}^d) = 0 \text{ for any } u \in \mathbb{R}, \tag{1}$$

$$\mathsf{P}(\mathcal{H}_{d-1}(\{s \in \mathbb{R}^d : \nabla X(s) = 0\}) > 0) = 0. \tag{2}$$

For a bounded Borel set $A \subset \mathbb{R}^d$ let $N_X(A, u) := \mathcal{H}_{d-1}(\{s \in A : X(s) = u\})$. By the first assumption, for each $u \in \mathbb{R}$ this defines a σ-finite random measure $N_X(\cdot, u)$ on Borel subsets of $\mathbb{R}^d$.

Theorem 8. *Let A be a rectangle in $\mathbb{R}^d$. Then, with probability 1, for any continuous function $f : \mathbb{R}^d \to \mathbb{R}$ the map $u \mapsto \int_A f(s) N_X(ds, u)$ is well-defined and continuous on $\mathbb{R}$.*

For $\Delta > 0$ set $T(\Delta) = \{j/\Delta \in \mathbb{R}^d : j \in \mathbb{Z}^d\} \subset \mathbb{R}^d$. Recall that a square-integrable random field $\xi = \{\xi(t), t \in \mathbb{R}^d\}$ is called (BL, θ)-dependent if there exist a non-increasing $\theta_\xi : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\theta_\xi(r) \to 0$ as $r \to \infty$ and for all large enough $\Delta > 0$, any disjoint finite $I, J \subset T(\Delta)$ and any pair of Lipschitz $f : \mathbb{R}^{|I|} \to \mathbb{R}$, $g : \mathbb{R}^{|J|} \to \mathbb{R}$, the inequality

$$|cov(f(\xi_I), g(\xi_J))| \leq Lip(f) Lip(g) (|I| \wedge |J|) \Delta^d \theta_\xi(r)$$

holds. Here $|M|$ is the cardinality of a set M, the notation $\xi_I = (\xi_i, i \in I)$ is employed, r is the distance between I and J, and the Lipschitz constants are with respect to the norm $\|z\|_1 = \sum_i |z_i|$.

Let $Y = \{Y(s), s \in \mathbb{R}^d\}$ be a centered stationary (BL, θ)-dependent continuous random field with integrable covariance function. Consider normalized integrals

$$Z_n(u) := n^{-d/2} \int_{[0,n]^d} Y(s) N_X(ds, u), \; n \in \mathbb{N}.$$

Theorem 9. *Random processes $\{Z_n(\cdot), n \in \mathbb{N}\}$ converge in distribution in $C(\mathbb{R})$, as $n \to \infty$, to a centered Gaussian random process Z with covariance function determined by X and Y.*

Acknowledgment. This research was partially supported by RFBR, grant No. 13-01-00612.

OCS15 Ecol. and Biomed. Data

Estimation of Shannon's Index When Samples are Taken without Replacement

TSUNG-JEN SHEN[*,†]

[*]Institute of Statistics, National Chung Hsing University, Taichung, Taiwan
[†]email: tjshen@nchu.edu.tw 421:Tsung-JenShen.tex,session:OCS15

Shannon's (entropy) index is widely employed in practical applications for the ecological monitoring and management. Most estimators of Shannon's index in the literature have been derived from models considering sampling with replacement or in an infinite population. However, such sampling devices may not be suitable for sedentary species where data are usually sampled without replacement from an assemblage. In the present work, we develop a promising estimator of Shannon's index based on data sampled without replacement being more efficient than with replacement in a finite assemblage. Furthermore, being tested with a simulation study, the proposed estimator is superior to the traditional one (derived from the maximum likelihood estimation) in terms of bias and RMSE.

Acknowledgment. This work was supported by the National Science Council of Taiwan with grant number NSC 101-2118-M-005-001.

CS38A Appl. Multivariate Tech.

A Distribution-Free Multivariate Control Chart for Phase I Analysis

CHING-REN CHENG[*], JYH-JEN HORNG SHIAU[*,†]

[*]National Chiao Tung University, Hsinchu, Taiwan
[†]email: jyhjen@stat.nctu.edu.tw 422:Jyh-JenShiau.tex,session:CS38A

In actual applications of statistical process control, the distributions of the underlying processes are often unavailable especially when process data are multivariate. The aim of this talk is to present a novel distribution-free control chart for monitoring the location parameter of a multivariate process in Phase I analysis. To be robust to the distribution of the data, the spatial sign statistic that defines the direction of a multivariate observation is considered. Based on the powerful spatial sign test, a Shewhart-type control chart is developed to determine which of the observations in the historical data set are out of control. A simulation study shows that our proposed method is more powerful in detecting out-of-control cases and quite robust in terms of the type-I error rate, when compared with the traditional Hotelling's T^2 control chart and other robust versions of the T^2 chart. The applicability of the proposed control chart is demonstrated with a set of real data.

OCS25 Long-mem. Time Ser.

Hypothesis Testing under Unknown Order of Fractional Integration

BENT JESPER CHRISTENSEN[*], ROBINSON KRUSE[*,†], PHILIPP SIBBERTSEN[†,‡]

[*]CREATES, Aarhus University, School of Economics and Management,
[†]Leibniz University Hannover, School of Economics and Management, Institute for Statistics
[‡]email: sibbertsen@statistik.uni-hannover.de
423:PHILIPP_SIBBERTSEN.tex,session:OCS25

This paper considers the question of testing for a regression parameter in a linear regression framework when the order of possible fractional integration of the error term is unknown. Specifically we consider the linear regression model

$$y_t = \beta' z_t + e_t,$$

where z_t are deterministic or exogenous stochastic regressors fulfilling some regularity conditions specified later, β is a vector of regression parameters, the prime denoting the transposed of the vector and e_t is an error term which is possibly fractionally integrated but the order of integration is unknown. We are concerned with testing hypotheses for β. It is well known that the behavior of standard tests for β such as the Wald test depend strongly on the order of integration of the error term. Neglecting the fractional behavior of the errors for instance can lead to spurious rejections of the null. On the other hand standard estimators for the order of integration are biased in the linear regression set-up.

For the I(0) - I(1) framework Vogelsang (1998a) suggested a testing approach which allows valid inference when the order of integration of the error term is unknown. The idea is originally embedded in a trend testing framework but was generalized to a test for mean shifts in Vogelsang (1998b). He suggests to correct the test statistic by a factor so that the critical values of the new test statistic are the same independent of whether the error term is integrated or not. Harvey et al. (2009) suggest an alternative approach in the context of testing for trend breaks. However, as will be shown in the following section these approaches do not hold when the order of integration is allowed to be fractional. The problem is that the Vogelsang approach is based on having a choice of two possible orders of integration of which one is the true order. In the fractional case we have a continuum of possible orders of integration which makes the suggested correction term of Vogelsang's approach being invalid. Vogelsang suggests a correction of the test statistic which depends on a correction parameter b. In the case of fractional integration this parameter depends on the unknown order of fractional integration which makes the approach infeasible in this situation.

To overcome this problem we suggest an LM-type test for hypotheses on the regression coefficient β. The framework of our approach is motivated by Robinson (1994). In a similar regression framework Robinson (1994) suggests an LM-type test for the order of integration. We adopt his approach to a test on the regression parameter when the order of integration is unknown under fairly mild regularity assumptions guaranteeing flexibility of our method.

References

[Harvey, D.I., Leybourne, S.J. and Taylor, A.M.R. (2009)] "Simple, robust and powerful tests of the breaking trend hypothesis." *Econometric Theory*, 25, 995 – 1029.

[Robinson, P. M. (1994)] "Efficient Tests of Nonstationary Hypotheses." *Journal of the American Statistical Association*, 89, 1420 – 1437.

[Vogelsang, T.J. (1998a)] "Trend function hypothesis testing in the presence of serial correlation." *Econometrica*, 66, 123 – 148.

Optimal Designs for Prediction of Shifted Ornstein-Uhlenbeck Sheets

CS20A R. Fields & Geom.

SÁNDOR BARAN*, KINGA SIKOLYA†,§, MILAN STEHLÍK‡

*University of Debrecen, Debrecen, Hungary,
†University of Heidelberg, Heidelberg, Germany,
‡Johannes Kepler University in Linz, Linz, Austria
§email: sikolya.kinga@inf.unideb.hu

424:SikolyaKinga.tex,session:CS20A

Computer simulations are often used to replace physical experiments to explore the complex relationships between input and output variables. We study the optimal design problem for prediction of a stationary Ornstein-Uhlenbeck sheet on a monotonic set with respect to integrated mean square

prediction error and entropy criterion. We show that there is a substantial difference between the shapes of optimal designs for Ornstein-Uhlenbeck processes [Baldi Antognini and Zagoraiou, 2010] and sheets. In particular, we show that optimal prediction based on integrated mean square prediction error not necessary leads to space-filling designs. Moreover, we present some numerical experiments to illustrate selected cases of optimal designs for small number of sampling locations.

Acknowledgment. This research has been supported by the Hungarian –Austrian intergovernmental S&T cooperation program TÉT_10-1-2011-0712 and partially supported the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[Baldi Antognini and Zagoraiou, 2010] Baldi Antognini, A., Zagoraiou, M., 2010. Exact optimal designs for computer experiments via Kriging metamodelling. *J. Statist. Plann. Inference* **140**, 2607–2617.

POSTER
Poster

Nonlinear Conditional U-Statistics

VALERIY SIMAKHIN*,†

*Kurgan State University, Kurgan, The Russian Federation

†email: sva_full@mail.ru 425:ValeriySimakhin_abstrac.tex,session:POSTER

Let $\{(x_i, y_i), i = \overline{1,N}\}$ be a sample i.i.d. from a $d = (s \times l)$ - dimensional random variable (X, Y) with continuous distribution function $F(x, y)$ and conditional distribution function $F(y/x)$, $H(\vec{t}, \vec{x}) = \prod_{j=1}^{m} F(t_j, x_j)$. We now consider the parameter θ in the form of a nonlinear conditional functional

$$\theta = \int \phi(\vec{t}, \partial^{\alpha} \vec{H}(\vec{t}/\vec{x}))dH(\vec{t}/\vec{x}), \tag{1}$$

where $\partial^{\alpha} \vec{H}(\vec{t}/\vec{x}) = \frac{\partial^{\alpha}}{\partial t_1^{\alpha_1} \ldots \partial t_m^{\alpha_m}}$, $\alpha_1 + \ldots + \alpha_m = \alpha$, and ϕ - is a continuous function.
Taking advantage of the substitution method, we obtain the estimate θ_N in the form

$$\theta_N = \int \phi(\vec{t}, \partial^{\alpha} \vec{H}_N(\vec{t}/\vec{x}))dH_N(\vec{t}/\vec{x}), \tag{2}$$

where $H_N(\vec{t}/\vec{x})$ and $\partial^{\alpha} \vec{H}_N(\vec{t}/\vec{x})$ are empirical distribution functions of the m th order from the class of discrete and continuous estimates, respectively [Stute (1991)],[Simahkin (2011, LAP)]. For $\phi(\bullet) = \phi(\vec{t})$, estimates (2) are the conditional U-statistics introduced in [Stute (1991)] as generalizations of the nonparametric Nadaraya-Watson estimates.

When a number of regularity conditions are fulfilled, the consistency and the asymptotic normality of the estimates are proved.

Examples.

1. Let $F(y/x) = 1 - F(2\theta - y/x)$ be symmetric about θ. Estimate (2) of the functional $\int \phi(1 - F_N(2\theta_N - t/x))dF_N(t/x) = 0.5$ at $\phi(\bullet) = t$ leads to the conditional Hodges-Lehmann estimate (the conditional median of the Walsh half-sums) of the form

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} Sign\left(\theta_N - \frac{y_i - y_j}{2}\right) W(x - x_i) W(x - x_j) = 0,$$

where $W(x - t)$ are the normalized kernel functions.
In some cases (robustness), the given estimate is preferable for nonparametric regression problems compared to the classical Nadaraya-Watson estimate.

2. The robust nonparametric estimates based on the weighed maximum likelihood method for nonparametric problems of regression and forecasting [Simahkin (2011, LAP)].

3. Conditional measures of the dependence [Simakhin (2011, Vestnik TymGU)] generate a wide class of nonlinear functionals (1).

References

[Stute (1991)] Stute W. Coditional U-statistics, *The Annals of Probability*, V. 19, No. 2, 1991, 812-825 pp.

[Simahkin (2011, LAP)] Simakhin V. A. Robust Nonparametric Estimators.– Germany: *LAP*, 2011. –292 p.

[Simakhin (2011, Vestnik TymGU)] Simakhin V. A. Conditional measures of a dependence, *Vestnik TyumGU*, No. 7, Series "Physical and Mathematical Sciences", 2011, 119-122 pp.

Semi-parametric Bayesian Partially Identified Models based on Support Function

CS8A Bayesian Semipar.

YUAN LIAO*, ANNA SIMONI†,‡

*Department of Mathematics, University of Maryland at College Park, College Park, MD 20742, USA

†CNRS and THEMA, Université de Cergy-Pontoise, 95011 Cergy-Pontoise, France

‡email: simoni.anna@gmail.com 426:AnnaSimoni.tex,session:CS8A

Bayesian partially identified models have received a growing attention in recent years in the econometric literature, due to their broad applications in empirical studies. The Bayesian approach in this literature has been assuming a parametric model, by specifying an ad-hoc parametric likelihood function. However, econometric models usually only identify a set of moment inequalities, and therefore using an incorrect likelihood function may result in inconsistent estimations of the identified set. On the other hand, moment-condition based likelihoods such as the limited information and exponential tilted empirical likelihood, though guarantee the consistency, lack of probabilistic interpretations. We propose a semi-parametric Bayesian partially identified model, by placing a nonparametric prior on the unknown likelihood function. Our approach thus only requires a set of moment conditions but still possesses a pure Bayesian interpretation. We study the posterior of the support function, which is essential when the object of interest is the identified set. The support function also enables us to construct two-sided Bayesian credible sets (BCS) for the identified set. It is found that, while the BCS of the partially identified parameter is too narrow from the frequentist point of view, that of the identified set has asymptotically correct coverage probability in the frequentist sense. We also develop the posterior concentration theory for the support function, and prove the semi-parametric Bernstein von Mises theorem. Moreover, we introduce an optimal estimation of the identified set based on the Bayesian decision theory. Finally, the proposed method is applied to analyzing a financial asset pricing problem.

Some Limiting Theorems for Signed Measures

Closing Closing Lecture

YAKOV G. SINAI*,†

*Princeton University, Princeton, NJ, USA

†email: sinai@math.princeton.edu 427:YakovSinai.tex,session:Closing

Number theory provides many interesting probabilistic problems which sometimes are connected with signed measures and lead to new limiting distributions and new asymptotical behavior.

Many examples appear in the theory of Moebius function which is one of the central objects in number theory. In the simplest case, statistical properties of the Moebius function are described with the help of the so-called Dickmann–De-Brujin distribution; in other cases the limiting distributions have complicated singularities.

The whole set of related problems will be the main content of the lecture. The basic results were obtained by M. Avdeeva, F. Cellarosi, Dong Li, and the present author.

IS26 Risk Anal.

A New Measure of Concentration: Its Role in Characterizing Distributions

NOZER D. SINGPURWALLA[*,†]

[*]Department of Systems Engineering and Engineering Management, Department of Management Science, City University of Hong Kong, Hong Kong
[†]email: nsingpur@cityu.edu.hk 428:NozerSingpurwalla.tex,session:IS26

Shaked and Shantifumar introduced a measure of income inequality which they called "excess wealth". A detailed investigation of this measure revealed that despite its name, the measure would be of little value to economists, econometricians, and social scientists. However, the measure motivated me to develop a new measure of income inequality which encapsulates excess wealth. An attractive by product of this measure is that it can be used to characterize probability distributions by their stochastic properties of ageing . The work should have appeal to applied probabilists, economists, and engineers working in reliability theory.

CS14A Stat. Neuronal Data

A New Estimator for Mutual Information

MARIA TERESA GIRAUDO[*], LAURA SACERDOTE[*], ROBERTA SIROVICH[*,†]

[*]Department of Mathematics “G. Peano", University of Torino, Italy
[†]email: roberta.sirovich@unito.it 429:RobertaSirovich.tex,session:CS14A

Mutual Information is a measure of multivariate association in a random vector. It has many attractive properties, in particular it is sensitive to non linear dependencies.

From a statistical point of view the direct estimation of mutual information can be difficult, especially if the marginal spaces have dimensions larger than one.

We propose here a new and competitive estimator for the mutual information in its general multidimensional definition. We deduce an equation that links the mutual information of a random vector to the entropy of the so called linkage function. The problem is hence reduced to the estimation of the entropy of a suitably transformed sample. The properties of the new estimator are illustrated through simulated examples and performances are compared to the best estimators in the literature.

CS35A Discretre Response M.

Testing Goodness of Fit for the Discrete Stable Family

LENKA SLÁMOVÁ[*,†], LEV KLEBANOV[*]

[*]Charles University in Prague, Czech Republic
[†]email: slamova.lenka@gmail.com 430:LenkaSlamova.tex,session:CS35A

Stable distributions play an important role both in the theory and applications. A lot of phenomenas are modeled by continuous stable distributions, even when the character of the data suggests a

discrete approach. Discrete stable distributions form discrete analogues to stable distributions, with ability to describe heavy tails and skewness of integer valued data.

In this talk we will speak about statistical methods to asses the goodness of fit of discrete stable distributions. We have to deal with several problems which inhibit the use of well known methods. The classic parametric χ^2 test assumes closed form of a probability function that is not available for the discrete stable family. The nonparametric Kolmogorov-Smirnov or Wilcoxon tests on the other hand assume continuity of the distribution and even thought methods that overcome this problem have been proposed in the literature, the nonexistence of closed form probability function remains a problem. We discuss methods proposed in the literature in the past and suggest a new nonparametric method based on a characterization of the discrete stable law.

Direct Semiparametric Estimation of Fixed Effects Panel Data Varying Coefficient Models

CS32A Nonparame

JUAN MANUEL RODRIGUEZ-POO*, ALEXANDRA SOBERON*,†

*University of Cantabria, Santander, Spain
†email: alexandra.soberon@unican.es 431:AlexandraSoberon.tex,session:CS32A

In this paper we present a new technique to estimate varying coefficient models of unknown form in a panel data framework where individual effects are arbitrarily correlated with the explanatory variables in a unknown way. The resulting estimator is robust to misspecification in the functional form of the varying parameters and it is shown to be consistent and asymptotically normal. Furthermore, introducing a transformation, it achieves the optimal rate of convergence for this type of problems and it exhibits the so called oracle efficiency property. Since the estimation procedure depends on the choice of a bandwidth matrix, we also provide a method to compute this matrix empirically. Monte Carlo results indicates good performance of the estimator in finite samples.

Acknowledgment. The authors acknowledge financial support from the Programa Nacional de Formación de Profesorado Universitario/ Spanish Ministery of Education. Ref. AP-2007-02209. We would like to thank also Wenceslao Gonzalez-Manteiga, Juan A. Cuesta-Albertos and Stefan Sperlich for their very helpful comments and suggestions. Of course, all errors are ours.

Nonparametric Estimation of a Conditional Covariance Matrix for Dimension Reduction

CS5A H-D Dim. Reduction

JEAN-MICHEL LOUBES*, CLEMENT MARTEAU*, MAIKOL SOLÍS*,†

*Institut de Mathématiques de Toulouse, France.
†email: msolisch@math.univ-toulouse.fr 432:Maikol_SOLIS.tex,session:CS5A

Nowadays in many fields such as biology, chemistry, economics, for instance; handling a huge quantity of data with many variables interacting mutually has become essential. Hence, reducing the complexity of high dimensional data have become a stringent issue, even at a computational level.

Assume for $\mathbf{X} \in \mathbb{R}^p$, $Y \in \mathbb{R}$, ψ an unknown regression function and $\varepsilon \in \mathbb{R}$ (usually some noise), we have the model

$$Y = \psi(X) + \varepsilon,$$

where p is much larger than the data available. Then, it is well-known that the *curse of dimensionality* problem arises. To cope with this issue, the model called Sliced Inverse Reduction (SIR) was proposed by [1] to reduce the dimensionality of this problem.

This technique transforms the original problem into,

$$Y = \varphi\left(\beta_1^\top \boldsymbol{X}, \ldots, \beta_k^\top \boldsymbol{X}, \varepsilon\right)$$

where k is much less than p, denoted as $k \ll p$, $\boldsymbol{X}$ is a p-dimensional random vector, the β's are unknown fixed vectors, ε is independent of $\boldsymbol{X}$ and $\varphi : \mathbb{R}^{k+1} \to \mathbb{R}$ is an arbitrary real valued function. This model allows, via projection, the extraction of all the Y's relevant information by only a k-dimensional subspace generated by the β's.

The main objective of the SIR method is the estimation of the largest eigenvalues with its corresponding eigenvectors (say the β's) of the unknown matrix

$$\Sigma = \mathrm{Cov}\left(\mathbb{E}\left[\boldsymbol{X}|Y\right]\right).$$

The directions generated by the β's are called Effective Dimension Reduction (EDR) directions and consequently the first k eigenvectors span the EDR space. In the literature, the algorithms to estimate the desired matrix do not give any explicit result about its rate of convergence.

This communication aims to estimate the conditional covariance Σ. To that purpose, we will use methods from covariance estimation (e.g. [3]) and plug a preliminary nonparametric estimate of the conditional density using a kernel estimator. We have proved two different behaviors for the estimate. On the one hand, if the conditional density is regular enough, we recover the parametric rate of convergence as if the density were known. On the other hand, the preliminary nonparametric estimate plays a role in the estimation of the conditional covariance, slowing its convergence rate.

References

[1] Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86(414)**, 316–327.

[2] Zhu, L. X. & Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression, *The Annals of Statistics*, **24(3)**, 1053–1068.

[3] Cai, T. T., Zhang, C.-H., & Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, **38(4)**, 2118–2144. DOI: 10.1214/09-AOS752

CS11A SDE-s

Approximation of the Solutions of Stochastic Differential Equations Driven by Multifractional Brownian Motion

ANNA SOÓS*,†

*Babes Bolyai University, Cluj Napoca, Romania

†email: asoos@math.ubbcluj.ro 433:AnnaSoos.tex,session:CS11A

The aim of this paper is to approximate the solution of a stochastic differential equation driven by multifractional Brownian motion using a series expansion for the noise. We use the M. Zähle's fractional calculus approach for Ito stochastic integral with respect the multifractional Brownian motion.

