THIRD CONFERENCE

OF THE **INTERNATIONAL SOCIETY**

FOR **NONPARAMETRIC STATISTICS** (ISNPS)

AVIGNON, PALACE OF THE POPES, 11-16 JUNE 2016

BOOK OF ABSTRACTS

1 Plenary and special invited speakers

1.1 The auto- and cross-distance correlation functions of a multivariate time series and their sample versions, *T. Mikosch, Univ. Copenhagen, Dk*, Sunday 9h

Abstract: Székely, Rizzo and Bakirov (2007) introduced the notion of distance covariance/correlation as a measure of independence/dependence between two vectors of arbitrary dimension and provided limit theory for the sample versions based on an iid sequence. The main idea is to use characteristic functions to test for independence between vectors, using the standard property that the characteristic function of two independent vectors factorizes. Distance covariance is a weighted version of the squared distance between the joint characteristic function of the vectors and the product of their marginal characteristic functions. Similar ideas have been used in the literature for various purposes: goodnes-of-fit tests, change point detection, testing for independence of variables,... ; see work by Meintanis, Huškova, and many others. In contrast to Székely et al. who use a weight function which is infinite on the axes, the latter authors choose probability density weights. Z. Zhou (2012) extended distance correlation to time series models for testing dependence/independence in a time series at a given lag. He assumed a "physical dependence measure". In our work we consider the distance covariance/correlation for general weight measures, finite or infinite on the axes or at the origin. These include the choice of Székely et al., probability and various Lévy measures. The sample versions of distance covariance/correlation are obtained by replacing the characteristic functions by their sample versions. We show consistency under ergodicity and weak convergence to an unfamiliar limit distribution of the scaled auto- and cross-distance covariance/correlation functions under strong mixing. We also study the auto-distance correlation function of the residual process of an autoregressive process. The limit theory is distinct from the corresponding theory of an iid noise process. We illustrate the theory for simulated and real data examples.
This is joint work with R.A. Davis, P. Wan (Columbia Statistics), and M. Matsui (Nagoya).

1.2 Permutation p-value approximation via generalized Stolarsky invariance, *A. Owen, Stanford University, USA* Tuesday 8h30

Abstract: It is extremely expensive to get a tiny p value via permutations. For linear test statistics, the permutation p value is the fraction of permuted data vectors in a given spherical cap. Stolarsky's invariance principal gives the root mean squared discrepancy between the observed and expected proportions of points in spherical caps. We specialize this formula to spherical caps of exactly the same size as the empirical p value, by applying a recent result of Brauchart and Dick. Then we extend it further by conditioning on more aspects of the data. The result is an approximate p value that is competitive with saddlepoint approximations and is accompanied by a finite sample error estimate. http://arxiv.org/abs/1603.02757
This is joint work with Hera He, Kinjal Basu and QIngyuan Zhao.

1.3 Neural Random Forests, *G. Biau, Univ. Paris VI, Fr,* Tuesday 9h30

Abstract: Decision tree learning is a popular data-modeling technique that has been around for over fifty years in the fields of statistics, artificial intelligence, and machine learning. The approach and its innumerable variants have been successfully involved in many challenges requiring classification and regression tasks, and it is no exaggeration to say that many modern predictive algorithms rely directly or indirectly on tree principles. The history of trees goes on today with random forests (Breiman, 2001), which are on the list of the most successful machine learning algorithms currently available to handle large-scale and high-dimensional data sets. It is sometimes alluded to that forests have the flavor of deep network architectures, insofar as ensemble of trees allow to discriminate between a very large number of regions. The richness of forest partitioning results from the fact that the number of intersections of the leaf regions can be exponential in the number of trees. That being said, the connection between random forests and neural networks is largely unexamined. My presentation will be divided into two parts. First, I will review some of the most recent theoretical and methodological developments for random forests, with a special emphasis on the mathematical forces driving the algorithm. Next, I will reformulate the random forest method into a neural network setting, and in turn propose two new hybrid procedures that we call « neural random forests ». In a nutshell, given an ensemble of random trees, it is possible to restructure it as a collection of (random) multilayered neural networks, which have sparse connections and are therefore easier to train. Also, while the original trees have constraints on the orientation of the decision boundaries, the companion networks have no such limitation. Their activation functions are soft nonlinear and thus expected to exhibit better generalization performance.
This is a joint work with E. Scornet and J. Welbl.