OCS1 Longitudin Models

Longitudinal Models with Outcome Dependent Follow-up Times

INÊS SOUSA*,†, LISANDRA ROCHA*

*University of Minho, Portugal

†email: isousa@math.uminho.pt 434:InesSousa.tex,session:OCS1

In longitudinal studies individuals are measured repeatedly over a period of time. Occasionally, individuals have different number of measurements assessed at different times. Usually, a patient is

measured according to their clinical condition. For example, in case of renal insufficiency, the follow-up times are decided according to the previous observed value of creatinine. High levels of creatinine in the blood, warn for a possible kidney failure. So, if the patient on the previous measurement had a high value of creatinine in the blood, the patient will be measured sooner than other patient with an expected creatinine value. Thus, in this situations, follow-up time process is considered dependent of the longitudinal outcome process and it should not be considered deterministic in the study design. In standard longitudinal models (Diggle et al. 2002), the follow up time process is assumed deterministic, meaning, the follow-up time process is noninformative about the outcome longitudinal process of interest. Therefore, this type of analysis does not considered the dependence that can exist between the follow-up time process and the longitudinal outcome process.

In this work we propose to joint model the longitudinal process and the follow-up time process. In 2002, Lipsitz et al. propose a model to study longitudinal data where the follow-up time process is conditioned by the longitudinal outcome process. They develop a likelihood-based procedure for estimating the regression parameters in models for continuous responses measured at irregular points in time. They assume that any follow-up time measurement depends only on the previous longitudinal observed outcome, not the time at which they were observed. So, under this assumption, the authors argue that the follow up time process can be ignored. We conducted a simulation study of longitudinal data and we estimate the model parameters taking into account the likelihood function proposed by Lipsitz et al. (2002).

Acknowledgment. The author acknowledges partial financial support from the project PTDC/MAT/112338/2009 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education.

References

[Diggle et al. (2002)] Diggle, P.J., Liang, K-Y., Zeger, S.L. 2002: Analysis of Longitudinal Data, Oxford: Clarendon Press.

[Lipsitz et al. (2002)] Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J.G., Gelber, R., Lipshultz, S. 2002: Parameter estimation in longitudinal studies with outcome-dependent follow-up, *Biometrics*, **58**, 50 - 59.

Data Disclosure: Sufficient Truth but Not the Whole Truth

IS26 Risk Anal.

KURT A. PFLUGHOEFT*, EHSAN S. SOOFI†,‡, REFIK SOYER§

*Maritz Research, St. Louis, MO, USA,
†University of Wisconsin-Milwaukee, Milwaukee, WI, USA,
‡George Washington University, Washington, DC, USA
§email: esoofi@uwm.edu

435:Soofi.tex,session:IS26

Preserving confidentiality of individuals in data disclosure is of a prime concern for public and business organizations. We propose an information-theoretic architecture for data disclosure problem in order to provide useful information to legitimate users for statistical analysis in a way that prevents the misuse by the intruders who seek information on individuals. This architecture consists of developing a maximum entropy (ME) model based on essential features of the actual data, checking the adequacy of the ME model for the data, and simulating disclosure data from the ME model, and minimizing the discrepancy between the essential features of actual data and the disclosure data. The framework can be used for univariate and multivariate data disclosure. Applications to a mortgage default data set and a bankŠs customersŠ data set will be illustrated.

IS26 Risk Anal.

Importance of Components for a System

NADER EBRAHIMI*, EHSAN S. SOOFI†, REFIK SOYER‡,§

*Northern Illinois University, DeKalb, IL, USA,
†University of Wisconsin-Milwaukee, Milwaukee, WI, USA,
‡George Washington University, Washington, DC, USA
§email: soyer@gwu.edu 436:Soyer.tex,session:IS26

Which component is most important for a system's survival? We answer this question by ranking the information relationship between a system and its components. The mutual information (M) measures dependence between the operational states of the system and a component for a mission time as well as between their life lengths. This measure ranks each component in terms of its expected utility for predicting the system's survival. We explore some relationships between the ordering of importance of components by M and by Zellner's Maximal Data Information (MDIP) criterion. For many systems the bivariate distribution of the component and system lifetimes does not have a density with respect to the two-dimensional Lebesgue measure. For these systems, M is not defined, so we use a modification of a mutual information index to cover such situations. Our results for ordering dependence are general in terms of binary structures, convolution of continuous variables, and order statistics.

SIL Spec. Invited Lecture

Removing Unwanted Variation: from Principal Components to Random Effects

TERRY SPEED*,†,‡, JOHANN GAGNON-BARTSCH*, LAURENT JACOB*

*Department of Statistics, University of California at Berkeley, USA, †Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
‡email: terry@stat.berkeley.edu 437:TerrySpeed.tex,session:SIL

Ordinary least-squares is a venerable tool for the analysis of scientific data originating in the work of A-M. Legendre and C. F. Gauss around 1800. Gauss used the method extensively in astronomy and geodesy. Generalized least squares is more recent, originating with A. C. Aitken in 1934, though weighted least squares was widely used long before that. At around the same time (1933) H. Hotelling introduced principal components analysis to psychology. Its modern form is the singular value decomposition. In 1907, motivated by social science, G. U. Yule presented a new notation and derived some identities for linear regression and correlation. Random effects models date back to astronomical work in the mid-19th century, but it was through the work of C. R. Henderson and others in animal science in the 1950s that their connexion with generalized least squares was firmly made.

These are the diverse origins of our story, which concerns the removal of unwanted variation in high-dimensional genomic and other "omic" data using negative controls. We start with a linear model that Gauss would recognize, with ordinary least squares in mind, but we add unobserved terms to deal with unwanted variation. A singular value decomposition, one of Yule's identities, and negative control measurements (here genes) permit the identification of our model. In a surprising twist, our initial solution turns out to be equivalent to a form of generalized least squares. This is the starting point for much of our recent work. In this talk I will try to explain how a rather eclectic mix of familiar statistical ideas can combine with equally familiar notions from biology (negative and positive controls) to give a useful new set of tools for omic data analysis. Other statisticians have come close to the same endpoint from a different perspectives, including Bayesian, sparse linear and random effects models.

Asymptotic Geometry of Excursion Sets of Non–Stationary Gaussian Random Fields

IS16 R. Fields, Geom.

EVGENY SPODAREV[*,‡], DMITRY ZAPOROZHETS[†]

[*]Institute of Stochastics, Ulm University, Germany,
[†]St. Petersburg Branch of Steklov Mathematical Institute, Russia
[‡]email: evgeny.spodarev@uni-ulm.de 438:Spodarev.tex,session:IS16

In this talk, we present some recent asymptotic results on the geometry of excursion sets of Gaussian random fields. Namely, we consider the asymptotics of the mean volume, surface area and the Euler characteristic of excursion sets of non–stationary sufficiently smooth Gaussian random fields with a unique point of maximum variance (lying on the boundary of the observation window) as the excursion level tends to infinity.

On Extremal Behaviour of Random Variables Observed in Renewal Times

CS19B Lim. Thms. Point Proc.

BOJAN BASRAK[*], DRAGO ŠPOLJARIĆ[†,‡]

[*]Department of Mathematics, University of Zagreb, Croatia,
[†]Faculty of Mining, Geology and Petroleum Engineering, University of Zagreb, Croatia
[‡]email: drago.spoljaric@rgn.hr 439:Spoljaric_Drago.tex,session:CS19B

The main aim is to understand the asymptotic behaviour of all upper order statistics of iid observations $X_1, X_2, \ldots$ until a random time $\tau(t)$ which is determined by a renewal process. This kind of problem appears in the analysis of general shock models (see [1] for instance) where X_i denotes the magnitude of the i-th shock and renewal process $(\tau(t))$ counts the number of shocks.

In general, the limiting behaviour of the distribution of the maximum of random variables observed in renewal times has been studied since 1960's. The basic problem was to understand the distribution of

$$M(t) = \max_{i \le \tau(t)} X_i$$

where $\tau(t)$ is a renewal process. When interarrival times of the renewal process have finite mean, the solution was given by Berman in [2]. Anderson in [1] was the first to study the limiting behaviour of $M(t)$ in the case of renewal process with infinite mean interarrival times. More recently, his result has been extended to describe the limiting behaviour of $(M(t))$ on the level of processes (see [3, 4] for instance).

We study asymptotic behaviour of all upper order statistics in the sequence X_i until time $\tau(t)$, rather than only the maximum. Our approach to the problem is different too, it mainly relies on theory of point processes and allows recovery of previously established results. We also permit certain types of dependence between observations and interarrival times, relaxing the conditions previously used in the literature. Finally, we show how our approach yields some well known and apparently new results concerning, for instance, the longest run of heads and the maximal sojourn time of a continuous time random walk.

References

[1] K. K. Anderson. Limit theorems for general shock models with infinite mean intershock times. J. Appl. Probab., 24(2):449-456, 1987.

[2] S. M. Berman. Limiting distribution of the maximum term in sequences of dependent random variables. Ann. Math. Statist., 33:894-908, 1962.

[3] M. M. Meerschaert and S. A. Stoev. Extremal limit theorems for observations separated by random power law waiting times. J. Statist. Plann. Inference, 139(7):2175-2188, 2009.

[4] D. S. Silvestrov and J. L. Teugels. Limit theorems for extremes with random sample size. Adv. in Appl. Probab., 30(3):777–806, 1998.

IS19 Shape & Image

Joint Registration and Shape Analysis of Functions, Curves and Surfaces

ANUJ SRIVASTAVA*,†

*Department of Statistics, Florida State University, Tallahassee, FL, USA

†email: anuj@stat.fsu.edu 440:AnujSrivastava.tex,session:IS19

I will present some recent developments in the area of functional data analysis, and shape analysis of curves and surfaces, with a focus on the registration problem. The registration is an important ingredient in comparing data objects – functions [4], curves [2, 3], and surfaces [1] – as it determines point-to-point correspondences across such objects, and governs the performances of ensuing statistical analyses – metrics for shape comparisons, statistical summaries, and even generative models for shapes.

When studying parameterized objects, the registration corresponds to matching of points that have the same parameter value. Consider a pair of arbitrarily parameterized curves $\beta_1, \beta_2 : [0,1] \to \mathbb{R}^n$. For any $t \in [0,1]$, the points $\beta_1(t)$ and $\beta_2(t)$ are said to be registered. Now, let $\gamma : [0,1] \to [0,1]$ be a boundary-preserving diffeomorphism; it is also called a re-parameterization function. If we re-parameterize β_2 using γ, i.e. we form $\beta_2(\gamma(t))$, then the point $\beta_1(t)$ is now registered to $\beta_2(\gamma(t))$. Thus, the re-parameterization function γ is used to control the registration between curves. The same idea applies to registration of functions, images, surfaces, and other objects.

I will present a Riemannian framework that performs both registration and statistical analysis of data objects. This framework is based on an extensions of Fisher-Rao Riemannian metric (non-parametric) form to arbitrary functions, curves in Euclidean spaces, and even 2D surface spaces. The single most important property of these Riemannian metrics is that they are invariant under identical re-parameterization of input curves. This allows us to fix parameterization of one curve and search over all re-parametrizations of the second curve reach optimal registration. The energy being minimized is also a proper distance on the shape space of objects, thus providing a tool for shape analysis. This distance is then used to generate shape summaries, shape models, and develop hypothesis tests for classification of shapes.

I will demonstrate these ideas using examples from growth curves, bioinformatics, medical image analysis, and computer vision.

References

[1] S. Kurtek, E. Klassen, J.C. Gore, Z. Ding, and A. Srivastava. Elastic geodesic paths in shape space of parametrized surfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1717–1730, 2012.

[2] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1415 –1428, 2011.

[3] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding. Statistical Modeling of Curves Using Shapes and Related Features. *Journal of American Statistical Association*, 107(499):1152 –1165, 2012.

[4] J. D. Tucker, W. Wu, and A. Srivastava. Generative Models for Functional Data Using Phase and Amplitude Separation. *Computational Statistics and Data Analysis*, 61:50 –66, 2013.

Strong Limit Theorems for Increments of Random Fields with Independent Components

OCS23 Strong Limit Thm.

ALLAN GUT*, ULRICH STADTMÜLLER†,‡

*University of Uppsala, Uppsala, Sweden,
†University of Ulm, Ulm, Germany
‡email: ulrich.stadtmueller@uni-ulm.de 441:stadtmueller.tex,session:OCS23

We will consider a $d-$dimensional random field on $\mathbb{Z}_+^d$ with iid components (X_n), $n \in \mathbb{Z}_+^d$ and partial sums (S_n). We study the increments of these partial sums. The aim of this talk is to give various strong laws and laws of iterated or single logarithms under exact moment conditions for these increments.

Statistical Inference for Financial Volatility Under Nonlinearity and Nonstationarity

OCS26 Resampling Nonstat T.S.

BARTOSZ STAWIARSKI*,†

*Institute of Mathematics, Cracow University of Technology, Cracow, Poland
†email: bstawiarski@pk.edu.pl 442:BartoszStawiarski.tex,session:OCS26

The presentation focuses upon statistical inference for financial volatility. The volatility of tradable, financial instruments plays a fundamental role not only in their pricing and associated risk management. Its behaviour can also convey a vital information concerning market conditions and economic environment as a whole, hence growing interest in econometric modelling and statistical inference for volatility. Obviously, proper modelling is crucial for justifiable, reliable statistical inference. Financial volatility, treated as a nonlinear and nonstationary process, exhibits such stylized phenomena as: clustering, spikes and mean-reversion. Moreover, recurrent structural breaks within financial markets dynamics seen since about 2007 translate into abrupt volatility changes. Detecting them is particularly challenging and involves using more sophisticated tools than plain GARCH time series. Especially, following the most recent research papers, we present ICSS-GARCH detection algorithm and its alternative proposed for (notoriously occurring in practice) nongaussian cases, namely Nonparametric Change Point Model. Besides, as resampling methods for nonstationary models draw rapidly growing attention, we give some notes on bootstrapping realized volatility.

Statistical Inference when Fitting Simple Models to High-Dimensional Data

CS5A H-D Dim. Reduction

LUKAS STEINBERGER*,†, HANNES LEEB*

*University of Vienna
†email: lukas.steinberger@univie.ac.at 443:LukasSteinberger.tex,session:CS5A

We study linear subset regression in the context of the high-dimensional overall model $y = \theta' Z + u$ with univariate response y and a d-vector of random regressors Z, independent of u. Here, 'high-dimensional' means that the number n of available observations may be much less than d. We consider simple linear submodels where y is regressed on a set of p regressors given by $X = B'Z$, for some $d \times p$ matrix B with $p \leq n$. The corresponding simple model, i.e., $y = \gamma' X + v$, can be justified by imposing appropriate restrictions on the unknown parameter θ in the overall model; otherwise,

this simple model can be grossly mis-specified. In this talk, we show that the least-squares predictor obtained by fitting the simple linear model is typically close to the Bayes predictor $E[y|X]$ in a certain sense, uniformly in $\theta \in \mathbb{R}^d$, provided only that d is large. Moreover, we establish the asymptotic validity of the standard F-test on the surrogate parameter which realizes the best linear population level fit of X on y, in an appropriate sense. On a technical level, we extend recent results from [2] on conditional moments of projections from high-dimensional random vectors; see also [1].

References

[1] Hall, P. and Li, K.-C. (1993). On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics* **21**, 2, 867–889.

[2] Leeb, H. (2013). On the Conditional Distributions of Low-Dimensional Projections from High-Dimensional Data. *The Annals of Statistics,* forthcoming.

IS23 Stat. Genetics, Biol.

Quantitative Analysis of Proteomics Mass Spectrometry Data

SEBASTIAN GIBB*, KORBINIAN STRIMMER*,†

*Institute of Medical Informatics, Statistics, and Epidemiology, University of Leipzig, Germany

†email: strimmer@uni-leipzig.de 444:KorbinianStrimmer.tex,session:IS23

In proteomics, mass spectrometry analysis is increasingly becoming an important tool, for example to identify biomarkers for cancer in clinical diagnostics. As with other high-throughput technologies, due to low signal to noise ratio and large systematic variations in the data the development of sophisticated statistical algorithms is essential in the analysis of mass spectrometry data.

Here, we will discuss the statistical and algorithmic challenges involved in the analysis of proteomics mass spectrometry data. Correspondingly, we present a complete analysis pipeline for proteomics mass spectrometry data that comprises all steps from importing of raw data, preprocessing (e.g. baseline removal), peak detection, non-linear peak alignment to relative calibration of mass spectra and feature selection and classification.

The suggested framework for analyzing mass spectrometry data is implemented in R in the package MALDIquant which is freely available from http://strimmerlab.org/software/maldiquant/ and CRAN under the terms of the GNU General Public License.

OCS7 Comp. Biology

Beyond Static Networks: A Bayesian Non-Parametric Approach

MICHAEL STUMPF*,†

*Theoretical Systems Biology, Imperial College London, UK

†email: m.stumpf@imperial.ac.uk 445:Stumpf.tex,session:OCS7

In biology molecular interaction networks are used to summarize the interactions between proteins and other molecules inside a cell. Generally they are considered as static objects, but in reality they are highly dynamic and change in response to environmental and internal signals and processes. Here we develop a set of approaches that capture this dynamic behaviour. Using *hierarchical Dirichlet processes* (HDP) we establish a framework in which the network structure is modelled as a hidden state, which is inferred from measured gene expression data. This turns out to be a very flexible framework that can be used to infer graphical models flexibly and reliably. We illustrate the use of this framework in the context of gene expression time-series data, where we use a Bayesian network formalism to infer the the gene-regulatory network, using a HDP hidden Markov model to capture the change in network structure over time. We then present a related approach which deals

with changes in the network between different conditions, and use this in the context of a Gaussian graphical model to identify systematic differences in gene regulatory networks between patients and controls for the case of sporadic inclusion body myositis. In both cases it turns out that the inference problem can be phrased in a manner that is amenable to very efficient Gibbs sampling procedures. In our applications we find that the resulting temporally-varying or condition-specific networks offer more detailed insights than approaches which impose a static structure onto the network.

A Study of Generalized Normal Distributions

CS39A Distributio Theory

NAN-CHENG SU*,‡, WEN-JANG HUANG†

*National Taipei University, New Taipei City, Taiwan,
†National University of Kaohsiung, Kaohsiung, Taiwan
‡email: sunc@gm.ntpu.edu.tw 446:NanChengSu.tex,session:CS39A

Recently, there are intense investigations about generalizing normal distributions. Among them, skew normal distributions and extended two-piece skew normal distributions are two classes of distributions, which both include standard normal distribution as special case. The later one is a class of skew distributions depending on two shape parameters. In this work, first some distributional properties of extended two-piece skew normal distributions will be presented. Next we revisit the special case, that is two-piece skew normal distributions. Then two distributions related to two-piece skew normal distributions will be studied. More precisely, we give some properties about generalized half normal distributions as well as a generalized Cauchy distribution. Finally, we discuss the distributions of linear combinations of two independent skew normal random variables.

Acknowledgment. This research was partially supported by the National Science Council of the Republic of China, grant No.: NSC 101-2118-M-006-003.

Profile Local Linear Estimation of Generalized Semiparametric Regression Model for Longitudinal Data

OCS21 Incomplete Longi. Data

YANQING SUN*,‡, LIUQUAN SUN†, JIE ZHOU†

*The University of North Carolina at Charlotte, Charlotte, USA,
†Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Beijing, China
‡email: yasun@uncc.edu 447:YanqingSun.tex,session:OCS21

We consider semiparametric modeling of covariate effects on a longitudinal response process based on repeated measurements observed at a series of sampling times. Suppose that there is a random sample of n subjects. For the ith subject, let $Y_i(t)$ be the response process and let $Z_i(t)$ and $X_i(t)$ be the possibly time-dependent covariates of dimensions $p \times 1$ and $q \times 1$, respectively, over the time interval $[0, \tau]$. We consider the following generalized semiparametric regression model for $Y_i(t)$, $0 \le t \le \tau$,

$$E\{Y_i(t)|X_i(t), Z_i(t)\} = g^{-1}\{\gamma^T(t)X_i(t) + \beta^T Z_i(t)\}, \qquad i = 1, \ldots, n, \tag{1}$$

where $g(\cdot)$ is a known link function, β is a p-dimensional vector of unknown parameters and $\gamma(t)$ is a q-dimensional vector of completely unspecified functions. The notation β^T represents transpose of a vector or matrix β. The first component of $X_i(t)$ is set to be 1, which gives a nonparametric baseline function. Under model (1), the effects of some covariates are constant while others are time-varying. Model (1) is more flexible than the parametric regression model where all the regression coefficients are time-independent and more desirable than the nonparametric model where every covariate effect

is an unspecified function of time. Different link functions can be selected to provide a richer family of models for longitudinal data. When the link function $g(\cdot)$ is the identity function, model (1) is known as the semiparametric additive model. The semiparametric additive model with longitudinal data has been studied extensively in recent years. When the link function is the natural logarithm function and $X_i(t) \equiv 1$, model (1) becomes the proportional means model. Model (1) unifies the semiparametric additive model and the proportional means model under the same umbrella.

This paper proposes a sampling adjusted profile local linear estimation method for the generalized semiparametric regression model (1). The paper has two main contributions. First, the proposed method automatically adjusts for heterogeneity of sampling times, allowing the sampling strategy to depend on the past sampling history as well as possibly time-dependent covariates without specifically model such dependence. Second, this paper presents an unified approach to the semiparametric model (1) with a general link function. The local linear estimation technique has been shown to be design-adaptive and more efficient in correcting boundary bias than the kernel smoothing approach for the cross-sectional data. We show that these features preserve under the proposed approach for longitudinal data. Large sample properties of the proposed estimators are investigated. Large sample pointwise and simultaneous confidence intervals for the regression coefficients are constructed. A formal hypothesis testing procedure is proposed to check whether the effect of a covariate is time-varying. A simulation study is conducted to examine the finite sample performances of the proposed estimation and hypothesis testing procedures. The method is illustrated with a data set from a HIV-1 RNA data set from an AIDS clinical trial.

Acknowledgment. The first author's research was partially supported by NSF grants DMS-0905777 and DMS-1208978, NIH grant 2 R37 AI054165-09 and a fund provided by UNC Charlotte. The second author's research was partly supported by the National Natural Science Foundation of China Grants (No. 10731010, 10971015 and 10721101), the National Basic Research Program of China (973 Program) (No. 2007CB814902) and Key Laboratory of RCSDS, CAS (No.2008DP173182).

POSTER
Poster

Attribute Diagrams for Diagnosing the Source of Error in Dynamical Systems

KAMONRAT SUPHAWAN*, RICHARD WILKINSON*, THEO KYPRAIOS*

*School of Mathematical Sciences, University of Nottingham, University Park Nottingham NG7 2RD, United Kingdom.

448:KSuphawan.tex,session:POSTER

In the past few decades, computer simulation has replaced physical experimentation in many areas of science. Whenever we try to simulate the behaviour of a real world process, we accept that the model is imperfect and we either ignore this imperfection, or occasionally, we try to quantify the error. Given an imperfect simulator, particularly a complex simulator, it can be extremely challenging to know which part of the model is incorrect. Numerous techniques have been suggested for model validation and diagnosing that there are forecast errors, but very little work has been done on diagnosing what the problem is given that one exists. In this poster, we describe methodology for diagnosing the source of the error. We try to attribute the error to either the simulator mean, the statistical model of the simulator discrepancy, the model of the measurement error process, or the variance of the measurement error, or a combination of all four.