1.4 Statistical Fusion Learning: combining inferences from multiple sources for more powerful findings, *R. Liu, Rutgers University, USA*, Tuesday 9h30

Abstract: Inferences from multiple databases or studies can often be fused together to yield a more powerful overall inference than individual studies alone. Fusion learning refers to the development of effective approaches for synergizing learnings from different data sources. Effective fusion learning is of vital importance, especially in light of the ubiquitous information and data collection nowadays. Decision-making processes in many domains such as medicine, life science, social studies, etc. often benefit greatly from considering data from different sources, possibly with varying forms of complexity and heterogeneity in their data structure.
This talk presents some new fusion methodologies for extracting and merging useful information. Some methodologies are motivated by challenges arising from massive complex structures from different data sources, while others by specific goal-directed applications, such as in precision medicine. Underlying those methodologies is the tool of "confidence distribution" (CD), which, simply put, is a versatile distributional inferential scheme (unlike the usual point or interval inferences) without priors. Some simulation and real applications are also presented.
This is joint work with Dungang Liu, University of Cincinnati, USA, and Jieli Shen and Minge Xie, Rutgers University, USA.

References

[1] Ehrenberg, A. C. S. (1982). Writing technical reports and papers. *The American Statistician*, **36**, 326–329.

[2] László Györfi and Michael Kohler and Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer. New-York.

[3] Lamport, L. (1986). *LaTeX A Document Preparation System.* Addison-Wesley. Boston.

1.5 The bootstrap in some novel environments, *P. Bickel, Univ. of California, Berkeley, USA* Wednesday 11h

Abstract: Peter Hall,in his tragically abridged life, wrote extensively on Efron's bootstrap and other Monte Carlo methods, culminating in his book,"The Bootstrap and Edgeworth Expansion" (1992). Although these are not "second order" situations of the type he focused on, I hope he would have enjoyed my review of the following work on the bootstrap in a number of high dimensional situations.

1. The Genome structural correction: Zhang, Huang, Boley, Brown and B. Ann.App.Stat.(2010)
2. Can we trust the bootstrap in high dimension? el Karoui and Purdom(2016)
3. Residual bootstrap for high dimensional regression with near low rank designs: Lopes NIPS(2014)
4. Subsampling bootstrap of count features of networks: Bhattacharyya and B. Ann.Stat.(2015))

1.6 Concentration in learning, *S. van de Geer, ETH Zürich, CH* Wednesday 16h45

Abstract: The title of this talk is inspired by the paper "Learning without concentration" Mendelson [2015]. The paper of Mendelson shows that one can prove good results for empirical risk minimizers without going over the sometimes restrictive route of concentration inequalities. In contrast, in this work (van de Geer and Wainwright [2016]) concentration is not only a tool but actually a result. It focusses on the concentration of the empirical risk minimizer itself. We consider an empirical risk function $\hat{R}_n$ on some space of functions $\mathcal{F}$ and a regularization penalty $\text{pen} : \mathcal{F} \to [0, \infty)$. The regularized ERM is

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \Big\{ \hat{R}_n(f) + \text{pen}(f) \Big\}.$$

Let $R(f) := \mathbb{E}\hat{R}_n(f)$, $f \in \mathcal{F}$ and let $f^0 := \arg\min_{f \in \mathcal{F}} R(f)$. The excess risk is $\mathcal{E}(f) := R(f) - R(f^0)$. The question we address is: when does $\mathcal{E}(\hat{f}_n)$ concentrate on a single point? More precisely and in an asymptotic setting, we say that $\mathcal{E}(\hat{f}_n)$ concentrates if for some positive sequence ρ_n converging to zero, the renormalized excess risk $\mathcal{E}(\hat{f}_n)/\rho_n$ converges in probability to a constant $c > 0$ (say $c = 1$). We present results for the Gaussian regression model, with convex $\mathcal{F}$ and penalty extending some findings of Chatterjee [2014], and also for some other regression and density estimation problems.
Joint work with Martin Wainwright, Dpt Statistics and Dpt of EECS, Univ. of California, Berkeley

References

[1] S. Chatterjee (2014). A new perspective on least squares under convex constraint. *Annals of Statistics*, 2340-2381.

[2] S. Mendelson (2015). Learning without concentration. *Journal of the ACM*, 3-21.

[3] S. van de Geer and M. Wainwright (2015). On concentration for (regularized) empirical risk minimisation (arXiv:1511.08698)

1.7 Geometric Feature Extraction, *W. Polonik, University of California, Davis, USA* Wednesday 17h45

Abstract: Extracting information about geometric features of an underlying distribution from a point cloud is of broad interest. Related methodologies address inference for level sets/curves, depth curves, ridgelines, integral curves, Morse complexes and modes. It is interesting to note that the objects in this list contain information on different scales, some are local and others are global by nature. In this talk we will first review some aspects of more recent work on inference for such objects and discuss several interrelations. Then we present a novel idea for extracting geometric information on different scales simultaneously. This multiscale methodology in particular also allows the visualization of certain aspects of the shape of a high-dimensional distribution. Applications to classification are discussed and some supporting theory is provided.