We base our approach on the classical attribute or reliability diagram, which are diagnostic plots used for forecasts of binary events. They evaluate the calibration of the forecasting system by showing the degree of correspondence between forecasts and observed outcomes. We introduce a version of the attribute diagram that can be used to diagnose the source of any error in continuous dynamical

systems. We describe how particular shapes arise in the attribute diagram depending on the source of error in the forecasting system. We demonstrate the power of these diagrams on a simple linear Gaussian model, and on a rainfall-runoff model used to model rainfall in Abercrombie, Australia.

Some New Results on the Empirical Copula Estimator with Applications

CS6H Copula Estim.

JAN WILLEM HENDRIK SWANEPOEL[*,†], JAMES SAMUEL ALLISON[*]

[*]North-West University, Potchefstroom, South Africa
[†]email: jan.swanepoel@nwu.ac.za 449:JanSwanepoel.tex,session:CS6H

We derive the joint distribution of the ranks associated with a given bivariate random sample. Using these results, exact non-asymptotic expressions and asymptotic expansions for the mean and variance of the classical empirical copula estimator are obtained. An explicit expression of the coefficient appearing in the $O(1/n)$-term for the mean can, for example, be found; a result that apparently does not appear in the existing literature. Furthermore, it is shown that similar explicit non-asymptotic expressions as well as asymptotic expansions can be derived for the rank-based Bernstein copula estimator. The independence and comonotone copulas will receive special attention.

An Approximation of One-Dimensional Itô Diffusions Based on Simple Random Walks

OCS24 Random Graphs

JOHN VAN DER HOEK[*], TAMÁS SZABADOS[†,‡]

[*]University of South Australia, Adelaide, Australia,
[†]Budapest University of Technology and Economics, Hungary
[‡]email: szabados@math.bme.hu 450:TamasSzabados.tex,session:OCS24

The aim of this work is to develop a sequence of discrete approximations to a one-dimensional Itô diffusion that almost surely converges to a weak solution of the given stochastic differential equation (SDE) $dX(t) = \mu(t, X(t))\,dt + \sigma(t, X(t))\,dB(t)$, $X(0) = x_0$. The solution of the SDE is reduced to the solution of an ordinary differential equation (ODE), plus an application of Girsanov's theorem to adjust the drift. The ODE we use depends only on the diffusion coefficient of the SDE: $\phi'_u(t, u) = \sigma(t, \phi(t, u))$, $\phi(t, 0) = x_0$. Our tentative solution of the SDE is obtained by substituting Brownian motion into the obtained solution of the ODE: $X(t) := \phi(t, B(t))$. Then X is an Itô process with the correct diffusion coefficient and initial value, but its drift coefficient is not the one we wanted.

In order to adjust the drift, we keep the above tentative solution, but we want to change the original probability measure $\mathbb{P}$ to another probability $\mathbb{Q}$. A drift term can be suitably adjusted only if the corresponding Radon-Nikodym derivative is a martingale as a function of time: this is the bottleneck of the applicability of our method. Typically, some Novikov-type condition should be satisfied which may not hold in certain applications. When we are able to adjust the drift, our X will be a weak solution of the above SDE: $dX(t) = \mu(t, X(t))\,dt + \sigma(t, X(t))\,dW(t)$, $X(0) = x_0$, where W is a $\mathbb{Q}$-Brownian motion.

The proposed discrete approximation is based on a specific strong approximation of Brownian motion by an embedded sequence of simple, symmetric random walks $(B_m(t))_{t\geq 0}$ ($m = 0, 1, 2, \dots$). This is basically the so-called "twist and shrink" method, see [1]. Another ingredient of our approximation is a discrete Itô's formula, see also [1]. Then one can show that the discrete approximation $X_m(t^m) := \phi(t, B_m(t^m))$, $t^m := \lfloor t2^{2m} \rfloor 2^{-2m}$ ($m = 0, 1, 2, \dots$) uniformly converges to $X(t)$ for $t \in [0, T]$, almost surely as $m \to \infty$.

Moreover, one may define an approximating sequence of probability measures $\mathbb{Q}_m$ and drifts so that the resulting nearest neighbor random walks W_m are $\mathbb{Q}_m$-martingales and they almost surely uniformly converge on $[0,T]$ to the above-mentioned $\mathbb{Q}$-Brownian motion W. Further, X_m approximately satisfies a difference equation that corresponds to the above weak solution:

$$X_m(t_n) - x_0 = \sum_{r=1}^{n} \sigma\left(t_{r-1}, X_m(t_{r-1})\right)\ \left(W_m(t_r) - W_m(t_{r-1})\right) + \sum_{r=1}^{n} \mu\left(t_{r-1}, X_m(t_{r-1})\right) 2^{-2m} + O(2^{-m}), \quad (1)$$

where the error term $O(2^{-m})$ is uniform for $t_n := n2^{-2m} \in [0,T]$ almost surely, but may depend on ω.

References

[1] Szabados, T. (1996) An elementary introduction to the Wiener process and stochastic integrals. *Studia Sci. Math. Hung.*, **31**, 249-297.

CS8B Bayesian Nonpar.

Bayes Procedures for Adaptive Inference in Nonparametric Inverse Problems

BARTEK KNAPIK*, BOTOND SZABÓ†¶, AAD VAN DER VAART‡, HARRY VAN ZANTEN§

*CEREMADE, Université Paris-Dauphine, France,
†Eindhoven University of Technology, The Netherlands,
‡Leiden University, The Netherlands,
§University of Amsterdam, The Netherlands
¶email: b.szabo@tue.nl

451:BotondSzabo.tex,session:CS8B

Recent years have seen an increasing number of applications of Bayesian approaches in nonparametric statistical inverse problems, for instance in genomics and medical image analysis. An advantage of Bayesian approach is that various computational methods exist to carry out the inference in practice, including MCMC methods and approximate methods like approximate Bayesian computation. However there is still a lack of fundamental understanding of Bayes procedures for nonparametric inverse problems. Only a few, recent paper deal with consistency, convergence rates, etcetera.

In Bayesian analysis nonparametric priors typically involve one or more hyper-parameters, that determine the degree of regularization. The optimal choice of the hyper-parameter crucially depends on the regularity of the unknown truth. Since the value of the regularity parameter in practice is usually not available, one has to use data driven methods. In practice the probably most commonly used adaptive Bayesian techniques are the empirical and full, hierarchical Bayes approaches.

In our work we focus on the problem of estimating an infinite-dimensional parameter in mildly ill-posed inverse problems (for terminology see for instance [Cavalier]). We study this problem in the context of the canonical signal-in-white-noise model. By singular value decomposition many nonparametric, linear inverse problems can be transformed into this form. Specifically we assume to observe an infinite sequence of coefficients $X = (X_1, X_2, ...)$ satisfying

$$X_i = \kappa_i \theta_{0,i} + \frac{1}{n} Z_i,$$

where $Z_1, Z_2, ...$ are independent standard normal random variables, $\theta_0 = (\theta_{0,1}, \theta_{0,2})$ is the unknown infinite dimensional parameter of interest and κ_i is a known sequence which may converge to zero as $i \to \infty$ and therefore complicates the inference. We restrict ourselves to mildly ill-posed inverse problems of order $p < 0$ in the sense that $(1/C)i^{-p} \leq \kappa_i \leq Ci^{-p}$, for some $C > 0$. The optimal minimax rate over Sobolev balls of regularity β is $n^{-\beta/(1+2\beta+2p)}$.

We endow the unknown sequence θ_0 with the prior distribution Π_α, where the hyper-parameter α quantifies the "regularity" of the prior. We study both the empirical and hierarchical Bayes methods in this model. In the empirical Bayes method the hyper-parameter is estimated in a frequentist way and the estimator is plugged in into to posterior distribution. In the full Bayes approach the hyper-parameter is endowed with a hyper-prior distribution itself and we work with this two level hierarchical prior distribution. We show that both methods achieve the optimal minimax rate of contraction over a collection of Sobolev-balls. We illustrate our findings with a short simulation study.

Acknowledgment. Research supported by the Netherlands Organization for Scientific Research.

References

[Cavalier] Cavalier, L., 2008: Nonparametric statistical inverse problems. *Inverse Problems 24*, **3**, 034004, 19.

Almost Sure Local Limit Theorems for Strictly Stable Densities

CS19A Lim. Thms. Heavy Tails

RITA GIULIANO ANTONINI*, ZBIGNIEW S. SZEWCZAK†,‡

*Dipartimento di Matematica, Università di Pisa, Italy,
†Nicholas Copernicus University, Faculty of Mathematics and Computer Science, Toruń, Poland
‡email: eu2013@congress.hu

452:Szewczak.tex,session:CS19A

In the recent paper [Weber (2011)], the author proves a correlation inequality and an Almost Sure Local Limit Theorem (ASLLT) for i.i.d. square integrable random variables taking values in a lattice. The sequence of partial sums of such variables are of course in the domain of attraction of the normal law, which is stable of order $\alpha = 2$.

The aim is to give an analogous correlation inequality for the more general case of random sequences in the domain of attraction of a stable law of order $\alpha \leqslant 2$ and to apply it for the purpose of extending the theory of ASLLT. Notice that in our situation the summands need not be square integrable. Our correlation inequality turns out to be of the typical form needed in the theory of Almost Sure (Central and Local) Limit Theorems. Our method is completely different from the one used in [Denker et al. (2002)], [Giuliano et al. (2011)], [Giuliano et al. (2013)] and [Weber (2011)].

Acknowledgment. The financial support of the Research Grant PRIN 2008 *Probability and Finance* and of the INDAM–GNAMPA and the Polish National Science Centre Grant N N201 608740 are gratefully acknowledged. The authors are grateful to M. Weber for his interest in their work.

References

[Denker et al. (2002)] Denker, M., Koch, S., 2002: Almost sure local limit theorems, *Statist. Neerlandica*, **56**, 143-151.

[Giuliano et al. (2011)] Giuliano Antonini, R., Weber, M., 2011: Almost Sure Local Limit Theorems with rate, *Stoch. Anal. Appl.*, **29**, 779-798.

[Giuliano et al. (2013)] Giuliano Antonini, R., Szewczak, Z. S., 2013: An almost sure local limit theorem for Markov chains, *Statist. Probab. Lett.*, **83**, 573-579.

[Giuliano et al. (2013+)] Giuliano Antonini, R., Szewczak, Z. S., 2013: A general correlation inequality and the Almost Sure Local Limit Theorem in the domain of attraction of a stable law, *submitted*.

[Weber (2011)] Weber, M., 2011: A sharp correlation inequality with applicationsto almost sure local limit theorem, *Probab. Math. Statist.*, **31**, 79-98.

OCS14 Hungarian Stat. Assoc.

Estimation on Non-Response Bias

ROLAND SZILÁGYI*,†,‡

*University of Miskolc, Hungary,
†Hungarian Statistical Society, Budapest, Hungary
‡email: roland.szilagyi@uni-miskolc.hu 453:RolandSzilagyi.tex,session:OCS14

Research based on samples and their conclusions play an increasing role in making business decisions and also in creating information. The spread of sampling is mainly due to the lower expenses and shorter time needed for an investigation. Surveys based on samples are becoming more popular not only on a micro-level, but also in case of macroeconomic investigations. However, the spread of sampling has a great risk because of the quality of the samples.

As the priority of my work I chose the exploration of the possible faults of surveys based on samples and the negative effect they have on the result. Consecutively, I searched for solution variants for handling of mistakes, primarily for non-random errors. Based on domestic and international research findings, we can state that the biggest problem when carrying out surveys is when answers are not given. Obviously, selective answering not only reduces sample quantity but also increases the variants of estimates and the scale of bias. That is why I dealt with investigating one of the most important non-sampling error types: non-response errors. I have carried out my experimental and simulation analyses based on household budget data estimating the consumption expenditure.

Identifying tendencies plays a significant role in eliminating the bias caused by non-response. It is worth examining the differences between response and non-response criterion tendencies. In order to achieve a successful procedure the sample should be grouped based on variables which are in stochastic relation with the examined criterion and they are generating nonresponse. The tendencies must be examined and modelled according to this. In my study I formed groups based on household income to estimate the consumption expenses. I identified the exponential tendencies of consumption expenses in households with different incomes, paying attention to the different scope of non-response. I experienced that at lower levels of non-response the explanatory feature of functions was better; however, at higher levels the average of consumption expenses was considerably underestimated. That was the reason why I created an estimating model of weighted tendencies, in which the estimated values of the above-mentioned tendencies were defined as average estimated values by weighting the explanatory features of functions. In comparison the average consumption expenses counted with weighting based on estimating non-response probability or omitting non-responses can show as much as a 40% bias at higher non-response levels. At the same time the estimate model of weighted tendencies shows only a 5% bias. The model for estimating weighting tendencies can be used not only in realised nonresponse cases but also gives opportunity to reduce costs considerably by cutting the planned samples.

NYA Not Yet Arranged

Power Expansions for Perturbed Tests

PIOTR MAJERSKI*, ZBIGNIEW SZKUTNIK*,†

*AGH University of Science and Technology, Kraków, Poland
†email: szkutnik@agh.edu.pl 454:ZbigniewSzkutnik.tex,session:NYA

For simplicity or tractability reasons one sometimes uses modified test statistics, which differ from the original ones up to $O_p(a_n)$ terms with $a_n \to 0$. Some technical conditions will be discussed under which a corresponding expansion for the powers of such perturbed tests holds. The necessity of some of these conditions will be illustrated by examples.

To be somewhat more specific, assume that for a fixed $\alpha \in (0,1)$ and for $n \in \mathbb{N}$, size α tests for testing problems $(H_0^{(n)}, H_1^{(n)})$ have been chosen with rejection regions $\{S_n > c_\alpha^{(n)}\}$, where S_n are real valued test statistics and $c_\alpha^{(n)}$ are critical values. (All asymptotic symbols are considered with $n \to \infty$.) Let $\hat{S}_n$ be another sequence of statistics for which, under both $H_0^{(n)}$ and $H_1^{(n)}$,

$$S_n = \hat{S}_n + O_p(a_n),$$

with $a_n = o(1)$. Consider a sequence of tests that reject $H_0^{(n)}$ when $\hat{S}_n > \hat{c}_\alpha^{(n)}$, where $\hat{c}_\alpha^{(n)}$ are chosen to provide size α tests. Having the asymptotic expansion for the test statistics S_n with the leading term $\hat{S}_n$, our goal is to find a corresponding expansion for the powers of the tests. In other words, we aim at studying the asymptotic effect of a small perturbation term added to the tests' statistics on the tests' powers.

If, among other conditions, $\limsup_{n\to\infty} E\left|(S_n - \hat{S}_n)/a_n\right|^r < \infty$ for some $r > 0$, then for the power functions one has

$$\beta_S^{(n)}(\alpha) = \beta_{\hat{S}}^{(n)}(\alpha) + O\left(a_n^{\frac{r}{r+1}}\right).$$

As an application, results for quasi-most powerful invariant tests for multivariate normality will be presented that nicely explain high powers of likelihood-ratio tests between separate group families of distributions. Consider the group G^* of affine transformations in R^p with upper triangular matrices with positive diagonal. Clearly, the induced group acts transitively on the family of p-variate Gaussian distributions. Take as the alternative the family of distributions generated by G^* from the uniform distribution on the unit cube. The (complicated) most powerful G^*-invariant (MPI) test is related to the (simple) lihelihood-ratio (LR) test and it can be shown that for $p = 2$, for example, the powers are related as follows

$$\beta_{MPI} = \beta_{LR} + O\left((n/\log^2 n)^{-a}\right),$$

with a arbitrarily close to one. Similar results hold true for approximations to MPI tests obtained via Laplace expansions applied to their integral representations and for other alternatives.

Change Detection in a Heston Type Model

OCS28 Stat. Affine Proc.

GYULA PAP*, TAMÁS T. SZABÓ*,†

*University of Szeged, Szeged, Hungary
†email: tszabo@math.u-szeged.hu

455:TamasTSzabo.tex,session:OCS28

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

$$\begin{cases} dY_t = (a - bY_t)\,dt + \sigma_1\sqrt{Y_t}\,dW_t, \\ dX_t = (m - \beta Y_t)\,dt + \sigma_2\sqrt{Y_t}\big(\varrho\,dW_t + \sqrt{1-\varrho^2}\,dB_t\big), \end{cases} \qquad t \in \mathbf{R}_+, \tag{1}$$

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 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. We introduce a two-dimensional martingale by deducting from our process the conditional expectation on (Y_0, X_0), which will form the base of a process which can be shown to converge to a Wiener process. The ML estimates can then be substituted into this martingale to obtain a change detection method. The appropriate modification of the

aforementioned process will then converge to a Brownian bridge. Under the null hypothesis (i.e., no change), the procedure can also be thought of as a model-fitting test.

Acknowledgment. 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 et al. (2013)] Barczy, M., Doering, L., Li, Z., and Pap, G. (2013). Parameter estimation for an affine two factor model. arXiv: 1302.3451.

[Overbeck (1998)] Overbeck, L. (1998). Estimation for continuous branching processes. *Scandinavian Journal of Statistics*, Vol. 25, pp. 111–126.

[Pap and T. Szabó (2013)] 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.

NYA
Not Yet
Arranged

On Finite Memory Estimation of Stationary Ergodic Processes

ZSOLT TALATA*,†

*University of Kansas, Lawrence, Kansas, USA

†email: talata@math.ku.edu

456:ZsoltTalata.tex,session:NYA

The presentation is concerned with the problem of estimating stationary ergodic processes with finite alphabet from a sample, an observed length n realization of the process, with the $\bar{d}$-distance being considered between the process and the estimated one. The $\bar{d}$-distance was introduced by Ornstein (1973) and became one of the most widely used metrics over stationary processes. Two stationary processes are close in $\bar{d}$-distance if there is a joint distribution whose marginals are the distributions of the processes such that the marginal processes are close with high probability. The class of ergodic processes is $\bar{d}$-closed and entropy is $\bar{d}$-continuous, which properties do not hold for the weak topology.

[Ornstein and Weiss (1991)] proved that for stationary processes isomorphic to i.i.d. processes, the empirical distribution of the $k(n)$-length blocks is a strongly consistent estimator of the $k(n)$-length parts of the process in $\bar{d}$-distance if and only if $k(n) \leq (\log n)/h$, where h denotes the entropy of the process.

First, we estimate the n-length part of a stationary ergodic process X by a Markov process of order k_n. The transition probabilities of this Markov estimator process are the empirical conditional probabilities, and the order k_n depends on the sample size n, $k_n \to +\infty$. We obtain a rate of convergence of the Markov estimator to the process X in $\bar{d}$-distance, which consists of two terms. The first one is the bias due to the error of the approximation of the process by a Markov chain. The second term is the variation due to the error of the estimation of the parameters of the Markov chain from a sample. (Joint work with Imre Csiszár.)

Then, the order k_n of the Markov estimator process is estimated from the sample. For the order estimation, penalized maximum likelihood (PML) with general penalty term is used. The resulted Markov estimator process finds a tradeoff between the bias and the variation as it uses shorter memory for faster memory decays of the process X. If the process X is a Markov chain, the PML order estimation recovers its order asymptotically with a wide range of penalty terms.

In both cases, not only an asymptotic rate of convergence result is obtained but also an explicit bound on the probability that the $\bar{d}$-distance of the above Markov estimators from the process X is greater than ε. It is assumed that the process X is non-null, that is, the conditional probabilities of the symbols given the pasts are separated from zero, and that the continuity rate of the process X is

summable and the restricted continuity rate is uniformly convergent. These conditions are usually assumed in this area [Marton (1998)].

The result in the second case relies upon a non-asymptotic analysis of the PML Markov order estimation for not necessarily finite memory processes.

Acknowledgment. This research was supported in part by NSF grant DMS 0906929.

References

[Ornstein and Weiss (1991)] Ornstein, D.S., Weiss, B., 1990: How sampling reveals a process, *Ann. Probab.*, **18**, 905 - 930.

[Marton (1998)] Marton, K., 1998: Measure Concentration for a Class of Random Processes, *Probab. Theory Related Fields*,, **110**, 427 - 439.

Permuting Fractional Factorial Designs for Screening Quantitative Factors

OCS12 Experiment Design

YU TANG[*,‡], HONGQUAN XU[†]

[*]Soochow University, Suzhou, China,
[†]University of California, Los Angeles, USA
[‡]email: ytang@suda.edu.cn 457:Yu_Tang.tex,session:OCS12

Fractional factorial designs are widely used in various screening experiments. They are often chosen by the minimum aberration criterion that regards factor levels as symbols. For designs with quantitative factors, however, level permutation of factors could alter their geometrical structures and statistical properties. Most existing standard fractional factorial designs are not optimal for screening quantitative factors. We consider permuting levels for fractional factorial designs to improve their efficiency and develop some general theory. Many of these permuted designs are more efficient than existing standard designs for screening quantitative factors.

Acknowledgment. This research was partially supported by the National Natural Science Foundation of China with grant No. 11271279, the Natural Science Foundation of Jiangsu Province with grant No. BK2012612 and the Qing Lan Project.

Robust Scale and Autocovariance Estimation

CS4B Time Series I.

GARTH TARR[*,†], SAMUEL MÜLLER[*], NEVILLE WEBER[*]

[*]University of Sydney
[†]email: garth.tarr@sydney.edu.au 458:GarthTarr.tex,session:CS4B

Given the increasing prevalence of automated statistical methods in finance and other industries, it is important to consider robust approaches to analysing time series data. This talk outlines an intuitive, robust and highly efficient scale estimator, P_n, derived from the difference of two U-quantile statistics based on the same kernel as the Hodges-Lehmann estimate of location, $h(x, y) = (x + y)/2$ (Tarr, Müller and Weber, 2012). Through the device proposed by Gnanadesikan and Kettenring (1972) and discussed further in Ma and Genton (2000), we extend the robust scale estimator, P_n, to a robust autocovariance estimator, γ_P and discuss its properties.

The asymptotic results for P_n in the iid setting follow from nesting P_n inside the class of generalised L-statistics (GL-statistics; Serfling, 1984) which encompasses U-statistics, U-quantile statistics and L-statistics and as such contains numerous well established scale estimators. The interquartile

range, the difference of two quantiles, fits into the family of GL-statistics; as does the standard deviation which can be written as the square root of a U-statistic with kernel $h(x, y) = (x - y)^2$. The trimmed and Winsorised variance also fall into the class of GL-statistics. The robust scale estimator, Q_n as introduced by Rousseeuw and Croux (1993), can be represented in terms of a U-quantile statistic corresponding to the kernel $h(x, y) = |x - y|$.

The primary advantage of P_n is its high efficiency at the Gaussian distribution whilst maintaining desirable robustness and efficiency properties at heavy tailed and contaminated distributions. It will be shown how all of the aforementioned scale estimators can be transformed to autocovariance estimators. Efficiency results and asymptotics for P_n and γ_P will be discussed under both short and long range dependence.