1.8 Cointegration: Bootstrap-based inference on rank and cointegration parameters, *A. Rahbek, University of Copenhagen, Dk* Wednesday 17h45

Abstract: We consider testing cointegration rank and hypothesis testing on cointegration parameters based on bootstrap. It is shown that the proposed bootstrap for rank testing and for general hypothesis testing lead to asymptotic consistency in cointegrated models with iid innovations. But also in cointegrated models with with (un)conditionally heteroscedastic innovations. For finite samples, it is demonstrated that empirical rejection probabilities are highly promising in terms of size and power and outperforms in particular standard asymptotic inference.

1.9 High Dimensional Learning and Deep Neural Networks, *S. Mallat, Ecole Normale Supérieure, Fr* Thursday 14h15

Abstract: Classification and regression require to approximate functions in high dimensional spaces. Avoiding the dimensionality curse opens many questions in statistics, probability, harmonic analysis and geometry. Recently, convolutional deep neural networks have obtained spectacular results for image analysis, speech understanding, natural languages and many other problems. We describe their architecture and analyze their mathematical properties, with many open questions. These architectures implement multiscale contractions, where wavelets have an important role. Applications will be shown for image and audio classification, and for regressions of molecular energies in quantum chemistry.

2 Invited sessions

2.1 Explicit solutions for the asymptotically-optimal bandwidth in cross validation

Karim M. Abadir[1,*] and Michel Lubrano[2]

[1] Imperial College London; k.m.abadir@imperial.ac.uk
[2] Aix-Marseille University and GREQAM-CNRS & EHESS; michel.lubrano@univ-amu.fr

Abstract: Least squares cross-validation (CV) methods are often used for automated bandwidth selection in density estimation. We show that they share a common structure which has an explicit asymptotic solution when the chosen kernel is asymptotically separable in the bandwidth h. Using the framework of density estimation, we consider unbiased, biased, and smoothed CV methods. For the Epanechnikov and Student $\mathrm{t}(\nu)$ kernels, the CV criterion becomes asymptotically equivalent to a simple polynomial. This leads to optimal-bandwidth solutions that dominate the usual CV methods, definitely in terms of simplicity of estimation and speed of calculation, but also often in terms of integrated squared error because of the robustness of our asymptotic solution, hence also alleviating the notorious sample variability of CV and its breakdown in the case of repeated observations. For the case of a $\mathrm{t}(\nu)$ kernel, we also provide a data-driven way of choosing ν, and simulations show that it outperforms the use of the Epanechnikov in finite samples. An empirical application involving the large database of Michigan State University yearly academic wages is provided.

Keywords: bandwidth choice; cross validation; explicit analytical solution; nonparametric density estimation, academic wage distribution.

2.2 Solution of linear inverse problems using exponential weights

F. Abramovich[1] and M. Pensky[2,*]

[1] Tel-Aviv University; felix@post.tau.ac.il;
[2] University of Central Florida; marianna.pensky@ucf.edu

Abstract: We consider solution of a general statistical linear inverse problem where solution is represented via a known (possibly overcomplete) dictionary with a sparse coefficient vector. We use exponential weights in both the model selection and the aggregation settings. We show that in both cases, under mild conditions on the dictionary, the estimator of the solution obeys oracle inequalities.

Keywords: Linear inverse problem; Model selection; Aggregation; Exponential weights.

2.3 Semiparametric Inference with Spatially Correlated Recurrent Event data

Dabo and Adekpedjou

Univ. Lille 3
Missouri Univ. of Science and Technology, US

Abstract: Consider n geographical regions that are monitored for the occurrence of a recurrent event. Further assume that various environmental factors, common to the regions trigger the recurrence of the event of interest leading to spatially correlated recurrent event data. In this talk, we first propose a model for the spatial correlation structure. Techniques for estimating the regression coefficients in a Cox-type model as well as the parameters of the correlation structure are presented. The estimators obtained have population interpretation and help identify risk factors that contribute to recurrence. The estimation procedures are facilitated by transforming the original gap-times yielding a multivariate Gaussian random field where the marginal Cox-type models for the original gap-times is preserved, and parameters are estimated via quasi-likelihood. Asymptotic properties of the estimators will be discussed. Results of a simulation study will be presented. The methods are illustrated with a real spatially correlated recurrent event data.

2.4 Local depth for functional data analysis

C. Agostinelli

Department of Mathematics, University of Trento, Italy; claudio.agostinelli@unitn.it

Abstract: Data depth proves successful in the analysis of multivariate data sets, in particular deriving an overall center and assigning ranks to the observed units. Two key features are: the directions of the ordering, from the center towards the outside, and the recognition of a unique center irrespective of the distribution being unimodal or multimodal. This behaviour is a consequence of the monotonicity of the ranks that decrease along any ray from the deepest point. Recently, a wider framework allowing identification of partial centers was suggested in [**?**]. The corresponding generalized depth functions, called *local depth functions* are able to record local fluctuations and can be used in mode detection, identification of components in mixture models and in cluster analysis. Functional data [**?**] are become common nowadays. Recently, [**?**] has proposed the half-region depth suited for functional data and for high dimensional data. A local half-region depth is introduced and its properties discussed. Several examples will illustrate the use of this new tool in data analysis.