References

[Gnanadesikan and Kettenring (1972)] Gnanadesikan, R. and Kettenring, J., 1972: Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, **28**, 81-124.

[Ma and Genton (2001)] Ma, Y. and Genton, M., 2001: Highly robust estimation of the autocovariance function, *J. Time Ser. Anal.*, **78**, 11-36.

[Rousseeuw and Croux (1993)] Rousseeuw, P. and Croux, C., 1993: Alternatives to the median absolute deviation, *J. Amer. Stat. Assoc.*, **88**, 1273-1283.

[Serfling (1984)] Serfling, R.J., 1984: Generalized L-, M-, and R-statistics, *Ann. Stat.*, **12**, 76-86.

[Tarr, Müller and Weber (2012)] Tarr, G., Müller, S. and Weber, N.C., 2012: A robust scale estimator based on pairwise means, *J. Nonparametr. Stat.*, **24**, 187-199.

CS16A Empirical processes

Volatility Occupation Times

VIKTOR TODOROV*, JIA LI†, GEORGE TAUCHEN†,‡

*Northwestern University, Evanston, USA,

†Duke University, Evanston, USA

‡email: george.tauchen@duke.edu 459:Tauchen.tex,session:CS16A

We propose nonparametric estimators of the occupation measure and its density of the diffusion coefficient (stochastic volatility) of a discretely observed Ito semimartingale on a fixed interval when the mesh of the observation grid shrinks to zero asymptotically. In a first step we estimate the volatility locally over blocks of shrinking length and then in a second step we use these estimates to construct a sample analogue of the volatility occupation time and a kernel-based estimator of its density. We prove the consistency of our estimators and further derive bounds for their rates of convergence. We use these results to estimate nonparametrically the quantiles associated with the volatility occupation measure.

Acknowledgment. We would like to thank Tim Bollerslev, Nathalie Eisenbaum, Jean Jacod, Andrew Patton and Philip Protter for helpful discussions. We are particularly grateful to Markus Reiss for suggesting the direct estimation approach adopted in the paper and the link with the uniform error in estimating the volatility path given in Lemma 2 of the paper.

Gaussian Suprema and a Significance Test for the LASSO

IS16 R. Fields, Geom.

RICHARD LOCKHART*, JONATHAN TAYLOR†,§, RYAN TIBSHIRANI‡, ROBERT TIBSHIRANI†

*Simon Fraser University, Burnaby, BC, Canada,
†Stanford University, Stanford, CA, USA,
‡Carnegie Mellon University, Pittsburgh, PA, USA
§email: jonathan.taylor@stanford.edu 460:JonathanTaylor.tex,session:IS16

In this talk we consider testing the significance of the terms in a fitted regression, the LASSO. We propose a novel test statistic for this problem, and show that it has a simple asymptotic null distribution. This work builds on the least angle regression (LARS) approach for fitting the LASSO , and the notion of degrees of freedom for adaptive models.

The analysis of this proposed test statistic is related to an overshoot related to a Gaussian process. The LASSO case corresponds to discretely indexed fields, while other examples, such as the group LASSO correspond to fields indexed on continuous parameter spaces.

A Generalization of Anderson's Procedure for Testing Partially Ranked Data

CS35A Discretre Response M.

JYH-SHYANG WU*, WEN-SHUN TENG*,†

*Tamkang University, New Taipei, Taiwan
†email: 121350@mail.tku.edu.tw 461:WENSHUN_TENG.tex,session:CS35A

In a consumer preference study, it is common to seek a ranking of a variety of alternatives or treatments (ice creams, beers, recreational facilities or cars etc.). When the number of alternatives, say r, is not too large, one may find it easy to rank all these alternatives simultaneously and statisticians appeal the procedure of Anderson (1959) to test the null hypothesis that each treatment has the same chance $1/r$ of receiving a given rank. Unfortunately, as r increases, it becomes progressively more confusing and undesirable for consumer to rank all r treatments simultaneously. In this presentation, we propose a new procedure that allows the setting where each surveyed consumers rank only her most preferred $k(k \leq r)$ treatments. Our proposed procedure includes Anderson's test as a special case as and is easy to implement our test as our proposed test statistics has a limit distribution of simple quadratic form in normal variables. We demonstrate our procedure with a set of (6/49) winning lottery numbers and Japanese sushi data. (This work is joint with Jyh-Shyang Wu)

Acknowledgment. This research was supported by the Taiwan National Science council, grant No.: 101-2118-M-032-001.

References

[ANDERSON (1959)] Anderson, R.L., 1959: Use of contingency tables in the analysis of consumer preference studies, *Biometrics*, **15**, 582- 590.

IS6 Financial Time Ser.

Conditional Correlation Models of Autoregressive Conditional Heteroskedasticity with Nonstationary GARCH Equations

CRISTINA AMADO*, TIMO TERÄSVIRTA*,†

*CREATES, Aarhus University, Denmark
†email: tterasvirta@econ.au.dk 462:TimoTerasvirta.tex,session:IS6

We investigate the effects of careful modelling the long-run dynamics of the volatilities of stock market returns on the conditional correlation structure. To this end we allow the individual unconditional variances in Conditional Correlation GARCH models to change smoothly over time by incorporating a nonstationary component in the variance equations. The modelling technique to determine the parametric structure of this time-varying component is based on a sequence of specification Lagrange multiplier-type tests derived in Amado and Ter?svirta (2011). The variance equations combine the long-run and the short-run dynamic behaviour of the volatilities. The structure of the conditional correlation matrix is assumed to be either time independent or to vary over time. We apply our model to pairs of seven daily stock returns belonging to the S&P 500 composite index and traded at the New York Stock Exchange. The results suggest that accounting for deterministic changes in the unconditional variances considerably improves the fit of the multivariate Conditional Correlation GARCH models to the data. The effect of careful specification of the variance equations on the estimated correlations is variable: in some cases rather small, in others more discernible. In addition, we find that portfolio volatility-timing strategies based on time-varying unconditional variances often outperforms the unmodelled long-run variances strategy in the out-of-sample. As a by-product, we generalize news impact surfaces to the situation in which both the GARCH equations and the conditional correlations contain a deterministic component that is a function of time.

CS13A Epidem. Models

Capture-Recapture Models in Epidemiology

JOANNE THANDRAYEN*,†

*University of Newcastle, Australia
†email: Joanne.Thandrayen@newcastle.edu.au 463:JoanneThandrayen.tex,session:CS13A

Capture-recapture models have long been applied to estimate the population size of wild animals. In more recent times these methods are now widely used to count human populations in various settings such as in epidemiology and in social sciences. One facet of epidemiological studies is concerned with estimating the number of people suffering from a particular disease. Traditionally, epidemiologists have tried to directly enumerate all individuals with a certain disease by merging records from various and distinct health services. These records may be available from hospital registrations, clinic enrolments, general practitioners lists, drug prescriptions and so on. However, this way of merging may lead to an undercount of the population as there are some cases that have not been identified by any of the lists. As an alternative to direct counting, the capture-recapture approach (also known as 'multiple-record systems') has been proposed to estimate the size of an epidemiological population.

The use of multiple lists gives rise to two main issues in the capture-recapture framework, namely list dependence and heterogeneity. The heterogeneity problem can be further classified as observed and unobserved heterogeneity. Here we review a series of new methods that we have proposed in the literature to account for these problems. We adopt covariate and latent class modelling techniques and perform maximum likelihood estimation via the EM algorithm. We illustrate our proposed methods by an application to data on diabetes patients in a town in northern Italy.

Acknowledgment. This is joint work with Yan Wang, RMIT University, Australia.

Analyzing Cell-to-Cell Heterogeneities in Gene Expression

IS23 Stat. Genetics, Biol.

FABIAN J. THEIS[*,†], FLORIAN BÜTTNER[*]

[*]Institute of Computational Biology, Helmholtz Center Munich, Germany,
[†]Department of Mathematics, Technical University Munich, Germany
email: fabian.theis@helmholtz-muenchen.de 464:Theis_Fabian.tex,session:IS23

Cell-to-cell variations in gene expression underlie many biological processes. Currently more and more experimental tools are becoming available in order to observe these variations, and to draw conclusions on underlying processes - for instance Munsky et al [MSB 2009] have shown that such information can be used for reducing model indeterminacies. However, given these experimental advances, we are now facing a series on computational questions dealing with these data, since classical analysis tools are often tailored to population averages.

Here I will discuss the analysis of single-cell qPCR expressions using nonlinear dimension reduction with an application to data from embryonic stem cell differentiation. The analysis is based on a recently proposed framework based on Gaussian process latent variable models (GPLVMs), which we extend by introducing gene relevance maps and gradient plots to provide interpretability as in the linear case. Furthermore, we take the temporal group structure of the data into account and introduce a new factor in the GPLVM likelihood which ensures that small distances are preserved for cells from the same developmental stage. As outlook, I briefly describe how to take additional data properties such as censoring into account.

Limit Theorems for the Empirical Distribution Function of Scaled Increments of Ito Semimartingales at high frequencies

IS11 Limit Thm. Appl.

VIKTOR TODOROV[*,‡], GEORGE TAUCHEN[†]

[*]Northwestern University, Evanston, USA,
[†]Duke University, Evanston, USA
[‡]email: v-todorov@northwestern.edu 465:Todorov.tex,session:IS11

We derive limit theorems for the empirical distribution function of "devolatilized" increments of an Ito semimartingale observed at high frequencies. These " devolatilized" increments are formed by suitably rescaling and truncating the raw increments to remove the effects of stochastic volatility and "large" jumps. We derive the limit of the empirical cdf of the adjusted increments for any Ito semimartingale whose dominant component at high frequencies has activity index of $1 < \beta \leq 2$, where $\beta = 2$ corresponds to diffusion. We further derive an associated CLT in the jump-diffusion case. We use the developed limit theory to construct a feasible and pivotal test for the class of Ito semimartingales with non-vanishing diffusion coefficient against Ito semimartingales with no diffusion component.

Acknowledgment. Research partially supported by NSF Grant SES-0957330.

CS20A R. Fields & Geom.

Kernel Density Estimators for Mixing Random Fields

CRISTINA TONE*,†

*University of Louisville, Louisville, Kentucky, USA

†email: cristina.tone@louisville.edu 466:ToneCristine.tex,session:CS20A

Density estimation and kernel density estimation for random processes have generated a considerable amount of interest and have been studied intensively in the literature. The extensive interest in kernel-type estimators of probability density is partly due to the fact that many useful stochastic processes, among them various time series models, satisfy a strong mixing property.

For a sequence of strictly stationary random fields that are uniformly ρ'-mixing and satisfy a Lindeberg condition, a central limit theorem is obtained for sequences of "rectangular" sums from the given random fields. The "Lindeberg CLT" is then used to prove a CLT for some kernel estimators of probability density for some strictly stationary random fields satisfying ρ'-mixing, and whose probability density and joint densities are absolutely continuous. The significance of our results on kernel estimators of probability density consists in having the feature that the normalizing constants are (asymptotically) the same as in the independent and identically distributed (i.i.d) case. Nevertheless, this fact shows that those procedures for estimating probability density are in a strong sense robust against a nontrivial departure (as the ρ'-mixing in our case) from the standard i.i.d. context.

CS22A R. Matrices

Estimation and Hypothesis Testing in High-Dimensional Transposable Data

ANESTIS TOULOUMIS*,‡, SIMON TAVARÉ†, JOHN MARIONI*

*European Bioinformatics Institute, Hinxton, United Kingdom,

†University of Cambridge, Cambridge, United Kingdom.

‡email: anestis@ebi.ac.uk 467:Touloumis.tex,session:CS22A

Transposable data refer to random matrices in which both the rows and the columns correspond to features of interest and dependencies might occur among and between the row and column variables. Transposable data are likely to occur in several fields, such as in genetics. For example, consider a cancer study where for each subject gene expression levels are measured in multiple tumor fragments and the tumor fragments satisfy a spatial and/or a temporal order. For each subject, we can write the data in a matrix form where the row variables correspond to genes and the column variables to tumor fragments. Interest might lie in drawing inference about the gene expression levels and the dependence structure between the genes and the tumor fragments.

Formally, let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be N independent and identically distributed transposable $p_1 \times p_2$ random matrices. Assume that N is a lot smaller than $p_1 \times p_2$ in order to reflect the 'small sample size, large number of parameters' situation. Challenges with high-dimensional transposable data are encountered in parsimonious modeling of the dependence structure between and among the row and column variables, in finding efficient estimating procedures and in developing testing procedures for accessing the structure of the row-wise and column-wise dependence pattern. A sensible choice for modelling transposable data is to utilize the matrix-variate normal distribution. This distribution defines three matrix parameters; the mean matrix $E[\mathbf{X}_i] = \mathbf{M}$, and two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ that satisfy the relation $Cov[vec(\mathbf{X}_i)] = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, where $vec()$ denotes the vec operator and $\otimes$ denotes the Kronecker matrix multiplication operator. The matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are identified as the covariance matrices that describe the dependence of the row variables and of the column variables, respectively.

In this talk, we present shrinkage estimators for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we discuss their properties and we compare them via simulation to penalized maximum likelihood based estimators. Further, we present testing procedures for the problem of hypothesis testing for the row covariance matrix $\mathbf{\Sigma}_1$ while treating the mean matrix $\mathbf{M}$ and the column covariance matrix $\mathbf{\Sigma}_2$ as 'nuisance'. The proposed tests are of a nonparametric nature as they do not specify the matrix-variate distribution of the transposable data and are suitable for testing the identity and the sphericity hypothesis for $\mathbf{\Sigma}_1$ with high-dimensional transposable data. In simulations, the proposed tests seem to preserve the nominal level when the null hypothesis is true and appear to be powerful against alternative hypotheses. Finally, we illustrate the above using an empirical example.

Adaptive Estimation of Quantiles in Deconvolution with Unknown Error Distribution

CS6E Function Est.

ITAI DATTNER*, MARKUS REISS†, MATHIAS TRABS†,‡

*EURANDOM, Eindhoven University of Technology, Eindhoven, The Netherlands,
†Humboldt-Universität zu Berlin, Berlin, Germany
‡email: trabs@math.hu-berlin.de 468:MathiasTrabs.tex,session:CS6E

We study the problem of quantile estimation in deconvolution with ordinary smooth error distributions. In particular, we focus on the more realistic setup of unknown error distributions. We develop a minimax optimal procedure and construct an adaptive estimation method under natural conditions on the densities. As a side result we obtain minimax optimal rates for the plug-in estimation of distribution functions with unknown error distributions. To prove our results, we study the deconvolution operator as random Fourier multiplier and we adopt Lepski's method. The estimation method is applied in simulations and on real data from systolic blood pressure measurements.

Solutions to Stochastic Heat and Wave Equation with Fractional Colored Noise: Existence, Regularity and Variations

IS27 SPDE

CIPRIAN TUDOR*,†,‡

*Université de Lillie 1, France,
†Academy for Economical Studies, Romania
‡email: tudor@math.univ-lille1.fr 469:Ciprian.Tudor.tex,session:IS27

We discuss recent results on the existence and the properties of the solutions to linear stochastic heat and wave equations driven by a Gaussian noise which behaves as a fractional Brownian motion in time and has correlated spatial covariance. Our presentation will include a discussion on the path regularity of the solution and the asymptotic behavior of its quadratic variations via Malliavin calculus.

Acknowledgment. The author ws supported by the CNCS grant PN-II-ID-PCCE-2011-2-0015 (Romania).

References

[Balan and Tudor (2007)] Balan, R., Tudor, C.A. 2007: The stochastic heat equation with fractional-colored noise: existence of the solution. *Latin Amer. J. Probab. Math. Stat.* 4, 57-87.

[Ouahhabi and Tudor (2013)] Ouahhabi, H., Tudor, C.A. 2013: Additive functionals of the solution to fractional stochastic heat equation. Preprint, to appear in *J. of Fourier Analysis and Applications.*

[Torres, Tudor and Viens (2013)] Torres, S., Tudor, C.A., Viens, F. 2013: Quadratic variations for the fractional-colored stochastic heat equation. Preprint.

POSTER
Poster

The use of Wildcards in Forensic DNA Database Searches

TORBEN TVEDEBRINK[*,†,‡]

[*]Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark,
[†]Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[‡]email: tvede@math.aau.dk 470:TvedebrinkTorben.tex,session:POSTER

Forensic genotyping is used as evidence and lead generating tool in many instances of the judicatory system. DNA is thought of as the golden standard within forensic science, as the chance of random matches between unrelated individuals attains values such as 1 in billions. Furthermore, the framework of population genetics enables the assessment of match probabilities between DNA profiles from various biological traces. Over the recent years, the national databases of DNA profiles have grown in size due to the success of forensic DNA analysis in solving crimes. The accumulation of DNA profiles implies that the probability of a random match or near match of two randomly selected DNA profiles in the database increases.

These reported near matches between supposedly unrelated individuals has caused some concern in the general public. However, Weir (2004, 2007) demonstrated elegantly that these near matches are close to what one would expect based on very simple population genetic models. Tvedebrink et al. (2012) derived computational efficient expressions for calculating the expectation and covariance of the near matches statistic for a given DNA database. In this work we show how the use of wildcards affect these quantities, and implement this in a R-package (`DNAtools`) for analysing DNA databases.

A forensic DNA profile can be thought of as a tuple of stochastic variables (list of genetic loci), where each entry is a two-dimensional vector containing of the genetic constitution (alleles) at a given position in the genome. The alleles of an individual is independently inherited from its parents' DNA profiles – one from the mother and one from the father. If the alleles are identical the locus is called homozygous and heterozygous otherwise. If one of the two different alleles at a heterozygous locus fail to be typed, a truly heterozygous profile will be typed as homozygote. This may be due to limited amounts of DNA or genetic anomalies, e.g. previously unseen mutations. An effect of such a mistyping is that the true contributor of the biological material may be excluded from further investigation due to a mismatch between the DNA profiles.

In order to avoid exclusions of DNA profiles due to mistyped heterozygous loci, the use of wildcards has been proposed and used by several national agencies. We show how the use of wildcards affect the probability of matches and near matches in a forensic DNA database.

Acknowledgment. I would like to thank Assoc. Prof. P.S. Eriksen[*], Prof. N. Morling[†] and Prof. J.M. Curran (University of Auckland) for valuable discussions, and Prof. P.D. Gill (University of Oslo) and Dr. J.S. Buckleton (ESR, New Zealand) for proposing an interesting research topic.

References

[Weir (2004)] Weir, BS., 2004. Matching and partially-matching DNA profiles. *J Forensic Sci* **49**, 1–6.

[Weir (2007)] Weir, BS., 2007. The rarity of DNA profiles. *Ann Appl Stat* **1**, 358–370.

[Tvedebrink et al. (2012)] Tvedebrink, T., Eriksen, PS., Curran, JM., Mogensen, HS., Morling, N., 2012. Analysis of matches and partial-matches in a Danish STR data set. *Forensic Sci Int-Gen.*, **6**, 387–392.

Some Characterizations for Mixed Poisson Processes in Terms of the Markov and the Multinomial Property

CS6D Dyn. Response Mod.

DEMETRIOS P. LYBEROPOULOS*, NIKOLAOS D. MACHERAS*, SPYRIDON M. TZANINIS*,†

*University of Piraeus, Department of Statistics and Insurance science
†email: stzaninis@unipi.gr 471:SpyridonTzaninis.tex,session:CS6D

It is well known that a mixed Poisson process (MPP for short) on a probability space (Ω, Σ, P) with mixing parameter a random variable Θ on Ω is always a MPP with mixing distribution P_Θ. The inverse implication is not always true, as it is in general not possible to construct regular conditional probabilities. We show that under an essential assumption both definitions coincide. As a consequence, the characterizations of MPPs with mixing distribution in terms of the multinomial and the Markov property can be carried over to MPPs with mixing parameter.

Moreover, the question whether a mixed renewal process (MRP for short) with mixing parameter a random variable Θ on Ω can be a MPP with the same mixing parameter is answered, under a mild assumption, to the positive and a characterization of MPPs as MRPs in terms of the Markov and the multinomial property is obtained. This result can be applied for further characterizations of MPPs as MRPs.

References

[Huang (1990)] Huang, W.J.: *On the Characterization of Point Processes with the Exchangeable and Markov Properties*, Sankhya, Volume 52, Series A, Pt. 1, pp. 16-27 (1990).

[Lyberopoulos - Macheras (2012)] Lyberopoulos, D.P. and Macheras, N.D. : *Some characterizations of mixed Poisson processes*, Sankhya, Volume 74, Series A, Pt. 1, pp. 57-79 (2012).

[Schmidt - Zocher (2003)] Schmidt, K.D. and Zocher, M.: *Claim Number Processes having the Multinomial Property*, Dresdner Schriften zur Versicherungsmathematik 1/2003.

Asymptotic Properties of Discriminant Functions for Stochastic Differential Equations from Discrete Observations

OCS6 Asympt. for Stoch Proc.

MASAYUKI UCHIDA*,†

*Osaka University
†email: uchida@sigmath.es.osaka-u.ac.jp 472:MasayukiUchida.tex,session:OCS6

We treat a discriminant analysis for stochastic differential equations based on sampled data. First we consider the situation where a discretely observed ergodic diffusion process $\mathbf{X}_n^{(0)}$ belongs to one of two diffusion models Π_1 and Π_2. One wishes to know whether the data $\mathbf{X}_n^{(0)}$ are obtained from Π_1 or Π_2 by using training data $\mathbf{X}_n^{(k)}$ from Π_k for $k = 1, 2$. The discriminant functions are constructed by the adaptive maximum likelihood type estimators derived from the training data ($\mathbf{X}_n^{(1)}$ and $\mathbf{X}_n^{(2)}$) and the quasi-likelihood functions, which are based on the approximation of the transition density by using the Ito-Taylor expansion, see Kessler (1997) for one-dimensional diffusions and Uchida and Yoshida (2012) for multi-dimensional diffusions. For details of maximum likelihood type estimators for both drift and volatility parameters of diffusion processes, see Yoshida (1992), Kessler (1995) and Uchida and Yoshida (2012). Based on the discriminant function, we propose a classification criterion and asymptotic distributions of the discriminant functions are shown under the two situations where the volatility functions are same or not. Next, we study a discriminant rule for stochastic differential

equations from sampled data observed on the fixed interval and show the asymptotic property of the discriminant function. We also prove that the misclassification probabilities based on the classification criteria converge to zero. This is a joint work with Nakahiro Yoshida.

References

[1] Kessler, M. (1995). Estimation des paramètres d'une diffusion par des contrastes corrigés. *C. R. Acad. Sci. Paris Ser. I Math.* **320**, 359–362.

[2] Kessler, M. (1997) *Estimation of an ergodic diffusion from discrete observations* , *Scandinavian Journal of Statistics*, **24**, 211–229.

[3] Uchida, M., Yoshida, N. (2012) *Adaptive estimation of an ergodic diffusion process based on sampled data* , *Stochastic Processes and their Applications*, **122**, 2885–2924.

[4] Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *J. Multivariate Anal.* **41**, 220–242.

CS17A Causal Inference

Learning DAGs Based on Sparse Permutations

GARVESH RASKUTTI*, CAROLINE UHLER†,‡

*SAMSI, Research Triangle Park, USA,
†IST Austria, Klosterneuburg, Austria
‡email: caroline.uhler@ist.ac.at 473:CarolineUhler.tex,session:CS17A

Determining causal structure among variables based on observational data is of great interest in many areas of science. While quantifying associations among variables is well-developed, inferring causal relations is a much more challenging task. A popular approach to make the causal inference problem more tractable is given by directed acyclic graph (DAG) models, which describe conditional dependence information and causal structure.

A popular way for estimating a DAG model from observational data employs conditional independence testing. Such algorithms, including the widely used PC algorithm [Spirtes et al. (2001)], require *faithfulness* (in addition to the standard Markov and causal minimality assumption) to recover the correct Markov equivalence class of the DAG. In [Uhler et al. (2013)] it is shown that the faithfulness assumption is very restrictive in many settings. A number of attempts have been made to weaken the faithfulness assumption (e.g. [Ramsey et al. (2006)]) and modify the PC algorithm to adjust for these weaker conditions. However, these relaxations of the faithfulness assumptions have ultimately led to weaker claims and all modifications of the PC algorithm do not guarantee discovery of the Markov equivalence class any longer.

We propose an alternative approach based on finding the permutation of the variables that yields the sparsest DAG. We prove that our sparsest permutation (SP) algorithm requires strictly weaker conditions for recovering the Markov equivalence class than the PC algorithm. In particular, we prove that the SP algorithm recovers the true Markov equivalence class under the Markov, causal minimality, and the sparsest Markov representation assumptions. These conditions are in fact necessary for consistency of any causal inference algorithm based on conditional independence testing. Through specific examples and simulations we also compare the SP algorithm to the PC algorithm in practice and show that the SP algorithm has better performance than the PC algorithm. Finally, we prove that when the variables are Gaussian, our SP algorithm is equivalent to finding the permutation of rows and columns for the inverse covariance matrix with the sparsest Cholesky decomposition. Using this connection, we show that in the oracle setting, where the true covariance matrix is known, our SP algorithm is equivalent to the penalized maximum likelihood approach in [van de Geer et al. (2013)].

References

[Ramsey et al. (2006)] . Ramsey, J., Zhang, J. and Spirtes, P., 2006: Adjancency-faithfulness and conservative causal inference, in *Uncertainty in Artificial Intelligence (UAI)*.

[Spirtes et al. (2001)] Spirtes P., Glymour, C. and Scheines, R., 2001: *Causation, Prediction and Search*, MIT Press.

[Uhler et al. (2013)] Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B., 2013: Geometry of faithfulness assumption in causal inference, to appear in *Annals of Statistics*.

[van de Geer et al. (2013)] van de Geer, S. and Bühlmann, P., 2013: Penalized maximum likelihood estimation for sparse directed acyclic graphs, to appear in *Annals of Statistics*.

A Model-Based Method for Analysing Attribute Measurement Systems

OCS11
ENBIS

EMESE VÁGÓ*,†, SÁNDOR KEMÉNY*

*University of Technology and Economics, Department of Chemical and Environmental Process Engineering, Budapest

†email: evago@mail.bme.hu 474:Vago.tex,session:OCS11

This presentation focuses on the analysis of measurement systems with two possible measurement outcomes: the measured part can be either rejected or accepted. Behind the attribute type decision there is often a continuous variable (reference value) that is not measured in practice, but it is known during the gauge analysis. I have investigated these cases in my research.

The most widely used approaches of attribute measurement system analysis can be divided into two main parts: the crosstabulation and the analytic method. In the crosstabulation method a random sample is taken from the parts, and different operators make repeated observations on each. Based on the results different countdown indexes are calculated. These are used to measure the agreement between the raters and also between the raters and the reference decision. The aim of the analytic method is the graphical estimation of the gauge performance curve. The gauge performance curve is an S-shaped curve that describes the connection between the reference value and the probability of acceptance. In my work I have revealed several theoretical errors of the crosstabulation and the analytic method.

Instead of the generally used AIAG method [1] I have proposed a new approach for attribute measurement system analysis. The basic novelty of the new method is that it uses the mathematical model of the gauge-performance-curve. Using the proposed logit model the practically interesting characteristics of the gauge performance curve can be estimated, thus it provides a theoretically correct alternative of the analytic method.

I have suggested the use of two conditional probabilities to measure the gauge capability. These are the probability that an accepted part is bad and the probability that a rejected part is good. These probabilities can be estimated both with the proposed model-based approach and the following the AIAG Manuals' crosstabulation method. I have compared the efficiency of the two estimation methods partly analytically and partly with simulation. I have found that sufficient improvement can be achieved in the estimation error with the new, model-based method.

References

[1] Automotive Industry Action Group, 2002. Measurement System Analysis; Reference Manual, 3rd ed. Detroit, MI: Automotive Industry Action Group, 125-140.

POSTER
Poster

Clustering Correlated Time Series via Quasi U-Statistics

MARCIO VALK*,†

*Federal University of Rio Grande do Sul, Porto Alegre, Brazil
†email: marciovalk@gmail 475:MarcioValk.tex,session:POSTER

Discrimination and classification time series becomes almost indispensable since the large amount of information available nowadays. The problem of time series discrimination and classification is discussed in [1]. In this work the authors propose a novel clustering algorithm based on a class of quasi U-statistics and subgroup decomposition tests. The decomposition may be applied to any concave time-series distance. The resulting test statistics is proved to be asymptotically normal for either i.i.d. or non-identically distributed groups of time-series under mild conditions. In practice there are many time series that are correlated among themselves. An example that can describe this fact is the financial markets globalization. When one of these markets is affected by an exogenous factor, a chain reaction can affect many others. So the independence condition fail.

We are interested in analyzing how the correlation among the groups of time series can affect classification and clustering methods especially the one proposed by [1]. Empirical results show that the proposed method is robust to the presence of correlation among time series. The convergence of the test statistic for dependent time series will be one of the goals in this work.

References

[1] Valk, M., Pinheiro, A. 2012: Time-series clustering via quasi U-statistics, *J. Time Ser. Anal.,* Vol. 33, 4, 608 - 619.

Opening
Opening Lecture

High Dimensional Statistics, Sparsity and Inference

SARA VAN DE GEER*,†

*Seminar for Statistics, ETH Zürich, Switzerland
†email: geer@stat.math.ethz.ch 476:vandeGeer.tex,session:Opening

High-dimensional models have gained a prominent role in statistical research. The methodology to deal with high dimensionality is generally based on sparsity-inducing regularization penalties such as the ℓ_1-penalty. The Lasso is an important example but there are many others. Most mathematical statistical results address issues such as compression and information theoretic bounds. Classical frequentist inference using confidence intervals and p-values is largely missing and will be the theme of this talk.

In the first part we will review some of the existing theory. We present some general oracle inequalities for sparsity-regularized estimators and briefly discuss the role of sparsity-inducing penalties versus the sparsity-enforcing ℓ_0-penalty.

The oracle inequalities say that the estimator adapts to unknown sparsity. However, accessing the accuracy of an adaptive estimator is generally not possible . This is one of the uncertainty principles in statistics. In this sense sparsity and inference seem not to go along.

In the second part we discuss this further and propose the semi-parametric approach as an escape route. We present some methods for deriving confidence intervals for lower-dimensional parameters of interest. We show that with sparsity-inducing methods the accuracy of a de-sparsified estimator can be estimated. Thus, sparsity may actually be crucial for inference in high dimensions. Also here, we compare inducing sparsity to enforcing sparsity.

Finally, thinking about uncertainty principles in high-dimensional models leads to considering asymptotic Cramer-Rao lower bounds in this context. We discuss what would be the semi-parametric efficiency bound and whether it can be attained.

Estimation of Conditional Ranks and Tests of Exogeneity in Nonparametric Nonseparable Models

IS8 Function Estim.

FRÉDÉRIQUE FÈVE*, JEAN-PIERRE FLORENS*, INGRID VAN KEILEGOM†,‡

*Toulouse School of Economics, France,
†Université catholique de Louvain, Louvain-la-Neuve, Belgium
‡email: ingrid.vankeilegom@uclouvain.be 477:Ingrid_van_Keilegom.tex,session:IS8

Consider a nonparametric nonseparable regression model $Y = g(Z, U)$, where $g(Z, U)$ is increasing in U and $U \sim U[0, 1]$. We suppose that there exists an instrument W that is independent of U. The observable random variables are Y, Z and W, all one-dimensional. The purpose of this paper is twofold. First, we study the asymptotic properties of a kernel estimator of the distribution of $V = F_{Y|Z}(Y|Z)$, which equals U when Z is exogenous. We show that this estimator converges to the uniform distribution at faster rate than the parametric $n^{-1/2}$-rate. Next, we construct test statistics for the hypothesis that Z is exogenous. The test statistics are based on the observation that Z is exogenous if and only if V is independent of W, and hence they do not require the estimation of the function g. The asymptotic properties of the proposed tests are proved, and a bootstrap approximation of the critical values of the tests is shown to work for finite samples via simulations. An empirical example using the U.K. Family Expenditure Survey is also given.

Acknowledgment. This research was supported by a grant from the European Research Council (ERC), an IAP research network grant of the Belgian government, and an ARC grant of the 'Communauté française de Belgique'.

Modelling Aftershock Sequences of the 2005 Kashmir Earthquake

CS19B Lim. Thms. Point Proc.

K. TÜRKYILMAZ*, MARIE-COLETTE N.M. VAN LIESHOUT*,‡, A. STEIN†

*CWI, Centre for Mathematics and Computer Science, Amsterdam, The Netherlands,
†Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
‡email: Marie-Colette.van.Lieshout@cwi.nl
478:MarieColettevanLieshout.tex,session:CS19B

We explore the pattern of aftershocks following the Kashmir earthquake. We discuss the Hawkes and the trigger process models and estimate their parameters by the minimum contrast method. It turns out that the trigger model is the better fit, with about two main shocks explaining the observed pattern. We refine the model by fitting a mixture model of two bivariate normal distributions.

Weighted Bootstrap Methods in Modelling Multivariate Financial Data

OCS26 Resampling Nonstat T.S.

LÁSZLÓ VARGA*,†, ANDRÁS ZEMPLÉNI*

*Eötvös Loránd University, Budapest, Hungary
†email: vargal4@cs.elte.hu 479:LaszloVarga.tex,session:OCS26

Financial time series are often modelled by complex structures like GARCH versions. To estimate the uncertainty of the fitted time series models is of fundamental importance. However, this is not

an easy task, and we suggest the application of the recently introduced weighted bootstrap to solve this problem.

We have shown in [1] that it possesses favourable asymptotical properties in the univariate case, and it can be analogously used in the multivatiate setup, for example for constructing confidence bounds simultaneously. In the talk we shall introduce the method and show its effectivity when applied to daily log-returns of stock exchange data. The weights are directly applied to the likelihood function, so it is not more complicated to estimate the parameters as for the original data. So it is a good candidate to investigate the strength of dependencies among stocks, possibly not detectable by classical methods, based on correlation.

References

[1] L. Varga and A. Zempléni, 2013. Weighted bootstrap in GARCH models. Submitted.

POSTER
Poster

Comparative Analysis of a One-Channel Microarray Dataset by Different Methods

DUYGU VAROL*, VILDA PURUTÇUOĞLU*,†

*Department of Statistics, Middle East Technical University, Ankara, Turkey
†email: vpurutcu@metu.edu.tr 480:DuyguVarol.tex,session:POSTER

The microarray experiments allow the researchers to investigate the differentially expressed genes and their levels in biologically interesting systems. In this study, firstly, we analyse a real one-channel microarray dataset which has not been evaluated statistically in detail. This dataset includes gene expressions of saccharomyces cerevisiae under certain stress, namely, the effect of heat shock and heat change. In the data, apart from these two stress groups, there is one control group as well. So there are two biological replicates for each treatment/stress group and the gene expressions for each group are measured after six-hour and one-hour, respectively. In the analysis, we use the `smida` package in the R programme language for the preprocessing of data before the analysis, which covers the spatial, background, within and across condition normalization, in order. As the outputs, we aim to find biologically interesting results for this yeast. In the assessment, we also implement clustering analysis after detecting significant genes to find their functional relationships. On the other hand, as the second novelty in this study, we normalize the dataset via a newly developed background normalization method, called *multi-RGX*. From the comparison based on benchmark data, it has been shown that the results of *multi-RGX* are promising in terms of accuracy and computational time with respect to its strong alternatives. Whereas its performance has not been yet evaluated in a real dataset. We consider that this method can help us to investigate new significant genes regarding previous findings during the normalization of the data and, thereby, enable us to discover novel functional relations between genes.

Acknowledgment. The authors would like to thank Dr. Remziye Yılmaz and Dr. Mehmet Cengiz Baloğlu for their valuable explanation about the design of the experiment and providing the data.

References

[1] Wit, E. & McClure, J., 2004: Statistics for Microarray: design, analysis and inference, Chichester, John Wiley and Sons.

[2] Yilmaz, R., Baloğlu, M.C., Ercan, O., Öktem, H.A., & Yücel, M., 2009: Detection of the gene expression levels of bakers' yeast (saccharomyces cerevisiae) under different heat stresses via microarray method (in Turkish), Proceeding of the 16th National Biotechnology Congress, 344-347, Antalya, Turkey.

[3] Purutçuoğlu, V. & Akal, T., 2012: multi-RGX: a novel background normalization method for oligonucleotides, Proceeding of the 8th World Congress in Probability and Statistics, İstanbul, Turkey.

Testing Granger Causality in Time-Varying Framework

OCS14 Hungarian Stat. Assoc.

GÁBOR RAPPAI*,‡, VIKTOR VÁRPALOTAI†

*University of Pécs, Hungary,
†Ministry for National Economy, Budapest, Hungary
‡email: rappai@ktk.pte.hu 481:RappaiVarpalotai.tex,session:OCS14

Widely used Granger causality test is based on assumption that causality relation is unchanged in the entire sample i.e. either it is present or it is absent in each point in time. However it is well documented in the empirical economic literature that co-movement of economic time series is subject to substantial change over time (see for example Stock-Watson 1996). Because of the potentially unstable relation among economic time series we consider the standard Granger causality test being oversimplified as it can yield black or white type answer only.

To overcome this oversimplification, in this paper we propose an extension of the standard Granger causality test with special focus on time varying property of economic time series. In this regard we follow the recent literature that test Granger causality where structural breaks are present in the data (see for example Christopoulos-Ledesma (2008), Lou et al (2012) and Balcilar-Ozdemir (2013)). The aforementioned papers combine regime switching VAR framework with Granger causality testing assuming at most two different regimes during the sample.

In our view, enabling only two regimes in a sample is still a very restrictive framework. Therefore we propose a time varying coefficient VAR (TV-VAR) framework in which the presence of Granger causality can be tested in each period without any restriction on potential number of regimes. Our framework is capable to detect not only the presence or absence of Granger causality but subsequent periods also where the Granger causality continuously strengthens or weakens. Our method is illustrated on Hungarian macroeconomic data.

Acknowledgment. This research was partially supported by the Social Renewal Operational Programme, grant No.: SROP-4.2.2.C-11/1/KONV-2012-0005, Well-being in the Information Society and the Hungarian Statistical Association (Magyar Statisztikai Társaság).

Kernel Type Estimator of a Bivariate Average Growth Function

CS6A Funct. Est., Kernel Meth.

ISTVÁN FAZEKAS*, ZSOLT KARÁCSONY†, RENÁTA VAS*,‡

*University of Debrecen, Debrecen, Hungary,
†University of Miskolc, Miskolc, Hungary
‡email: vas.renata@inf.unideb.hu 482:RenataVas.tex,session:CS6A

The usual regression problem is considered. A bivariate average growth function with general error is studied. That is, the random field observed is of the form $Y(u,v) = f(u,v) + \varepsilon(u,v)$ where f is the unknown bivariate function to be estimated and ε is the error random field. The Gasser-Müller method is used to construct the estimator of the unknown function. Under general and realistic conditions on the covariance structure of the error random field an upper bound is obtained for the mean squared error. The upper bound is expressed in terms of the derivatives of the covariance function. The proof of our theorem is based on appropriate versions of the Taylor expansion (avoiding the singularities of the covariance function). The result can be used for several particular covariance structures. Numerical evidence is also presented by simulation. The results obtained are the two-dimensional versions of the ones presented in [Benhenni and Rachdi, 2007].

Acknowledgment. The work was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

References

[Benhenni and Rachdi, 2007] Benhenni, K., Rachdi, M., Nonparametric estimation of average growth curve with general nonstationary error process. *Communications in Statistics - Theory and Methods* **36** (2007), 73-86.

IS28 Stoch. in Biol.

The Effects of a Weak Selection Pressure in a Spatially Structured Population

ALISON ETHERIDGE*, AMANDINE VÉBER†,§, FENG YU‡

*University of Oxford, UK,
†Ecole Polytechnique, Palaiseau, France,
‡University of Bristol, UK.
§email: amandine.veber@cmap.polytechnique.fr 483:AMANDINE_VBER.tex,session:IS28

One of the motivations for the introduction of the Fisher-KPP equation was to model the wave of advance of a favourable (genetic) type in a population spread over some continuous space. This model relies on the fact that reproductions occur very locally in space, so that if we assume that individuals can be of two types only, the drift term modelling the competition between the types is of the form $sp_{t,x}(1-p_{t,x})$. Here, s is the strength of the selection pressure and $p_{t,x}$ is the frequency of the favoured type at location x and time t. However, large-scale extinction-recolonisation events may happen at some nonnegligible frequency, potentially disturbing the wave of advance. In this talk, we shall address and compare the effect of weak selection in the presence or absence of occasional large-scale events, based on a model of evolution in a spatial continuum called the *spatial Lambda-Fleming-Viot process*.

CS6H Copula Estim.

Bernstein Estimator for a Copula and its Density

PAUL JANSSEN*, JAN SWANEPOEL†, NOËL VERAVERBEKE*,†,‡

*Hasselt University, Hasselt, Belgium,
†North-West University, Potchefstroom, South Africa
‡email: noel.veraverbeke@uhasselt.be 484:NoelVeraverbeke.tex,session:CS6H

Copulas are functions that couple the multivariate distribution function $H(x,y)$ of a random vector (X,Y) to its one-dimensional marginals $F(x)$ and $G(y)$. According to Sklar's theorem, there exists a bivariate function C, called copula, such that $H(x,y) = C(F(x),G(y))$. Several papers deal with the estimation of C, based on a random sample of size n from (X,Y).

In this talk we discuss the asymptotic properties of the so called Bernstein estimation method. This nonparametric smoothing method approximates $C(u,v)$ by a polynomial of degree m in (u,v). Asymptotics are considered as n and m tend to infinity. The estimator for the copula C leads in a very natural way to an estimator for the corresponding density c of C. Asymptotic normality is obtained and optimal order of the degree m is discussed. Compared to the existing results our theorem does not assume known marginals.

Minimax Risks for Sparse Regressions: Ultra-High Dimensional Phenomenons

IS13 Model Selection

NICOLAS VERZELEN*,†

*INRA, UMR 729 MISTEA, Montpellier
†email: nicolas.verzelen@supagro.inra.fr 485:NicolasVerzelen.tex,session:IS13

Consider the standard Gaussian linear regression model $\mathbf{Y} = \mathbf{X}\theta_0 + \boldsymbol{\epsilon}$, where $\mathbf{Y} \in \mathbb{R}^n$ is a response vector and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix. Numerous work have been devoted to building efficient estimators of θ_0 when p is much larger than n. In such a situation, a classical approach amounts to assume that θ_0 is approximately sparse. In this talk, we study the minimax risks of estimation and testing over classes of k-sparse vectors θ_0. These bounds shed light on the limitations due to high-dimensionality. The results encompass the problem of prediction (estimation of $\mathbf{X}\theta_0$), the inverse problem (estimation of θ_0) and linear testing (testing $\mathbf{X}\theta_0 = 0$). Interestingly, an elbow effect occurs when the number of variables $k \log(p/k)$ becomes large compared to n. Indeed, the minimax risks and hypothesis separation distances blow up in this ultra-high dimensional setting. We also prove that even dimension reduction techniques cannot provide satisfying results in an ultra-high dimensional setting. Moreover, we compute the minimax risks when the variance of the noise is unknown. The knowledge of this variance is shown to play a significant role in the optimal rates of estimation and testing. All these minimax bounds provide a characterization of statistical problems that are so difficult so that no procedure can provide satisfying results.

Stationary Solution of 1D KPZ Equation

CS22A R. Matrices

ALEXEI BORODIN*, IVAN CORWIN*, PATRIK L. FERRARI†, BÁLINT VETŐ†,‡

*Massachusetts Institute of Technology, Department of Mathematics, Cambridge, MA, USA,
†Bonn University, Institute for Applied Mathematics, Bonn, Germany
‡email: vetob@uni-bonn.de 486:BalintVeto.tex,session:CS22A

The KPZ equation is believed to describe a variety of surface growth phenomena that appear naturally, e.g. crystal growth, facet boundaries, solidification fronts, paper wetting or burning fronts. In the recent years, serious efforts were made to describe the solution with different types of initial data. In the present work, we derive an explicit solution for the equation with stationary, i.e. two-sided Brownian motion initial condition.

Our approach to the solution for the KPZ equation is via the continuum directed random polymer model, since the Hopf – Cole solution to the KPZ equation can be represented as the logarithm of the partition function of this model. As a semi-discrete analogue of the continuum directed polymer model with the appropriate boundary data, we consider an extended version of the O'Connell – Yor semi-discrete polymer model which scales to the continuum model under the intermediate disorder scaling. By providing integral formulas for the action of Macdonald difference operators, we characterize explicitly the Laplace transform of the partition function of the semi-discrete polymer model by giving a Fredholm determinant formula. Via a certain double limit of semi-discrete polymer models, we obtain a formula for the free energy of the continuum directed polymer model with stationary boundary perturbation which gives the solution to the stationary KPZ equation.

In the large time limit of the solution, we recover the distribution obtained for the limiting fluctuations of the height function of the stationary totally asymmetric simple exclusion process (TASEP).

IS11 Limit Thm. Appl.