Keywords: Clustering; Functional Data; Half-region Depth; Local Depth; Time Series

2.5 Least Squares Estimation of the Central Mean Subspace

M. Akritas[1,*] and J. Lin[1]

[1] Penn State University; mga@stat.psu.edu, jul268@psu.edu

Abstract: [1] introduced the single index model (SIM) and showed that the ordinary least squares (OLS) estimator captures the direction of the parametric component of the model under a "linearity" condition. This remarkable property was extended to more general loss functions by [2]. They also showed that in the case of a multi index model (MIM) the direction captured belongs in the central mean subspace, and developed the Iterative Hessian Transformation (IHT) method to generate additional directions in the Central Mean Subspace (CMS). This paper examines the use the non-linear least squares for estimating the direction in the SIM, or a direction in the CMS. In the case of the SIM it is shown that the link function can be chosen so the resulting estimator is more efficient than the OLS estimator. Additional directions in the CMS can be estimated both iteratively and non-iteratively. Iterative estimation is more suitable for estimating the order of the MIM. Numerical results comparing the proposed method with the IHT are presented.

Keywords: Dimension reduction; Index models; Non-linear least squares.

References

[1] Brillinger, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In: Peter J. Bickel, Kjell A. Doksum, and J.L. Hodges, eds., A festschrift for Erich L. Lehmann, Wordsworth International Group, Belmont, CA.

[2] Cook, R.D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Annals of Statistics*, **30** 455-474.

2.6 Testing the adequacy of semiparametric transformation models

M. Hušková [1], S.G Meintanis[2] and J.S. Allison[3,*]

[1] Charles University of Prague, Department of Statistics, Czech Republic ; huskova@karlin.mff.cuni.cz
[2] Department of Economics, National and Kapodistrian University of Athens, Athens, Greece ; simosmei@econ.uoa.gr
[3] Unit for Business Mathematics and Informatics, North–West University, Potchefstroom, South Africa; james.allison@nwu.ac.za

Abstract: We consider a semiparametric model whereby the response variable following a transformation can be expressed by means of a nonparametric regression model. In this model the form of the transformation is specified analytically but incorporates an unknown transformation parameter. We develop testing procedures for the null hypothesis that this semiparametric model adequately describes the data at hand. In doing so, the test statistic is formulated on the basis of Fourier–type conditional expectations. The asymptotic distribution of the test statistic is obtained under the null as well as under alternative hypotheses. Since the limit null distribution is nonstandard, a bootstrap version is utilized in order to actually carry–out the test procedure. Monte Carlo results are included that illustrate the finite–sample properties of the new method.

Keywords: Transformation model; Goodness–of–fit test; Nonparametric regression; Bootstrap test

2.7 Nonparametric estimation of a transition probability matrix from left-truncated and right-censored data

J. de Uña-Álvarez

Department of Statistics and OR & Biomedical Research Center (CINBIO), University of Vigo; jacobo@uvigo.es

Abstract: Nonparametric estimation of a transition probability matrix for a general progressive multi-state model from possibly left-truncated and right-censored data is considered. To introduce a suitable estimator, the transition probability matrix is represented as depending on certain weighted (sub-)distributions of the entry and exit times for the several states. Asymptotic results, simulations, and real data illustrations are included. This work somehow extends de Uña-Álvarez and Meira-Machado (2015) to the left-truncated setting.

Keywords: Cross-sectional sampling; Multi-state models; Random truncation; Survival Analysis; Time-to-event data.

References

[1] de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. *Biometrics*, **61**, 364–375.

2.8 Non parametric estimation of space-varying distribution

A. Amiri[1], and S. Dabo[1]

[1] Univ. Lille, UMR 9221 aboubacar.amiri@univ-lille3.fr, sophie.dabo@univ-lille3.fr

Abstract: Let $N \geq 1$ an integer number and $\mathcal{D} \subset \mathbb{R}^N$ a N-dimensional spatial domain. We consider a spatial process $\{X_{\mathbf{s}}, \mathbf{s} \in \mathcal{D}\}$, with spatially varying distribution, where the restrictions of the distribution function are locally separable meaning that the sampled data do not come from the same distribution of interest but rather from a slowly changing distribution with respect to the space. The evolution of such a process

is modeled over a sequence of spatial regions, thereby the spatial domain $\mathcal{D}$ is segmented into a set of n pairwise disjoint regions $\mathcal{D}_i, i = 1, \ldots, n$.

It is assumed that inside each of these regions, a number of observations is made, such as cases for some disease or the magnitude of an earthquake on the Richter scale, color of pixels.