Statistical Inference on Lévy Measures and Copulas

BÜCHER, AXEL*, VETTER, MATHIAS*,†

*Ruhr-Universität Bochum, D-44780 Bochum, Germany

†email: mathias.vetter@rub.de 487:Vetter.tex,session:IS11

In this talk nonparametric methods to assess the multivariate Lévy measure are introduced. Starting from high-frequency observations of a Lévy process $\mathbf{X}$, we construct estimators for its tail integrals and the Pareto Lévy copula and prove weak convergence of these estimators in certain function spaces. Given n observations of increments over intervals of length Δ_n, the rate of convergence is $k_n^{-1/2}$ for $k_n = n\Delta_n$ which is natural concerning inference on the Lévy measure. Besides extensions to non-equidistant sampling schemes analytic properties of the Pareto Lévy copula which, to the best of our knowledge, have not been mentioned before in the literature are provided as well. We conclude with a short simulation study on the performance of our estimators and apply them to real data.

Acknowledgment. This research was supported by the German Research Foundation (DFG) via SFB 823 "Statistics for nonlinear dynamic processes".

OCS30 Stoch. Neurosci.

Delayed Feedback Results in non-Markov Statistics of Neuronal Activity

KSENIIA KRAVCHUK*, ALEXANDER VIDYBIDA*,†

*Bogolyubov Institute for Theoretical Physics, Kyiv, Ukraine

†email: vidybida@bitp.kiev.ua 488:AlexanderVidybida.tex,session:OCS30

It is observed that the sequence of interspike intervals (ISI) $\{t_1, t_2, \dots\}$ of a neuron can be nonrenewal and even non-Markov, see, e.g. [Ratnam and Nelson (2000)]. We expect that the nonrenewal statistics arises due to delayed feedback communications ubiquitous in real neuronal nets. In order to check this hypothesis, we consider the simplest possible "network", namely, a single binding neuron with delayed feedback, Fig. 1.

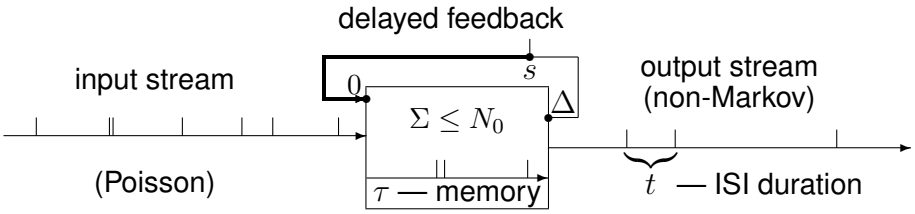

Fig. 1. Binding neuron with delayed feedback. Here Δ denotes total delay in the feedback line; s is the time required for an impulse to reach the output end of the line; τ is the time during which an impulse is kept in the neuron; N_0 is the threshold: if the number of kept impulses is equal to N_0 the neuron fires a spike. In this work, $N_0 = 2$.

For this construct we introduce the stream of events (t, s): $\mathbf{ts} = \{\dots, (t_i, s_i), \dots\}$, where s_i is the time to live of the impulse in the feedback line at the moment, when ISI t_i starts.

Lemma 10. *Stream* **ts** *is the 1-st order Markovian:*

$$\forall_{n\geq 0}\forall_{t_0>0}\forall_{s_0\in]0;\Delta]}\dots\forall_{t_{n+1}>0}\forall_{s_{n+1}\in]0;\Delta]}P(t_{n+1}, s_{n+1} \mid t_n, s_n; \dots; t_0, s_0) = P(t_{n+1}, s_{n+1} \mid t_n, s_n),$$

where $\{t_0, \dots, t_{n+1}\}$ *is the set of successive ISIs, and* $\{s_0, \dots, s_{n+1}\}$ *are the corresponding times to live.*

This Lemma allows us to calculate exact expressions for multievent probabilities $P(t_n, s_n; \ldots; t_0, s_0)$ for the $\mathbf{ts}$, then $P(t_n; \ldots; t_0)$ (as marginal probabilities) and finally conditional probabilities $P(t_{n+1} \mid t_n; \ldots; t_0)$ for the output stream of ISIs as functions of Δ, τ and λ — the input stream intensity. Having exact expressions for $P(t_{n+1} \mid t_n; \ldots; t_0)$ we prove the following theorem:

Theorem 11. *The output ISIs stream of binding neuron with delayed feedback under Poisson stimulation cannot be represented as a Markov chain of any finite order.*

References

[Ratnam and Nelson (2000)] Ratnam, R., Nelson, M.E., 2000: Nonrenewal statistics of electrosensory afferent spike trains: implications for the detection of weak sensory signals, *The Journal of Neuroscience*, **20**, 6672-6683.

[Vidybida and Kravchuk (2012)] Vidybida, A. and Kravchuk, K., 2012: Delayed feedback makes neuronal firing statistics non-markovian, *Ukrainian Mathematical Journal*, **64**, 1587-1609.

Convergence Properties of Pseudo-Marginal Markov Chain Monte Carlo Algorithms

IS1 Bayesian Comp.

CHRISTOPHE ANDRIEU*, MATTI VIHOLA†,‡

*University of Bristol, UK,
†University of Jyväskylä, Finland
‡email: matti.vihola@iki.fi

489:MattiVihola.tex,session:IS1

Pseudo-marginal Markov chain Monte Carlo (MCMC) is a generic emerging class of algorithms for computationally challenging Bayesian inference (Beaumont, *Genetics* 164, 2003; Andrieu & Roberts, *Ann. Statist.* 37, 2009). In general, the pseudo-marginal methods may be applied when the density of interest $\pi(x)$ cannot be evaluated (or is computationally expensive to evaluate), but it can be estimated at any point x in an unbiased manner. More specifically, if for any x it is possible to generate a non-negative estimate $\hat{\pi}(x)$ which has the expectation $\pi(x)$, then it is possible to implement pseudo-marginal MCMC which is exact in the sense that it will be ergodic with respect to an invariant distribution with the correct marginal distribution $\pi(x)$.

Perhaps the most well-known instance of such an algorithm is the particle MCMC (Andrieu, Doucet & Holenstein, *J. R. Stat. Soc. Ser. B* 72, 2010), which uses the particle filter to generate the unbiased estimates $\hat{\pi}(x)$. The particle MCMC has opened new possibilities of Bayesian inference in time-series (state-space) models. Another important example is the algorithm for inference of discretely observed diffusions (Beskos, Papaspiliopoulos, Roberts & Fearnhead *J. R. Stat. Soc. Ser. B* 68, 2006), which can also seen as a pseudo-marginal algorithm.

Our work on the theoretical convergence properties of the pseudo-marginal MCMC helps to understand when the methods may be useful and what are their fundamental limitations. Our first finding is an expected limitation: the pseudo-marginal MCMC is always worse in terms of asymptotic variance than the corresponding (ideal) 'marginal' algorithm, where $\pi(x)$ is directly available. On the other hand, when the estimators $\hat{\pi}(x)$ are made more accurate, the resulting pseudo-marginal Markov chain will approximate the marginal algorithm in arbitrary precision in terms of the asymptotic variance.

We consider also various sufficient conditions which guarantee certain rates of convergence of a pseudo-marginal algorithm in terms of the properties of the 'weights' $W_x = \hat{\pi}(x)/\pi(x)$ and the properties of the marginal algorithm. Particularly, if the weights $\{W_x\}$ are uniformly bounded and the marginal algorithm is geometrically ergodic, the pseudo-marginal algorithm is also geometrically ergodic. If the marginal algorithm is uniformly ergodic, we deduce sub-geometric convergence rates

whenever the weights $\{W_x\}$ are uniformly integrable. We also consider more subtle conditions for the rates of convergence for the independent Metropolis-Hastings and the random-walk Metropolis algorithms, where the estimators $\hat{\pi}(x)$ may worsen towards the tail of $\pi(x)$. Our results on convergence rates imply central limit theorems.

Acknowledgment. The work of C. Andrieu was partially supported by an EPSRC advanced research fellowship and a Winton Capital research award. M. Vihola was supported by the Academy of Finland (project 250575) and by the Finnish Academy of Science and Letters, Vilho, Yrjö and Kalle Väisälä Foundation.

CS19D Lim. Thms. Processes

Functional Limit Theorems for the Quadratic Variation of a Continuous Time Random Walk and for Certain Stochastic Integrals

ENRICO SCALAS*, NOÈLIA VILES†,‡

*Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale Amedeo Avogadro, Alessandria, Italy and BCAM-Basque Center for Applied Mathematics, Bilbao, Basque Country, Spain,

†Department of Probability, Logics and Statistics, Universitat de Barcelona, Barcelona, Spain

‡email: nviles@ub.edu 490:VILESNoelia.tex,session:CS19D

A continuos time random walk (CTRW) can be formally defined as a random walk subordinated to a counting renewal process. CTRWs became a widely used tool for describing random processes that appear in a large variety of physical models and also in finance.

The main motivation of our work comes from the physical model given by a damped harmonic oscillator subject to a random force (Lévy process) studied in the paper of Sokolov.

We study the convergence of a class of stochastic integrals with respect to the compound fractional Poisson process.

Under proper scaling and distributional hypotheses, we establish a functional limit theorem for the integral of a deterministic function driven by a time-changed symmetric α-stable Lévy process with respect to a properly rescaled continuous time random walk in the Skorokhod space equipped with the Skorokhod M_1-topology. The limiting process is the corresponding integral but with respect to a time-changed α-stable Lévy process where the time-change is given by the functional inverse of a β-stable subordinator.

Acknowledgment. This research was partially supported by the Italian grant PRIN 2009 "Finitary and non-finitary probabilistic methods in economics", Ref. 2009H8WPX5-002 and the Spanish grant MEC-FEDER Ref. MTM2009-08869 from the Dirección General de Investigación, MEC (Spain).

References

[Scalas and Viles (2012)] Scalas, E., Viles, N, 2012: A functional limit theorem for stochastic integrals driven by a time-changed symmetric α-stable Lévy process, *submitted*.

[Sokolov (2011)] Sokolov, I.M., 2011: Harmonic oscillator under Lévy noise: Unexpected properties in the phase space, *Phys. Rev. E. Stat. Nonlin Soft Matter Phys* , **83**, 041118.

CS7A Spatio-Temp. Stat I.

Kernel Estimation of Mean Densities of Random Closed Sets

FEDERICO CAMERLENGHI*, VINCENZO CAPASSO†, ELENA VILLA†,‡

*Dept. of Mathematics, Università degli Studi di Pavia, Italy,

†Dept. of Mathematics, Università degli Studi di Milano, Italy

‡email: elena.villa@unimi.it 491:ElenaVilla.tex,session:CS7A

Many real phenomena may be modelled as random closed sets in $\mathbb{R}^d$, of different Hausdorff di-

mensions. Of particular interest are cases in which their Hausdorff dimension, say n, is strictly less than d, such as fiber processes, boundaries of germ-grain models, and n-facets of random tessellations. A crucial problem is the estimation of pointwise mean densities of absolutely continuous, and spatially inhomogeneous random sets, as defined by the authors in a series of recent papers [Ambrosio et al. (2009), Villa (2012)]. The case $n = 0$ (random vectors, point processes, etc.) has been, and still is, the subject of extensive literature. In this talk we face the general case of any $n < d$; pointwise density estimators which extend the notion of kernel density estimators for random vectors are proposed. We study the unbiasedness and consistency properties, and identify optimal bandwidths for the proposed estimators, under sufficient regularity conditions. We show how some known results in literature follow as particular cases; a series of examples to illustrate various relevant situations are also provided.

References

[Ambrosio et al. (2009)] Ambrosio, L., Capasso, V., Villa, E., 2009: On the approximation of mean densities of random closed sets, *Bernoulli*, **15,** 1222 - 1242.

[Camerlenghi et al. (2012)] Camerlenghi, F., Capasso, V., Villa, E., 2012: On the estimation of the mean density of random closed sets, *Submitted*.

[Villa (2012)] Villa, E., 2012: On the local approximation of mean densities of random closed sets, *Bernoulli*. In press.

The Sound of Random Graphs

IS17 Random Matrices

BALINT VIRÁG*,†

*University of Toronto, Canada

†email: balint@math.toronto.edu 492:Virag.tex,session:IS17

The sound of random graphs and Cayley graphs

Infinite random graphs, such as Galton-Watson trees and percolation clusters, may have real numbers that are eigenvalues with probability one, providing a consistent "sound". These numbers correspond to atoms in their density-of-states measure.

In random matrix theory, these correspond to atoms in the analogue of Wigner semicircle law for the sparse Bernoulli setting.

When does the sound exist? When is the measure purely atomic? I will review many examples and show some elementary techniques developed in joint works with Charles Bordenave and Arnab Sen.

Option Price Decomposition under Stochastic Volatilty Models: Applications to Model Selection and Calibration

CS25A Stoch. Finance I.

JOSEP VIVES*,†

*Universitat de Barcelona

†email: josep.vives@ub.edu 493:vives-.tex,session:CS25A

In this talk we present two different decomposition methods for the price of a plain vanilla option under stochastic volatility models with or without jumps. These decompositions, or Hull and White type formulas, allow to distinguish between the impacts of correlation and jumps in the price. One of the methods is based on Malliavin calculus for Lévy processes and it is applied to a very

general stochastic volatility model with infinite activity jumps, both in the price and in the volatility. The use of Malliavin calculus allow to assume on the stochastic volatility, only minimal conditions, and markovianity is not necessary. As an application we obtain results related with the short time behaviour of the price of a plain vanilla option and of the implied volatility, that can be useful for model selection. The other method is based on Itô Calculus. As an example, it is applied to Heston model. We obtain the limit behaviour of the implied volatility surface, both in short and long time, and we develop a calibration model methodology. Comparison between the two methods is also given. The talk is a synthesis of recent papers [ALPV08], [ASV12] and [JV12].

Acknowledgment. This research was partially supported by the spanish grant MEC FEDER MTM 2009-07203.

References

[ALPV08] Alós, E., León, J. A., Pontier, M., Vives, J., 2008: A Hull and White formula for a general stochastic volatility jump-diffusion model and some applications, *JAMSA*, **Volume 2008**, Article ID 359142, 17 pages.

[ASV12] Alós, E., De Santiago, R., Vives, J. 2012: Calibration of stochastic volatility models via second order approximation: the Heston case, *Submitted*.

[JV12] Jafari, H., Vives, J., 2012: A Hull and White formula for a stochastic volatility Lévy model with infinite activity, *Submitted*.

CS26B Life and Failure Time

On Characterization of Generalized Mixtures of Weibull Distributions

NARAYANASWAMY BALAKRISHNAN*, MANUEL FRANCO†,§, DEBASIS KUNDU‡, JUANA-MARIA VIVO†

*McMaster University Hamilton, Ontario, Canada,
†University of Murcia, Murcia, Spain,
‡Indian Institute of Technology Kanpur, Kanpur, India
§email: mfranco@um.es 494:ManuelFranco.tex,session:CS26B

Mixtures of Weibull distributions play a great role in the reliability theory to model lifetime and failure time data, since they can incorporate wide varieties of failure rate functions. These mixtures forms have been generalized by allowing negative mixing weights, which arise under the formation of some structures of systems, and provide more versatile distributions for modelling dependent lifetimes from heterogeneous populations, see [Jiang et al. (1999)] among others.

Moreover, the generalized mixtures of Weibull distributions can be considered as extensions of the generalized mixtures of exponentials, which have been characterized in terms of the mixing weights and the parameters of its exponential components, and related results have been obtained by several authors, among others, see [Baggs and Nagaraja (1996)] and [Franco and Vivo (2006)]. In this setting, [Franco and Vivo (2009)] and [Franco et al. (2011)] have expanded the study of generalized mixtures of exponential components to the case of Weibull components with a common shape parameter.

In this work, we propose to study generalized mixtures of Weibull distributions with different shape parameters, by establishing its characterization in terms of the mixing weights and the parameters of its Weibull components to be a valid probability model.

Acknowledgment. This research was partially supported by the Fundación Séneca of the Regional Government of Murcia (Spain), grant No.: 11886/PHCS/09.

References

[Baggs and Nagaraja (1996)] Baggs, G.E., Nagaraja, H.N., 1996: Reliability properties of order statistics from bivariate exponential distributions, *Comm. Statist. Stoch. Models*, **12**, 611–631.

[Franco and Vivo (2006)] Franco, M., Vivo, J.M., 2006: On log-concavity of the extremes from Gumbel bivariate exponential distributions, *Statistics*, **40**, 415–433.

[Franco and Vivo (2009)] Franco, M., Vivo, J.M., 2009: Constraints for generalized mixtures of Weibull distributions with a common shape parameter,*Statist. Probab. Lett.*, **79**, 1724–1730.

[Franco et al. (2011)] Franco, M., Vivo, J.M., Balakrishnan, N., 2011: Reliability properties of generalized mixtures of Weibull distributions with a common shape parameter, *J. Statist. Plann. Infer.*, **141**, 2600–2613.

[Jiang et al. (1999)] Jiang, R., Zuo, M.J., Li, H.X., 1999: Weibull and inverse Weibull mixture models allowing negative weights, *Reliab. Engrg. Syst. Safety*, **66**, 227–234.

Some Recent Multivariate Extensions of the Exponential Model based on Optimization Procedures

CS26B Life and Failure Time

JUANA-MARIA VIVO*,†, MANUEL FRANCO*

*University of Murcia, Murcia, Spain
†email: jmvivomo@um.es 495:Vivo.tex,session:CS26B

Multivariate lifetime data frequently arise in many fields such as medicine, biology, public health, epidemiology, engineering, economic and demography. Undoubtedly, it is important to considerer different multivariate distributions that could be used to model such multivariate lifetime data, and their properties are also useful to carry out that purpose.

In the univariate case, exponential and Weibull distributions have been some of the most frequently applied statistical distributions in reliability and survival analysis, and multivariate extensions have been constructed from these models. However, the multidimensional extension of a distribution model is not unique, i.e., different derivations through the construction methods could be called multivariate extensions.

In this work, we focus on the construction of the multivariate models based on the method of the latent random factors which was used by Marshall and Olkin in 1967 to introduce a multivariate exponential distribution through the minimization process, among others, see [3], [2] and [1]. A brief review about this technique is shown along with some recent extensions and their properties.

Acknowledgment. This research was partially supported by the Fundación Séneca of the Regional Government of Murcia (Spain), grant No.: 11886/PHCS/09.

References

[1] Franco, M., Kundu, D., Vivo, J.M., 2011: Multivariate extension of the modified Sarhan-Balakrishnan bivariate distribution, *J. Statist. Plann. Infer.*, **141**, 3400–3412.

[2] Franco, M., Vivo, J.M., 2010: A multivariate extension of Sarhan and Balakrishnan bivariate distribution and its ageing and dependence properties, *J. Multivar. Anal.*, **101**, 491–499.

[3] Marshall, A.W., Olkin, I., 1967: A multivariate exponential distribution, *J. Amer. Statist. Assoc.*, **62**, 30–44.

D-Optimal Design for Nonlinear Models with Diffuse Prior Information

OCS9 Design of Experiments

TIM WAITE*,†

*Statistical Sciences Research Institute, University of Southampton, UK
†email: tww1g08@soton.ac.uk 496:Tim_Waite.tex,session:OCS9

In recent years there has been focus on applying more complex, often nonlinear, statistical models to the analysis of experimental data. Examples include generalized linear models in industrial

experiments, sometimes including random effects, and nonlinear compartmental models in pharmacokinetics. These more complex models have in common the property that design performance, for instance under D-optimality, may depend on the unknown values of the model parameters. Several methods have been proposed to derive designs which are reasonably efficient under a range of plausible values for the parameters, such as maximin and Bayesian approaches.

We consider the construction of designs in problems with a particular technical condition on the prior possibilities for the parameters. Specifically we consider the case where it is possible a priori for the parameters to be arbitrarily close to a (parameter) *singularity*, θ_s. Essentially, a singularity is a parameter vector such that any fixed design becomes uninformative as $\theta \to \theta_s$. From a design perspective, this represents a practically substantial amount of uncertainty since, for any fixed design, the worst-case efficiency over possible parameter values is equal to zero. As a consequence, maximin criteria will be degenerate. We show that these conditions can also be unfavourable to the most popular formulation of the Bayesian D-optimality criterion, which can therefore break down in a wider range of scenarios than has previously been explicitly acknowledged. We give an example of a prior distribution for the exponential model which has bounded support and for which all finitely-supported designs are singular with respect to Bayesian D-optimality. To overcome this problem, we consider (i) the use of alternative pseudo-Bayesian optimality criteria which are better behaved, and (ii) the use of designs with infinite support, defined through a continuous probability density function.

CS19D Lim. Thms. Processes

Limiting Spectral Distribution of a Symmetrized Auto-Cross Covariance Matrix

BAISUO JIN*, CHEN WANG†,¶, Z.D. BAI†,‡, K. KRISHNAN NAIR§, MATTHEW HARDING§

*University of Science and Technology of China, Heifei, China,
†National University of Singapore, Singapore,
‡Northeast Normal University, Changchun, China,
§Stanford University, USA
¶email: a0030755@nus.edu.sg

497:ChenWang.tex,session:CS19D

This paper studies the limiting spectral distribution (LSD) of a symmetrized auto-cross covariance matrix. The auto-cross covariance matrix is defined as $M_\tau = \frac{1}{2T}\sum_{k=1}^{T}(\mathbf{e}_k\mathbf{e}_{k+\tau}^* + \mathbf{e}_{k+\tau}\mathbf{e}_k^*)$, where τ is the lag and $\mathbf{e}_k = (\varepsilon_{1k}, \cdots, \varepsilon_{Nk})'$, $k = 1, 2, ..., T+\tau$, are N-vectors of independent standard complex components with $\sup_{1\le i\le N, 1\le t\le T+\tau} \mathrm{E}|\varepsilon_{it}|^{2+\delta} \le M < \infty$ for some $\delta \in (0, 2]$, and for any $\eta > 0$,

$$\frac{1}{\eta^{2+\delta}NT}\sum_{i=1}^{N}\sum_{t=1}^{T+\tau}\mathrm{E}(|\varepsilon_{it}|^{2+\delta}I(|\varepsilon_{it}| \ge \eta T^{1/(2+\delta)})) = o(1).$$

M_0 is well studied in the literature whose LSD is the Marčhenko-Pastur (MP) Law. The contribution of this paper is in determining the LSD of M_τ where $\tau \ge 1$. It should be noted that the LSD of the M_τ does not depend on τ. This study will facilitate the model selection of any large dimensional model with a lagged time series structure which are central to large dimensional factor models and singular spectrum analysis.

Acknowledgment. This research was supported by NSF China Young Scientist Grant 11101397, NSF China 11171057 as well as Stanford Presidential Fund for Innovation in International Studies.

References

[Bai and Silverstein (2010)] Bai, Z.D., Silverstein, J.W., 2010: Spectral Analysis of Large Dimensional Random Matrices, 2nd ed., *Springer Verlag, New York.*

Evaluating Recurrent Marker Processes with Competing Terminal Events

OCS21 Incomplete Longi. Data

MEI-CHENG WANG*,‡, KWUN-CHUEN G. CHAN†, YIFEI SUN*

*Johns Hopkins University, Baltimore, USA,
†University of Washington, Seattle, USA
‡email: mcwang@jhsph.edu

498:MeiChengWang.tex,session:OCS21

In follow-up or surveillance studies, marker measurements are frequently collected or observed conditioning on the occurrence of recurrent events. In many situations, marker measurement exists only when a recurrent event took place. Examples include medical cost for inpatient or outpatient cares, length-of-stay for hospitalizations, and prognostic measurements repeatedly measured at incidences of infection. A recurrent marker process, defined between an initiating event and a terminal event, is composed of recurrent events and repeatedly measured markers. This talk considers the situation when the occurrence of terminal event is subject to competing risks. Statistical methods and inference of recurrent marker process are developed to address a variety of questions/applications for the purposes of (i) estimating and comparing real-life utility measures, such as medical cost or length-of-stay in hospital, for different competing risk groups, and (ii) by counting time backward from terminal event, evaluating recurrent marker performance for different competing risk groups. A SEER-Medicare-linked database is used to illustrate the proposed approaches.

Truncated Offspring Distributions and Classification of Derived Coalescent Processes

IS28 Stoch. in Biol.

ALISON ETHERIDGE*, BJARKI ELDON†, SHI-DONG WANG*

*University of Oxford, UK,
†Technische Universität Berlin, Germany

499:ShidongWang.tex,session:IS28

We consider a model to obtain coalescent processes from supercritical Galton-Watson processes proposed by Schweinsberg (2003). Truncated offspring distributions and associated coalescent processes are of our interest. By a minimum truncation on each juveniles distribution, which is up to size of any finite integer for Kingman's coalescent, of order N for Beta coalescent, of order $N \log N$ for Bolthausen-Sznitman coalescent, and of order $N^{1/\gamma}$ $(0 < \gamma < 1)$ for Ξ-coalescent, we classify four kinds of (incomplete) coalescent processes on appropriate time scales. Furthermore, we combine them into an all-in-one offspring distribution, and give a complete classification of the coalescent limits depending on their skewed fraction of offspring distributions in terms of population size N. In the end, we study the rates of convergence to Beta coalescent, and show how small offspring size distribution alters the coalescent limit.

References

[1] J. Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch. Proc. Appl.*, 106:107–139, 2003.

OCS20 H-D Longitudinal Data

Multivariate t Linear Mixed Models for Multiple Repeated Measures with Missing Outcomes

WAN-LUN WANG[*,†]

[*]Department of Statistics, Feng Chia University, Taichung, Taiwan
[†]email: wlunwang@fcu.edu.tw 500:WanLunWang.tex,session:OCS20

The multivariate t linear mixed model (MtLMM; Wang and Fan 2011, 2012) has been shown a robust approach to modeling multi-outcome continuous repeated measures in the presence of outliers or heavy-tailed noises. Missing responses or irregularly timed multivariate longitudinal data frequently occur in clinical trials or biomedical studies. In the study, I present a framework for fitting the MtLMM with an arbitrary missing pattern across outcome variables and scheduled occasions. Motivated by several distinguishing features in longitudinal data analysis which differs from the classical time series analysis, a damped exponential correlation (DEC; Muñoz et al. 1992) structure is considered to address the serial correlation among the within-subject errors. Under the missing at random (MAR; Rubin 1976) mechanism, I develop an efficient alternating expectation-conditional maximization (AECM; Meng and van Dyk 1997) algorithm for carrying out maximum likelihood estimation of parameters and retrieving each missing outcome with a single-valued imputation. The techniques for the estimation of random effects and the prediction of further values given past observed responses are also investigated. The utility of the proposed methodology is illustrated through real and simulated examples.

Acknowledgment. This work was supported by the National Science Council under grant number NSC101-2118-M-035-003-MY2 of Taiwan.

References

[Meng and van Dyk (1997)] Meng, X.L. and van Dyk, D., 1997: The EM algorithm – an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59**, 511–567.

[Muñoz et al. (1992)] Muñoz, A., Carey, V., Schouten, J.P., Segal, M. and Rosner, B., 1992: A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **48**, 733–742.

[Rubin (1976)] Rubin, D.B., 1976: Inference and missing data. *Biometrika* **63**, 581–592.

[Wang and Fan (2010)] Wang, W.L. and Fan, T.H., 2010: ECM-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors. *Comput. Statist. Data Anal.* **54**, 1328–1341.

[Wang and Fan (2011)] Wang, W.L. and Fan, T.H., 2011: Estimation in multivariate t linear mixed models for multiple longitudinal data. *Statist. Sinica* **21**, 1857–1880.

OCS16 Interacting Particles

The Airy_1 Process for Brownian Motions Interacting through One-Sided Reflection

PATRIK FERRARI[*], HERBERT SPOHN[†], THOMAS WEISS[†,‡]

[*]Universität Bonn,
[†]Technische Universität München
[‡]email: thomas-weiss@mytum.de 501:ThomasWeiss.tex,session:OCS16

We consider brownian motions with one-sided reflection, which means that each particle is reflected at its right neighbour. For a finite number of particles there is a Schütz-type formula for the transition probability. We investigate an infinite system with periodic initial configuration, i.e. particles occupying the integer lattice at time zero. Via a description as a signed determinantal point process, it is shown that this diffusion converges to the Airy_1 process in the proper large time scaling limit.

Sequential Monte Carlo with Constrained Interaction

IS1 Bayesian Comp.

NICK WHITELEY*,†

*School of Mathematics, University of Bristol
†email: nick.whiteley@bristol.ac.uk 502:NickWhiteley.tex,session:IS1

The increasing availability of decentralized computational architectures raises many new questions about how algorithms should be designed and implemented. Several popular families of Monte Carlo algorithms operate by imposing and exploiting interaction between a collection of simulated processes. It is then natural to ask if this interaction can be relaxed in some sense, so as to more readily admit decentralized implementation, without completely destroying the attractive theoretical properties of the algorithms.

Motivated by such questions, this talk will report on an investigation of a class of Sequential Monte Carlo algorithms in which the interaction between particles occurs subject to constraints which may arise from a pre-determined parameterization of dependence, or may be constructed on-the-fly as the data are processed and the simulation progresses. Some convergence properties of these algorithms will be discussed, in particular, their variance-growth properties can be explained in terms of certain colliding Markov chains on graphs.

Solutions of Martingale Problems for Lévy-Type Operators and Stochastic Differential Equations Driven by Lévy Processes with Discontinuous Coefficients

CS19A Lim. Thms. Heavy Tails

PETER IMKELLER†, NIKLAS WILLRICH*,‡

*Weierstrass-Institut Berlin, Germany,
†Institut für Mathematik, Humboldt-Universität zu Berlin, Germany
‡email: willrich@wias-berlin.de 503:NiklasWillrich.tex,session:CS19A

We show the existence of Lévy-type stochastic processes in one space dimension with characteristic triplets that are either discontinuous at thresholds, or are stable-like with stability index functions for which the closures of the discontinuity sets are countable [IW12]. For this purpose, we formulate the problem in terms of a Skorokhod-space martingale problem associated with non-local operators with discontinuous coefficients. These operators are approximated along a sequence of smooth non-local operators giving rise to Feller processes with uniformly controlled symbols. They converge uniformly outside of increasingly smaller neighborhoods of a Lebesgue nullset on which the singularities of the limit operator are located. In our approach the characterization of Feller processes via generalizations of the Lévy-Khinchin formula plays a crucial role. They are based on the theory of pseudo-differential operators. In effect, they allow a representation of the non-local operators associated with Feller processes by a type of Fourier inversion formula in which the generalization of the characteristic exponent $q(\xi), \xi \in \mathbb{R}$, of the Lévy-Khinchin formula, the so-called *symbol* $q(x, \xi), \xi, x \in \mathbb{R}$, appears. It can be seen as a space (x)-dependent version of the exponent in the Fourier transform of the corresponding Feller process. Our method of construction of the Lévy-type processes solving at least the martingale problem for the associated operator A starts in a general framework. It is later shown to agree with both the scenario of Lévy triplets that have a discontinuity at a fixed threshold, as well as with the one of purely stable-like behavior with a discontinuous stability index function. In our main approximation result we approach A by a sequence of operators $(A_n)_{n\in\mathbb{N}}$ which possess smooth coefficient triplets, and corresponding symbols $(q_n)_{n\in\mathbb{N}}$. As an essential ingredient of our method, we have to control the sequence of symbols *uniformly in* n. This

is possible with the help of a recent result by Schilling and Wang [SW11]. In consequence we first obtain tightness of the unique solutions of the martingale problems $(\mathbb{P}_n)_{n\in\mathbb{N}}$ associated with $(A_n)_{n\in\mathbb{N}}$. The approximation further has to guarantee that the singularities of A are contained in $\cap_{m\in\mathbb{N}}U_m$, with a decreasing sequence $(U_m)_{m\in\mathbb{N}}$ of sets with asymptotically vanishing Lebesgue measure, and that the A_n converge to A compactly uniformly on the complements of the U_m. To deduce that a cluster point $\mathbb{P}$ of the sequence $(\mathbb{P}_n)_{n\in\mathbb{N}}$ solves the martingale problem as well, on U_m one uses once more the uniform bounds on the transition densities resulting from the control on the symbols. The construction of the appropriate sequences $(A_n)_{n\in\mathbb{N}}$ and $(U_m)_{m\in\mathbb{N}}$ in the two scenarios mentioned can be achieved. In the scenario of glueing together two Lévy processes techniques developed in Kurtz [K11] further reveal that the solution of the martingale problem also gives rise to a weak solution of the corresponding stochastic differential equation.

References

[K11] Kurtz, T.G., Equivalence of Stochastic Equations and Martingale Problems, Stochastic Analysis 2010. Dan Crisan, Ed. (2011), 113-130

[SW11] Schilling, R., Wang, J., Some Theorems on Feller Processes: Transience, Local Times and Ultracontractivity, To appear in: Trans. Am. Math. Soc. (2011)

[IW12] Imkeller, P., Willrich N., Solutions of martingale problems for Lévy-type operators and stochastic differential equations driven by Lévy processes with discontinuous coefficients, Preprint, arXiv: 1208.1665 (2012)

OCS8 Time Series

Weak Transport Inequalities and Applications to Exponential and Oracle Inequalities

OLIVIER WINTENBERGER*,†

*Université Paris Dauphine, CREST

†email: wintenberger@ceremade.dauphine.fr 504:Wintenberger.tex,session:OCS8

We extend the weak transport as defined in Marton (1996) [1] to other metrics than the Hamming distance. We obtain new weak transport inequalities for non product measures. Many examples are provided to show that the euclidian norm is an appropriate metric for many classical time series. The dual form of the weak transport inequalities yield new exponential inequalities and extensions to the dependent case of the classical result of Talagrand (1995) [2] for convex functions that are Lipschitz continuous. Expressing the concentration properties of the ordinary least square estimator as a conditional weak transport problem, we derive from the weak transport inequalities new oracle inequalities with fast rates of convergence. We also provide a new aggregation procedure when multiple models are considered.

References

[1] MARTON, K. (1996) A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6 (3)**, 556–571.

[2] TALAGRAND, M. (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, **81**, 73–205.

A Stopping Rule for Empirical Horvitz-Thompson Estimation with Application to Fixed-Cost Sampling

POSTER
Poster

WOJCIECH GAMROT*,†

*University of Economics, Katowice, Poland
†email: wojciech.gamrot@ue.katowice.pl 505:Gamrot.tex,session:POSTER

The knowledge of first-order inclusion probabilities characterizing a sampling scheme is essential in design-based estimation of finite population totals. However, sometimes the complexity of the sampling scheme prevents the exact computation of these probabilities. This is particularly true for various sequential spatial sampling designs where inclusion of a certain population unit in the sample rules out inclusion of neighboring units and for sequential fixed-cost schemes. The same problem also arises for variants of rejective sampling schemes. In such a situation a simulation study may be carried out in order to estimate unknown inclusion probabilities and corresponding sampling weights from the set of independent sample replications generated through the same sampling scheme. By replacing inclusion probabilities with estimates in the original Horvitz-Thompson statistic the empirical Horvitz-Thompson estimator is obtained. Often the knowledge of the sampling scheme facilitates analytical calculation of upper and lower bounds for individual inclusion probabilities. A natural way to incorporate this auxiliary knowledge into the estimation process is to use the restricted maximum likelihood principle. A question then arises: how to assess the number of sample replications needed to achieve a desired level of precision measured by variance or by the probability exceeding a prescribed margin of error. In this presentation such a problem is addressed. The distribution of the restricted maximum likelihood for the inclusion probability is examined and a method for computing the sample size controlling for the relative error of the population total estimator is proposed. An application to the fixed cost sequential sampling scheme is presented.

Acknowledgment. This research was supported by the grant No.: N N111 558540 from the Polish National Science Centre.

Spectrum Estimation in Large Dimensions

CS5D
H-D
Inference

OLIVIER LEDOIT*, MICHAEL WOLF*,†

*University of Zurich, Zurich, Switzerland
†email: michael.wolf@econ.uzh.ch 506:Wolf.tex,session:CS5D

We are interested in estimating the set of eigenvalues of a covariance matrix (also known as the *spectrum*) when the number of variables is of a magnitude similar to the sample size, or even larger. Our approach is shown to be consistent even in the absence of strong structural assumptions. Finite-sample performance is studied via Monte Carlo simulations. In addition, applications to covariance matrix estimation and principal component analysis (PCA) are given.

Acknowledgment. This research has been supported by the NCCR Finrisk project "New Methods in Theoretical and Empirical Asset Pricing".

OCS21 Incomplete Longi. Data

Dynamic Modeling of High-Dimensional Pseudo-Longitudinal Data for Gene Regulatory Networks

HULIN WU*,†

*The University of Rochester, Rochester, USA
†email: Hulin_Wu@urmc.rochester.edu 507:Hulin_Wu.tex,session:OCS21

Genome-wide time course gene expression data are collected to construct gene regulatory networks using microarray and RNA-Seq technologies in many biomedical studies. These high-dimensional time course data are often clustered into groups for the purposes of dimension reduction and functional module identification. Those genes in each group are considered as a supergene and the time course gene expression data in each group can be treated as "pseudo-longitudinal data," since the expression data for different genes are not independent compared to the independent data from different subjects in a standard longitudinal data set. We apply a nonparametric mixed-effects model with a mixture distribution to cluster the genome-wide gene expression data. Mixed-effects ordinary differential equations (ODE) are used to link "supergenes" into a regulatory network. The stochastic approximation EM (SAEM) algorithm is employed to refine the mixed-effects ODE model parameter estimates. In contrast, the stochastic differential equation (SDE) models can also be applied to the pseudo-longitudinal data. Comparisons between the two models will be discussed. The proposed models and methods are applied to several time course gene expression data for illustration.

Acknowledgment. This work is partially supported by the NIAID/NIH grants HHSN272201000055C and AI087-135 as well as two University of Rochester CTSI pilot awards (UL1RR024160) from the National Center For Research Resources.

CS30A Inf. on Distributions

Detection of Non-Gaussianity

TERESA LEDWINA*, GRZEGORZ WYŁUPEK†,‡

*Institute of Mathematics, Polish Academy of Sciences, Poland,
†Institute of Mathematics, University of Wrocław, Poland
‡email: wylupek@math.uni.wroc.pl 508:GrzegorzWylupek.tex,session:CS30A

Detection of non-Gaussianity, though being a classical problem, is still a subject of intensive research. This is motivated by wide applications and the existing needs to detect non-standard and subtle deviations from Gaussianity. For discussion and some evidence see for example Graham et al. (1995), Jin et al. (2005), Güner et al. (2009), Romão et al. (2010), Tarongi and Camps (2010) and references therein.

We develop two tests sensitive to various departures from composite goodness-of-fit hypothesis of normality. The tests are based on the sums of squares of some components naturally arising in decomposition of the Shapiro-Wilk type statistic. Each component itself has diagnostic properties. The numbers of squared components in sums are determined via some novel selection rules based on the data. The new solutions prove to be effective tools in detecting a broad spectrum of sources of non-Gaussianity. We also discuss two variants of the new tests adjusted to verification of simple goodness-of-fit hypothesis of normality.

References

[1] Graham, P., Turok, N., Lubin, P. M., and Schuster, J. A. (1995), A simple test for non-gaussianity in cosmic microwave background radiation measurements, *The Astrophysical Journal*, **449**, 404 - 412.

[2] Güner, B., Frankford, M. T., and Johnson, J. T. (2009), A study of the Shapiro-Wilk test for the detection of pulsed sinusoidal radio frequency interference, *IEEE Transactions on Geoscience and Remote Sensing*, **47**, 1745 - 1751.

[3] Jin, J., Starck, J. -L., Donoho, D. L., Aghanim, N., and Forni, O. (2005), Cosmological non-gaussian signature detection: comparing performance of different statistical tests, *EURASIP Journal on Applied Signal Processing*, **15**, 2470 - 2485.

[4] Romão, X., Delgado, R., and Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality, *Journal of Statistical Computation and Simulation*, **80**, 545 - 591.

[5] Tarongi, J. M., and Camps, A. (2010), Normality analysis for RFI detection in microwave radiometry, *Remote Sensing*, **2**, 191 - 210.

Matching Quantiles Estimation

OCS8 Time Series

NIKOLAOS SGOUROPOULOS*, QIWEI YAO†,‡, CLAUDIA YASTREMIZ*

*QA Exposure Analytics, Barclays Bank, London, UK,
†London School of Economics, London, UK
‡email: q.yao@lse.ac.uk 509:QIWEI_YAO.tex,session:OCS8

Motivated by a backtesting problem for counterparty credit risk management, we propose a new Matching Quantiles Estimation (MQE) method, for selecting representative portfolios. An iterative procedure based on the ordinary least squares estimation (LSE) is proposed to compute the MQE. The convergence of the algorithm and the asymptotic properties of the estimation are established. The finite sample properties are illustrated numerically by both simulation and a real data example on selecting a counterparty representative portfolio. The proposed MQE also finds applications in portfolio tracking, which demonstrates the potential usefulness of combing the MQE with LASSO.

On Estimation of the Number of Factors from High-Dimensional Data

OCS13 H-D Stat, R. Matrices

DAMIEN PASSEMIER*, JIANFENG YAO†,‡

*Université de Rennes 1, France,
†The University of Hong Kong, China
‡email: jeffyao@hku.hk 510:Yao.tex,session:OCS13

We study a spiked population model introduced in Johnstone (2001) where the population covariance matrix has all its eigenvalues equal to a constant value except for a few fixed eigenvalues (spikes). The classical factor models with random factors and homoscedastic errors is one particular instance of a spiked populaiton model. Determining the number of spikes or factors is a fundamental problem which appears in many scientific fields, including signal processing (linear mixture model) or economics (factor model). Several recent papers studied the asymptotic behavior of the eigenvalues of the sample covariance matrix when the dimension of the observations and the sample size both grow to infinity so that their ratio converges to a positive constant. Using these results, we propose a new estimator of the number of spikes (or factors) based on the difference between two consecutive sample eigenvalues.

Taking the view of hypothesis testing, we also present a test statistic to detect the presence of spikes. The test statistic is shown to have an asymptotic normal distribution.

References

[1] Z. D. Bai and J. F. Yao, Central limit theorems for eigenvalues in a spiked population model, *Ann. Inst. H. Poincar Probab. Stat.* **44(3)** (2008) 447–474.

[2] I. M. Johnstone, On the distribution of the largest eigenvalue in principal component analysis, Ann. Stat. **29** (2001) 295–327.

[3] A. Onatski, Testing hypotheses about the number of factors in large factors models, *to appear in Econometrica* (2008).

[4] S. Kritchman and B. Nadler, Determining the number of components in a factor model from limited noisy data, *Chem. Int. Lab. Syst.* **94** (2008) 19–32.

[5] D. Passemier and J. Yao, 2012. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrix: Theory and Applciations* **1**, 1150002

[6] Q. Wang, J. Silverstein and J. Yao, 2013. A note on the CLT for linear spectral statistics of sample covariance matrix from a spiked population model. *Preprint*

NYA Not Yet Arranged

Quantitative Analysis of the Pricing of Food Products in Kazakhstan

BAZHAN TUREBEKOVA*, ZHANAR YESZHANOVA*,†

*Eurasian National University named by L. Gumilev, Astana, Kazakhstan

†email: eszhan78@mail.ru 511:BazhanTurebekova.tex,session:NYA

The article is devoted to one of the most topical issues of the modern economy of Kazakhstan - the problem of an effective pricing policy for food products. Under market conditions, the state should not pursue a policy of pricing, but should regulate the food market, creating the economic conditions of agricultural production in the right quantities and proportions. Price is a means, and not subject to state regulation.

State compensates agricultural producers deviation from the market price in size needed for their activities at a certain level of profitability. Pricing policy in the agricultural sector should be market-based pricing, combined with a reasonable protectionism. The price policy has to assume, first of all, tracking dynamics of a number of economic indicators – costs of production, ensuring parity of the prices of means of production for agriculture and on agricultural production, and also profitabilities of farms, branches of production and all agriculture.

According to the economists, one of the main reasons of crisis position of domestic agro-industrial complex is the disparity of the prices on industrial and agricultural products. The prices of manufactured goods in 2000-2011 grew 4-5 times quicker, than on agroproduction. The village wasn't able to pay back costs of production. The greatest specific weight in material inputs of agricultural enterprises annually is the share of production of the industry including the fixed and revolving business assets.

Analysis of a sample survey of agricultural producers, presented in the paper reflects the price disparity, formed in the agricultural and industrial sectors of the economy.

Growth of retail prices for agricultural products is connected not only with rise of the prices of an industrial output, but also a considerable gain happens at a stage of realization of production in a chain from the producer to the end user that is connected with numerous intermediary links. The domestic retail market prices are higher than export prices that show the absence of a marketing and management of the vegetable market.

The contents of article briefly can be characterized as follows: first, the author gives the short description of a price situation in domestic market of food production, secondly is carried out the comparative analysis of growth rates of prices for products of agricultural producers and an industrial output, pricing process on food production in a chain from the producer to the end user is thirdly considered. In article the correlation and regression analysis of the change in price of food production from various factors that gave the chance of identification of most significant of them in the course of pricing on agricultural production is carried out.

The contents of article briefly can be characterized as follows: first, the author gives the short description of a price situation in domestic market of food production, secondly is carried out the comparative analysis of growth rates of prices for products of agricultural producers and an industrial output, pricing process on food production in a chain from the producer to the end user is thirdly considered. In article the correlation and regression analysis of the change in price of food production is carried out. The analysis gave the chance of identification the most significant of them in the course of pricing on agricultural production.

The conducted research gives the grounds to the author of article to make the conclusion that pricing process on domestic agricultural production – the unbalanced chaotic process which practically isn't supervised by authorities that leads to rise in price of many kinds of food. At the end of article the main ways of improvement of a pricing policy on food production that will allow providing its economic availability to the country population are considered.