These observations form a sample from $F(\cdot, i)$, the distribution applicable in the domain $\mathcal{D}_i$. The intention of the exercise is to provide a space-varying estimation of the distribution $F(\cdot, i)$.

Keywords: Kernel smoothing; Probability distribution; Spatially varying.

References

[1] Grillenzoni, C. (2006). Sequential Kernel Estimation of the Conditional Intensity of Nonstationary Point Processes. *Statistical Inference for Stochastic Processes*, **9**(2), 135–160.

[2] Hall, P., Muller H-G. and Wu, P-S. (2006). Real-Time Density and Mode Estimation With Application to Time-Dynamic Mode Tracking. *Journal of Computational and Graphical Statistics*, **15**(1), 82–100.

[3] Harvey, A. and Oryshchenko, V. (2012). Kernel density estimation for time series data. *International Journal of Forecasting*, **28**, 3–14.

[4] Nieuwenhuis, C. and Cremers, D. (2013). Space-Varying Color Distributions for Interactive Multiregion Segmentation: Discrete versus Continuous Approaches *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **35**(5), 1234–1247.

2.9 Estimating under order restrictions on two dimensions

D. Anevski

Centre for Mathematical Sciences, Lund University, Sweden; dragi@maths.lth.se

Abstract: I discuss the problem of estimating a regression function or density function, defined on the two-dimensional reals and taking values in the reals, under the assumption that the unknown function is monotone with respect to the partial order on the two-dimensional reals. Of particular interest is a manageable characterization of the estimator, as well as limit distributions.

Keywords: Monotone, NPMLE, Isotonic regression, Brownian sheet

2.10 On the Consistency of the Crossmatch Test

Ery Arias-Castro[1,*] and Bruno Pelletier[2]

[1] Department of Mathematics, University of California, San Diego, CA, USA
[2] IRMAR – UMR CNRS 6625, Université Rennes II, France

Abstract: [2] proposed the crossmatch test for two-sample goodness-of-fit testing in arbitrary dimensions. We prove that the test is consistent against all fixed alternatives. In the process, we develop a general consistency result based on [3] that applies more generally.

Keywords: Two-sample goodness-of-fit testing; graph-based methods; permutation tests.

References

[1] Arias-Castro, E. and B. Pelletier (2015). On the Consistency of the Crossmatch Test *Journal of Statistical Planning and Inference* (in press).

[2] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(4), 515–530.

[3] Henze, N. and M. D. Penrose (1999). On the multivariate runs test. *Annals of statistics*, 290–298.

2.11 Nonparametric instrumental regression: Adaptive estimation in presence of dependence

N. Asin[1,*], and J. Johannes[2]

[1]ISBA, Université catholique de Louvain; `nicolas.asin\string@uclouvain.be`
[2]Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg; `johannes\string@math.uni-heidelberg.de`

Abstract: We consider in nonparametric instrumental regression the estimation of the structural function, which models the dependence of a response Y on the variations of an endogenous explanatory variable Z in the presence of an instrument W. Given an iid. sample of (Y, Z, W) a lower bound is derived for a maximal weighted mean integrated squared error assuming the structural function belongs to an ellipsoid linked in a certain sense to the conditional expectation operator of Z given W. We propose an estimator of the structural function based on a dimension reduction and an additional thresholding. Assuming either an iid. sample of (Y, Z, W) or a sufficiently weak dependence characterized by fast decreasing mixing coefficients it is shown that the lower bound provides also an upper bound of the estimators' maximal risk over a wide range of ellipsoids. We illustrate these results by considering classical smoothness assumptions. However, the proposed estimator requires an optimal choice of a dimension parameter depending on certain characteristics of the structural function and the conditional expectation operator of Z given W. As these are unknown in practice, we investigate a fully data-driven choice of the tuning parameter which combines model selection and Lepski's method. It is inspired by the recent work of [1]. It is shown that the adaptive estimator with data-driven choice of the dimension parameter can attain the lower minimax risk bound up to a constant, and this over a variety of classes of structural functions and conditional expectation operators.

Keywords: Nonparametric instrumental regression; Minimax theory; Projection estimation; Adaptive estimation; Mixing.

References

[1] Goldenshluger A. and Lepski O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, **39**, 1608–1632.

2.12 Change Point Analysis in Neuroimaging Data

J.A.D. Aston

Statistical Laboratory, DPMMS, University of Cambridge; j.aston@statslab.cam.ac.uk

Abstract: Change point analysis is becoming increasingly prevalent in neuroimaging, as more experiments with complex or even no underlying design (such as resting state fMRI) are being used. It is well known, however, that it is difficult to determine change points in situations were there are complex spatial and temporal dependencies. In this talk, a number of different approaches to determine the presence or absence of change points in either the first or second moment structure will be presented. These will include taking a functional approach to the data, where the brain image is seen as a discrete observation from a continuous 3-D function. It will be shown that taking such an approach can be both interesting from a statistical and applied point of view.