References

[1] About level and dynamics of prices for products of agriculture and construction production in the Republic of Kazakhstan. Analytical report. Statistics agency of Kazakhstan. 2011 http://www.stat.gov.kz

[2] About the change in price for agriculture production in the Republic of Kazakhstan in 2012: Express information Statistics agency of Kazakhstan. 2012 http://www.stat.gov.kz

[3] Esenbayev M.(2012) Agrarian adaptation. Kazakhstan truth. 2012: 3-5

[4] Imangazhin Sh., Ismailov A. (2012) About the prices and food security of the country. Economy and statistics. 1 2012: 23-28

[5] Mukhtarova K.S. Kenzhebayeva A.R. (2010) Economic safety in the conditions of globalization. Almaty: Lawyer, 2010:150

[6] Kuchukov R. (2011) The state support of agriculture in the countries with the developed market economy. The Economist. 11 2011: 12-19

[7] Zinchenko A.P., Nazarenko V. I (2011) Agrarian policy / under the editorship of A.P.Zinchenko. M: Colossus, 2011: 304

A Functional Generalized Method of Moments Approach for Longitudinal Studies with Missing Responses and Covariate Measurement Error

OCS21 Incomplete Longi. Data

G. Y. YI[*,‡], Y. MA[†], R. J. CARROLL[†]

[*]University of Waterloo, Waterloo, N2L 3G1, Canada,
[†]Texas A&M University, College Station, Texas 77843-3143, U.S.A.
[‡]email: yyi@uwaterloo.ca 512:GraceYi.tex,session:OCS21

Covariate measurement error and missing responses are typical features in longitudinal data analysis. There has been extensive research on either covariate measurement error or missing responses, but relatively little work has been done to address both simultaneously. In this talk, I will discuss a simple method for the marginal analysis of longitudinal data with time-varying covariates, some of which are measured with error, while the response is subject to missingness. Both theoretical justification and numerical results will be presented.

IS25
Stat. SDE

Quasi Likelihood Analysis of Volatility and its Applications

NAKAHIRO YOSHIDA*,†

*Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, Japan
†email: nakahiro@ms.u-tokyo.ac.jp 513:NakahiroYoshida.tex,session:IS25

The **quasi likelihood analysis** (QLA) here means a systematic analysis of the quasi likelihood random field and the associated estimators, i.e., quasi maximum likelihood estimator and quasi Bayesian estimators, with a large deviation method that gives precise tail probability estimates for the random field and estimators. The QLA is necessary when one develops the basic fields in theoretical statistics such as asymptotic decision theory, prediction, information criteria, asymptotic expansion, higher-oder inference etc. for stochastic processes.

A polynomial type large deviation inequality was generally proved for the locally asymptotically quadratic quasi log likelihood random field under mild conditions that are easily verified for non-linear stochastic processes [1]. This scheme was applied to stochastic differential equations to form QLA under ergodicity ([1], [4], [2]).

The nondegeneracy of the statistical random field is crucial in estimation theory for the volatility parameter of a sampled semimartingale in finite time horizon. This is a non-ergodic setting where the degree of separation of statistical models is random even in the limit, and hence the nondegeneracy is random. An analytic criterion and a geometric criterion for the nondegeneracy were presented in [5]. Thus the QLA was established in this problem.

The **martingale expansion** plays an essential role to develop a higher order inference theory in non-ergodic statistics. The expansion formula is expressed by adjoint operation (to a certain functional involving the Donsker-Watanabe's delta functional) by certain random symbols, namely, the adaptive random symbol and the anticipative random symbol. The latter is written by the Malliavin calculus ([6]). An application was the asymptotic expansion of the realized volatility under finite time horizon. It can also be applied to parametric estimation of the volatility. Combining the martingale expansion with QLA, we can obtain asymptotic expansion of QLA estimators.

References

[1] Yoshida, N.: Polynomial type large deviation inequality and its applications. ISM Research Memorandum 1021

[2] Ogihara, T., Yoshida, N.: Quasi-likelihood analysis for the stochastic differential equation with jumps. Statistical Inference for Stochastic Processes 2011, 14, 3, 189-229

[3] Uchida, M., Yoshida, N.: Estimation for misspecified ergodic diffusion processes from discrete observations. ESAIM: Probability and Statistics, DOI: 10.1051/ps/2010001

[4] Uchida, M., Yoshida, N.: Adaptive estimation of an ergodic diffusion process based on sampled data. Stochastic Processes and their Applications, 122, 8, 2885-2924S

[5] Uchida, M., Yoshida, N.: Nondegeneracy of random field and estimation of diffusion. arXiv: 1212.5715.

[6] Yoshida, N.: Martingale expansion in mixed normal limit. arXiv: 1210.3680.

Double-Conditional Smoothing of High-Frequency Volatility Surface in a Spatial Multiplicative Component Garch with Random Effects

CS7A Spatio-Temp. Stat I.

YUANHUA FENG*,†

*Faculty of Business Administration and Economics, University of Paderborn
†email: yuanhua.feng@wiwi.upb.de

514:YuanhuaFeng.tex,session:CS7A

Let $r_{i,j}$ denote the high-frequency (log-)returns observed at the trading time t_j on the i-th trading day, where $i = 1, ..., n_x$ and, $j = 1, ..., n_t$. This paper introduces the following spatial framework

$$r_{i,j} = n_t^{-1/2} \sigma(x_i, t_j) Y_{i,j} \tag{1}$$

for high-frequency returns, where $x_i = (i - 0.5)/n_x$ is a re-scaled variable of the observation day, $\sigma(x, t) > 0$ is a slowly changing deterministic volatility surface as an entire figure of the long-term and intraday volatility dynamics and $Y_{i,j}$ is a stationary random field with zero mean and unit variance. Our main purpose is to estimate $\sigma^2(x, t)$ using bivariate kernel regression. An equivalent, much faster double-conditional smoothing is developed, where the data is first smoothed in one dimension, the smoothing results are then smoothed again in the other dimension, which also helps us to find that high-frequency returns exhibit multiplicative random effects. We hence propose to model $Y_{i,j}$ by introducing multiplicative random effects into the model of Engle and Sokalska (2012):

$$Y_{i,j} = \sqrt{\omega_i h_i \lambda_j q_{i,j}} \varepsilon_{i,j}, \tag{2}$$

where $\omega_i > 0$ and $\lambda_j > 0$ are i.i.d. random variables with unit means and finite variance standing for the random effects, $\varepsilon_{i,j}$ are i.i.d. random variables with zero mean, unit variance and $E(\varepsilon_{i,j}^4) < \infty$, h_i is a daily conditional variance component and $q_{i,j}$ are unit intraday GARCH components. It is assumed that these components are all independent of each other. The daily volatility component h_i may be governed by a separate stochastic process. Model (1) together with (2) extend the multiplicative component GARCH of Engle and Sokalska (2012) in different ways. Firstly, the the deterministic diurnal patter, $\sigma(x_i, t_j)$ for given i, is now allowed to change slowly over the observation period. And a long-term deterministic volatility trend, $\sigma(x_i, t_j)$ for given j, is introduced, which can also change slowly over the trading time points. Secondly, multiplicative random effects in both dimensions are introduced into the stochastic part. Following this ides, volatility in high-frequency returns is decomposed into different deterministic or stochastic components under the spatial framework. The proposed model is hence call a spatial multiplicative component GARCH with random effects.

Stationarity of $Y_{i,j}$ and $Y_{i,j}^2$, and properties of the sample variance of $Y_{i,j}$, $\hat{\sigma}^2(x, t)$ and the nonparametric estimators of $\sigma^2(x, t)$ in the first stage are all investigated in detail. It is found that multiplicative random effects affect the autocovariances of the squared process $Y_{i,j}^2$ and the variance of the sample variance very strongly. The asymptotic properties of the nonparametric estimators of the volatility surface are also changed correspondingly. In particular, the smoothing results in the first stage are now inconsistent and the final smoother converges much slower than in the case without random effects. Application to real data examples shows that the long-term volatility dynamics before, during and after the 2008 financial crisis at given time point form a *volatility arch* (volatility bridge) with a very sharp peak, which together with the daily *volatility smiles* build a *volatility saddle*.

References

[Engle and Sokalska (2012)] Engle, R.F. and Sokalska, M.E., 2012: Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *J. Fin. Econometr.*, **10**, 54-83.

OCS12 Experiment Design

D-optimal Designs for Multiresponse Linear Models with Qualitative Factors

RONG-XIAN YUE[*,§], XIN LIU[†], KASHINATH CHATTERJEE[‡]

[*]Department of Mathematics of Shanghai Normal University, Scientific Computing Key Laboratory of Shanghai Universities, and Division of Scientific Computation of E-Institute of Shanghai Universities, Shanghai 200234, China,
[†]College of Science. Donghua University, Shanghai, China,
[‡]Department of Statistics, Visva-Bharati University, Santiniketan, India
[§]email: yue2@shnu.edu.cn 515:Rong-Xian_Yue.tex,session:OCS12

Consider a linear regression model with both quantitative and qualitative factors and an m-dimensional response variable Y whose components are equicorrelated for each observation. The D-optimal design problem is investigated when the qualitative factors interact with the quantitative factors. It is shown that the determinant of the information matrix a product design can be separated into two parts corresponding to the two marginal designs. Especially, for the hierarchically ordered system of regression models the D-optimal design does not depend on the covariance matrix of Y.

Acknowledgment. This work was partially supported by NSFC grant (11071168, 11101077), Special Funds for Doctoral Authorities of Education Ministry (20103127110002), Innovation Program of Shanghai Municipal Education Commission (11zz116), E-Institutes of Shanghai Municipal Education Commission (E03004), Shanghai Leading Academic Discipline Project (S30405), the Fundamental Research Funds for the Central Universities.

OCS27 Infer. Censored Sample

Interval Estimation for some Life Distributions Based on Progressively Censored Sample

YUNUS AKDOĞAN[*,†], AHMET ÇALIK[*], ILKAY ALTINDAĞ[*], COŞKUN KUŞ[*], ISMAIL KINACI[*]

[*]Selcuk University, Konya, Turkey
[†]email: yakdogan@selcuk.edu.tr 516:YunusAkdogan.tex,session:OCS27

In this study, we obtained the exact confidence intervals and exact statistical tests for parameters of Weibull, Gompertz and Burr XII distributions based on progressively censored sample. There are some existing approaches for obtaining exact confidence intervals and statistical test for these distributions. The second discussion of this presentation is to compare the new and existing results in terms of coverage probability, lenght of intervals and power of tests through simulation study. A numerical example is also provided for illustration.

Acknowledgment. This research was partially supported by Selcuk University BAP Office.

OCS18 Lifetime Data Anal.

The Coxian Phase Type Distribution in Survival Analysis

ADELE H. MARSHALL[*], MARIANGELA ZENGA[†,‡]

[*]Centre for Statistical Sciences and Operational Research, David Bates Building, Queen's University Belfast, Belfast, Northern Ireland, U.K,
[†]Department of Statistics and Probability, Milano Bicocca University, Milan, Italy
[‡]email: mariangela.zenga@unimib.it 517:MariangelaZenga.tex,session:OCS18

Coxian phase-type distribution (CPT) (Cox, 1955) is a special type of Markov model that can be used to represent duration of time as a stochastic process consisting of phases through which ele-

ments in the model progress until they leave the system completely at any stage into a final absorbing phase. Some researches have shown as Coxian phase-type distributions could well represent survival times as the length of time until a certain event occurs, where the phases are considered to be stages in the survival and the absorbing, final stage, the event that occurs causing the individual or element to leave the system completely. For instance, this event could be a patient recovering from an illness, a patient having a relapse, an individual leaving a certain type of employment, a piece of equipment failing, or a patient dying. Faddy (1994) illustrates how useful the Coxian phase-type distributions are in representing survival times for various applications such as the length of treatment spell of control patients in a suicide study, the time prisoners spend on remand and the lifetime of rats used as controls in a study of ageing. Faddy and McClean (1999) used the Coxian phase-type distribution to find a suitable distribution for modelling the duration of stay of a group of male geriatric patients in hospital. They found that the phase-type distributions were ideal for measuring the lengths of stay of patients in hospital and showed how it was also possible to consider other variables that may influence the duration. More recently, Marshall and McClean (2003) have demonstrated how the Coxian phase-type distribution can, unlike alternative approaches, adequately model the survival of various groups of elderly patients in hospital uniquely capturing the typical skewed nature of such survival data in the form of a conditional phase-type model which incorporates a Bayesian network of inter-related variables.

In this work we will introduce the CPT distribution, with regards to the problem of the parameters estimation and we will show several applications in the field of the survival analysis.

References

[Cox, 1955] Cox , D.R., 1955: A use of complex probabilities in the theory of stochastic processes. *Cambridge Philosophical Society* **51**, 313–319.

[Faddy, 1994] Faddy, M., 1994: Examples of fitting structured phase-type distributions. *Applied Stochastic Models and Data Analysis*, **10**, 247–255.

[Faddy, Mcclean, 1999] Faddy, M., McClean S.I., 1999: Analysing data on lengths of stay of hospital patients using phase-type distribution. *Applied Stochastic Models in Business and Industry*, **14**(4), 311–317.

[Marshall, Mcclean, 2003] Marshall A.H., McCleanS.I., 2003: Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research* **10**, 565–576.

Portfolio Investment Based on Mixture Experiments Designs

OCS12 Experiment Design

CHONGQI ZHANG*,†

*Department of Probability and Statistics, Guangzhou University, Guangzhou, China
†email: cqzhang@gzhu.edu.cn 518:Chongqi_Zhang.tex,session:OCS12

Experiments with mixtures are special types of experiments in which the response depends only on the proportions of input variables $x = (x_1; \ldots; x_q)$, not on the total amount of the ingredients. Mixture experimental designs have been widely used in industry, agriculture and science test. This paper mainly applies mixture experimental designs theory and method to the portfolio investment, which is a new perspective to research on the portfolio investment. We analyze portfolio investment at homoscedastic and heteroscedastic mixture model and give the explanation of portfolio risk and portfolio prot from the mixture designs theory. According to the theory analysis, the maximum portfolio risk will be minimum at the G-optimal experiment design, and we can calculate the investment proportional coefficient when portfolio risk minimum through some examples.

IS13 Model Selection

Statistical Inference with High-Dimensional Data

CUN-HUI ZHANG*,†

*Rutgers University

†email: czhang@stat.rutgers.edu 519:CunHuiZhang.tex,session:IS13

We propose a semi low-dimensional (LD) approach for statistical analysis of certain types of high-dimensional (HD) data. The proposed approach is best described with the following model statement:

model = LD component + HD component.

We develop statistical inference procedures for the LD component, including efficient estimators, p-values and confidence regions. Model selection problems can be then treated as multiple testing problems based on efficient low-dimensional procedures. A number of specific problems will be considered, including linear regression, partial correlation and graphical model selection, tests of group effects, generalized linear models, and proportional hazards regression.

OCS13 H-D Stat, R. Matrices

A Central Limit Theorem for Linear Spectral Statistics of Large Dimensional General Fisher-Matrices

SHURONG ZHENG*,‡, ZHIDONG BAI*, JIANFENG YAO†

*School of Mathematics and Statistics, Northeast Normal University, Changchun, P. R. China,

†Department of Statistics and Actuarial Science, Hong Kong University, P. R. China

‡email: zhengsr@nenu.edu.cn 520:ShurongZheng.tex,session:OCS13

It becomes obvious that central limit theorems (CLT) for linear spectral statistics of F matrix from large dimensional random matrices constitute a key tool in multivariate and high-dimensional data analysis. They have received considerable attention in the literature in recent years. Besides high-dimensional data analysis, their applications occur in various disciplines such as number theory, mathematical finance, economics and wireless communication networks In a milestone work, Bai and Silverstein (2004) establishes a CLT for LSS of general sample covariance matrices of form $\mathbf{S}_1\mathbf{T}_p$, where $\mathbf{S}_1$ is a sample covariance matrix from a population with i.i.d. (independently and identically distributed) components and $\mathbf{T}_p$'s are nonnegative definite non-random Hermitian matrices. For a Fisher matrix of simpler form, $\mathbf{F} = \mathbf{S}_1\mathbf{S}_2^{-1}$ where both populations have i.i.d. components, so in particular $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, the CLT of Bai and Silverstein (2004) may be applied with $\mathbf{T}_p = \mathbf{S}_2^{-1}$. However, the centring term in the CLT would then depend on a random distribution which is a function of the eigenvalues of $\mathbf{S}_2^{-1}$ and this function is complex and non explicit. This fact makes such a CLT useless for concrete data analysis. To overcome this difficulty, Zheng (20120 establishes a CLT for LSS of the Fisher matrix $\mathbf{F} = \mathbf{S}_1\mathbf{S}_2^{-1}$ where the centring terms are defined as a function of a limiting spectral distribution, thus is non-random and can be calculated explicitly. However, this CLT requires the equality $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. This means that for the two sample Fisher test above, the CLT in Zheng (2012) helps us to find the null distribution but not the power function where $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ under the alternative hypothesis. More detailed discussions on this test can be found in Bai et al. (2009).

The main purpose of the paper is to establish a CLT for LSS of the general Fisher matrix $\mathbf{F}$, i.e. the population covariance matrices $\mathbf{\Sigma}_j$ are not necessarily equal. Besides the discussed two-sample test, this new CLT can also be applied to one-sample test of the hypothesis that a population covariance matrix has a given structure, e.g. diagonal, band matrix. Such situations are again not covered by the CLT in Zheng (2012).

Acknowledgment. This research was partially supported by NECT-11-0616, NSFC 11171058 and NSFC 11071035.

References

[Bai et al.(2009)] Bai, Z., Jiang, D., Yao, J., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37, 3822–3840.

[Bai and Silverstein (2004)] Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, **32**(1A), 553-605.

[Zheng (2012)] Zheng, S. R. (2012). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F-Matrix. *Annales de l'Institut Henri Poincare-Probabiliteset Statistiques*, **48**(2), 444-476.

Uniform Confidence Bands in Local Estimation

CS6A Funct. Est., Kernel Meth.

VLADIMIR SPOKOINY[*,†,‡], MAYYA ZHILOVA[*,‡,§]

[*]Weierstrass Institute, Berlin,
[†]Humboldt University Berlin,
[‡]Moscow Institute of Physics and Technology
[§]email: zhilova@wias-berlin.de

521:MayyaZhilova.tex,session:CS6A

Uniform confidence bands based on the local maximum likelihood estimate for the generalized regression model are being constructed. Local estimation allows weakening of the parametric assumption on the model, but demands the proper choice of the degree of localization, characterized by bandwidth value. Too large bandwidth leads to a bad approximation, whereas too small bandwidth – to a bigger variance. Choice of the optimal bandwidth leading to the bias-variance tradeoff is one of the central problems in nonparametric statistics. This motivates construction of the uniform in bandwidth confidence sets which show how the band's size changes for different bandwidths with the same level of confidence. The main questions here are how to attain the uniformity and how much it does increase the band's size. In the presented work we address the both questions and obtain the uniform in bandwidth confidence bands in non-asymptotic form.

Acknowledgment. The authors are supported by Predictive Modeling Laboratory, MIPT, RF government grant, ag. 11.G34.31.0073. Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged.

Zero-One k-Law for Large Denominator

CS28A Random Graphs

MAKSIM ZHUKOVSKII[*,†,‡]

[*]Moscow State University, Moscow, Russia,
[†]MIPT, Moscow, Russia
[‡]email: zhukmax@gmail.com

522:MaksimZhukovskii.tex,session:CS28A

We study asymptotical behaviour of the probabilities of first-order properties for Erdős–Rényi random graphs $G(N, p)$.

It was proved in 1969 by Y.V. Glebskii, D.I. Kogan, M.I. Liagonkii and V.A. Talanov (and independently in 1976 by R.Fagin) that for any first order property L either "almost all" graphs satisfy this property as N tends to infinity or "almost all" graphs don't satisfy the property. In other words, if p doesn't depend on N, then for any first-order property L either the random graph satisfies the property L almost surely or it doesn't satisfy (in such cases the random graph is said to *obey zero-one law*. Moreover, the following statement holds. In this work we consider the probabilities $p = p(N)$, where $p(N) = N^{-\alpha}$, $N \in \mathbb{N}$, for $\alpha \in (0, 1)$. The zero-one law for such probabilities was proved by S. Shelah and J.H. Spencer (see [1]). They proved that for irrational α from $(0, 1)$ the random graph

$G(N, N^{-\alpha})$ obeys the zero-one law. When $\alpha \in (0, 1)$ is rational the zero-one law in ordinary sense for the graph doesn't hold.

Let k be a positive integer. Denote by $\mathcal{L}_k$ the class of the first-order properties of graphs defined by formulae with quantifier depth bounded by the number k (the sentences are of a finite length). Let us say that the random graph obeys *zero-one k-law*, if for any first-order property $L \in \mathcal{L}_k$ either the random graph satisfies the property L almost surely or it doesn't satisfy. In 2012 we prove [2] that $G(N, N^{-\alpha})$, $\alpha \in (0, 1/(k-2))$, obeys zero-one k-law. Recently we extend this law and prove that zero-one k-law holds if the denominator of rational number α is rather large. We also obtain $n_0 = n_0(k)$ such that if $\alpha = \frac{m}{n} \in (0, 1)$ is an irreducible fraction, $n > n_0$ then zero-one k-law holds.

References

[1] Shelah, S., Spencer, J.H., 1988: Zero-one laws for sparse random graphs, *J. Amer. Math. Soc.*, **1**: 97 - 115.

[2] Zhukovskii, M.E., 2012: Zero-One k-Law, *Discrete Mathematics*, **312**: 1670 - 1688.

IS21 Spatial Point Proc.

Stable Point Processes: A Model for Bursty Spatial Data

SERGEI ZUYEV[†]

*Department of Mathematical Sciences Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden.

[†]email: sergei.zuyev@chalmers.se 523:ZuyevSergei.tex,session:IS21

A recent challenge in modelling contemporary complex systems is to take into account an often observed *burstiness* when the structures at different even close locations may differ drastically. Modelling temporal bursty phenomena, such as the internet traffic, gave rise to development of models based on fractional Brownian motion and Lévy processes, while spatial burstiness has hardly been addresses so far. To take into account extreme spatial variability, we study thinning-stable point processes. They can be considered as a generalisation of discrete-stable integer random variables, this is why they are also called *discrete alpha-stable point processes*, or DaS. DaS processes arise as a limit in superposition-thinning of i.i.d. point processes. When the intensity measure of the thinned point processes exist, the limit is a Poisson process. However, in the case when intensity measures of the summands assume infinite values, the limit is DaS which manifests in a bursty structure. By using recent results on the cluster representation of DaS processes [1] we develop statistical tools to estimate their parameters: the exponent alpha, intensity of cluster centres and the spectral measure governing the distribution of the underlying Sibuya point process.

References

[1] Yu. Davydov, I. Molchanov and S. Zuyev. Stability for random measures, point processes and discrete semigroups, *Bernoulli*, **17**(3), 1015-1043, 2011

Author index

Session index