Keywords: CUSUM; functional Magnetic Resonance Imaging; Multiple Change Points; Mean Stationarity; Covariance Stationarity

2.13 Testing optimal dimension for suitable Hilbert valued processes

Jean-Baptiste Aubin[1,*], Enea G. Bongiorno[2] and Rosaria Ignaccolo[3]

[1] INSA-Lyon, ICJ, 20, Rue Albert Einstein, 69621 Villeurbanne Cedex, France; jean-baptiste.aubin@insa-lyon.fr,
[2] Università del Piemonte Orientale, Italy; enea.bongiorno@uniupo.it
[3] Dipartimento di Economia e Statistica, UniTO, Torino, Italy; rosaria.ignaccolo@unito.it

Abstract: The small–ball probability of a Hilbert valued process is considered. Recent works have shown that, for a fixed number d and as the radius ε of the ball tends to zero, the small–ball probability is asymptotically proportional to (a) the joint density of the first d principal components (PCs) evaluated at the center of the ball, (b) the volume of the d–dimensional ball with radius ε, and (c) a correction factor weighting the use of a truncated version of the process expansion. Under suitable assumptions on the decay rate of the eigenvalues of the covariance operator of the process, it has been shown that the correction factor in (c) tends to 1 as the dimension increases.
In this work, the properties of the correction factor are studied and a consistent estimator is introduced. Features of such estimator allow to conservatively test whenever the correction factor equals 1. This implicitly implies that, for the class of processes whose eigenvalues of the covariance operator decay hyper–exponentially, an optimal dimension can be defined allowing to use a "finite–dimensional" approach in approximating the small–ball probability and, hence, providing a natural model advantage.

Keywords: Small–ball probability.

2.14 Dating structural breaks in functional data without dimension reduction

A. Aue[1,*], G. Rice[2] and O. Sönmez[1]

[1] University of California, Davis; aaue@ucdavis.edu, osonmez@ucdavis.edu
[2] University of Waterloo; grice@uwaterloo.ca

Abstract: An estimator for the time of a break in the mean of stationary functional data is proposed that is fully functional in the sense that it does not rely on dimension reduction techniques such as functional principal component analysis (fPCA). A thorough asymptotic theory is developed for the estimator of the break date for fixed break size and shrinking break size. The main results highlight that the fully functional procedure performs best under conditions when analogous fPCA based estimators are at their worst, namely when the feature of interest is orthogonal to the leading principal components of the data. The theoretical findings are confirmed by means of a Monte Carlo simulation study in finite samples. An application to one-minute intra-day cumulative log-returns of Microsoft stock data highlights the practical relevance of the proposed fully functional procedure.

Keywords: Change-point analysis; Functional principal components; Functional time series; Intra-day financial data.

2.15 Bayesian Nonparametric Graph Clustering

Sayantan Banerjee[1], Rehan Akbani[1] and Veera Baladandayuthapani[1*]

[1] Dept. of Biostatistics, UT MD Anderson Cancer Center; veera@mdanderson.org

Abstract: We present clustering methods for multivariate data exploiting the underlying geometry of the graphical structure between variables. As opposed to standard approaches that assume known graph structures, we first estimate the edge structure of the unknown graph using Bayesian neighborhood selection approaches, wherein we account for the uncertainty of graphical structure learning through model-averaged estimates of the suitable parameters. Subsequently, we develop a nonparametric graph clustering model on the lower dimensional projections of the graph based on Laplacian embeddings using Dirichlet process mixture models. In contrast to standard algorithmic approaches, this fully probabilistic approach allows incorporation of uncertainty in estimation and inference for both graph structure learning and clustering. More importantly, we formalize the arguments for Laplacian embeddings as suitable projections for graph clustering by providing theoretical support for the consistency of the eigenspace of the estimated graph Laplacians. We develop fast computational algorithms that allow our methods to scale to large number of nodes. Through extensive simulations we compare our clustering performance with standard clustering methods. We apply our methods to a novel pan-cancer proteomic data set, and evaluate protein networks and clusters across multiple different cancer types.

Keywords: Dirichlet process mixture models; Graph clustering; Graph Laplacian; Graphical models; Proteomic data.

2.16 A general frequency domain method for assessing spatial covariance structure

M. Van Hala[1], S. Bandyopadhyay[2,*], S. N. Lahiri[3] and D. J. Nordman[1]

[1] Iowa State University; mvanhala@iastate.edu, dnordman@iastate.edu
[2] Lehigh University; sob210@lehigh.edu
[3] North Carolina State University; snlahiri@ncsu.edu

Abstract: Current methods for testing spatial covariance are often intended for specialized inference scenarios, usually with spatial lattice data. We propose instead a general method for estimation and testing of spatial covariance structure, which is valid for a variety of inference problems *and* applies to a large class of spatial sampling designs with irregular data locations. In this setting, spatial statistics have limiting distributions with complex standard errors depending on the rate of spatial sampling, the distribution of sampling locations, and the process covariance. The proposed method has the advantage of providing valid inference in the frequency domain without knowledge of such factors or estimation of standard errors. To illustrate, we develop the method for formally testing isotropy and separability in spatial covariance and consider confidence regions for spatial parameters in variogram model fitting; extensions to other testing applications are also presented. The approach uses spatial test statistics, based on an extended version of empirical likelihood, having simple chi-square limits. We demonstrate the proposed method through several numerical studies.

Keywords: Spatial periodogram; Spatial testing; Spectral moment conditions; Stochastic sampling.

2.17 Sample-Splitting in Non-Standard Problems and the Superefficiency Phenomenon

Moulinath Banerjee[1]

[1] Department of Statistics, University of Michigan; moulib@umich.edu

Abstract: We study the implications of the divide and conquer technique in monotone function estimation: specifically, how the global isotonic estimator based on a large dataset compares to a pooled estimator obtained by averaging isotonic estimates from a number of subsamples resulting from randomly splitting the original sample. The spectre of superefficiency is seen to rear its 'ugly' head: the pooled estimator beats the isotonic estimator for any given function but its performance suffers in terms of the maximal mean-squared error over a class of functions in a neighborhood of the given function. This is joint work with Cecile Durot and Bodhisattva Sen.

Keywords: monotone function, pooled estimator, sample splitting, superefficiency

2.18 On records and record times of stationary heavy tailed sequences

Bojan Basrak[1,*], Hrvoje Planinić[1] and Philippe Soulier[2]

[1] Department of Mathematics, University of Zagreb; bbasrak@math.hr, hrvoje.planinic@math.hr
[2] Université Paris Ouest Nanterre; philippe.soulier@u-paris10.fr

Abstract: We study records times and other order dependent functionals of stationary regularly varying sequences. Using point processes theory adapted to some non standard spaces, we are able to describe asymptotic distribution for the extremes and records in such a sequence, as long as the dependence between the observations is sufficiently weak. In particular, we obtain a remarkably simple structure for the limiting distribution of the record times.

Keywords: Regular variation; Point processes; Records; Extremal process

2.19 Network Modeling of High-dimensional Time Series with Applications to System-wide Risk Monitoring

Sumanta Basu[1,*], Sreyoshi Das[2], George Michailidis[3] and Amiyatosh Purnanandam[4]

[1] Department of Statistics, University of California, Berkeley; sumbose@berkeley.edu
[2] Department of Economics, University of Michigan; sreyoshi@umich.edu
[3] Department of Statistics, University of Florida; gmichail@ufl.edu
[4] Ross School of Business, University of Michigan; amiyatos@umich.edu

Abstract: Measuring connectedness among financial institutions is central in many aspects of financial economics, including system-wide risk monitoring and identifying systemically risky institutions. In this work, we present a unified framework for measuring connectivity among firms or asset classes from multivariate time series data. The proposed framework relies on regularized estimation of high-dimensional vector autoregressive models (VAR), is flexible enough to incorporate grouping and latent structures, allows parallel implementation for large data sets and enjoys strong theoretical guarantees under high-dimensional setting. We apply our method to analyze connectivity among stock returns and volatilities of leading financial firms in the U.S. before, during and after the financial crisis of 2007-2008, and show that the estimated networks can be used to identify important systemic events and systemically risky institutions.

Keywords: high-dimensional time series; graphical models; lasso; structured sparsity; systemic risk

2.20 Aggregation of supports along the Lasso path

Pierre C. Bellec[1]

[1] CREST-ENSAE, pierre.bellec@ensae.fr

Abstract: In linear regression with fixed design, we propose two procedures that aggregate a data-driven collection of supports. The collection is a subset of the 2^p possible supports and both its cardinality and its elements may depend on the data. The procedures satisfy oracle inequalities with no assumption on the design matrix. Then we use these procedures to aggregate the supports that appear on the regularization path of the Lasso in order to construct an estimator that mimics the best Lasso estimator. If the restricted eigenvalue condition on the design matrix is satisfied, then this estimator achieves optimal prediction bounds. Finally, we discuss the computational cost of these procedures.

Keywords: Fixed design regression, sparsity, regularization path of the Lasso, aggregation.

2.21 Trade-offs in statistical learning

Q. Berthet[1]

[1] University of Cambridge, DPMMS, Statslab

Abstract: I will explore the notion of constraints on learning procedures, and discuss the impact that they can have on statistical precision. This is inspired by real-life concerns such as limits on time for computation, on reliability of observations, or communication between agents. I will show how these constraints can be shown to have a concrete cost on the statistical performance of these procedures, by describing several examples. Joint work with Philippe Rigollet, Tengyao Wang, Richard J. Samworth, Venkat Chandrasekaran, Jordan Ellenberg

Keywords: High-dimensional statistics, Trade-offs

2.22 Nonparametric estimation of pregnancy outcome probabilities using a stabilized Aalen-Johansen estimator

J. Beyersmann

[1] Ulm University, Institute of Statistics, Helmholtzstrasse 20, D-89081 Ulm; jan.beyersmann@uni-ulm.de

Abstract: Estimating pregnancy outcome probabilities based on observational cohorts has to account for both left-truncation, because the time scale is gestational age, and for competing risks, because, e.g., a spontaneous abortion may be precluded by an elective termination. The applied aim of this work was to investigate drug safety in statin exposed pregnancies using data from a Teratology Information Service. Using the standard Aalen-Johansen estimator of the cumulative event probabilities suggested the medically implausible finding that statin exposure decreased the probability of elective termination and led to more live births. The reason was an early elective termination in a very small risk set, leading to unstable estimation which propagated over the whole time span. We suggest a stabilized Aalen-Johansen estimator which discards contributions from too small risk sets. The new estimator leads to a more meaningful analysis of the statin data. We also show that the new estimator enjoys the same asymptotic properties as the original Aalen-Johansen estimator - if one uses a little information from the future. We discuss why the present conditioning on the future does not compromise our analyses and investigate small sample properties in extensive simulations. Time permitting, we also discuss choice of tuning parameters using a left-truncated Brier score, semi-parametric modelling of dependent left-truncation and extensions to more general right-censored and/or left-truncated multistate models. This is joint work with Sarah Friedrich, Arthur Allignol, Martin Schumacher and Ursula Winterfeld.

Keywords: Survival analysis; Lai-Ying estimator

2.23 Random warping functions for landmark-constrained alignment of functional data

Karthik Bharath

University of Nottingham; karthik.bharath@nottingham.ac.uk

Abstract: Pairwise registration of functional or curve data defined on a domain D is can be carried out by determining a warping or reparameterization function from D to D that best matches the functions. In the presence of fixed landmarks, fixed-point constraints are imposed on the warping functions. In this situation, a case is made for the Dirichlet process as a candidate distribution on the requisite set of warping functions, notwithstanding the fact its sample paths are discrete with probability one. Some examples from simulated and real datasets will be presented.

Keywords: Curve registration; point process; diffeomorphisms

2.24 Fast sampling with Gaussian scale-mixture priors

A. Bhattacharya[1,*], A. Chakraborty[1] and B. Mallick[1]

[1] Texas A&M University; anirbanb@stat.tamu.edu, antik@stat.tamu.edu, bmallick@stat.tamu.edu

Abstract: We propose an efficient way to sample from a class of structured multivariate Gaussian distributions which routinely arise as conditional posteriors of model parameters that are assigned a conditionally Gaussian prior. The proposed algorithm only requires matrix operations in the form of matrix multiplications and linear system solutions. We exhibit that the computational complexity of the proposed algorithm grows linearly with the dimension unlike existing algorithms relying on Cholesky factorizations with cubic orders of complexity. The algorithm is broadly applicable in settings where Gaussian scale mixture priors are used on high dimensional model parameters.

Keywords: Bayesian; Confidence interval; High-dimensional; Shrinkage

2.25 Robust confidence intervals in high-dimensional censored regression

Jelena Bradic

Univ. California at San Diego, USA

Abstract: This paper develops the properties of robust one-step estimation in high-dimensional and censored regression models. Type I censored regression models are extremely common in prac- tice where a competing event makes the variable of interest unobservable. De-biasing technique is an appealing method but is not directly applicable to censored data. In this paper, we develop an approximate Fisher's scoring technique for the censored, least absolute deviation loss. The proposed method augments the naive de-biasing such that the resulting estimator is adaptive to the censoring and is more robust to the misspecification of the error distribution. A novel Type II censored regression approach is proposed to obtain asymptotically efficient estimation of the variability of each parameter estimates. We propose a unified class of robust estimators, including Mallow's, Schweppe's and Hill-Ryan's one-step estimators. In the ultra-high-dimensional setting, where the dimensionality can grow exponentially with the sample size, we show that as long as the preliminary estimator converges faster than $n^{+1/4}$ the one-step estimator inherits asymptotic distribution of fully iterated version. Moreover, we show that the size of the residuals of the Bahadur representation match those of the simple linear models, $s^3/4(log(p \vee n))^3/4/n^1/4$ – that is, the effects of censoring asymptotically disappear. Simulation studies demonstrate that our method is adaptive to the censoring level and asymmetry in the error distribution, and does not lose efficiency when the errors are from the symmetric distribution. Finally, we analyze a real data set from the MAQC-II repository that is related to the HIV-1 study.

2.26 Regression with Selectively Missing Covariates

Christoph Breunig[1]

[1] Humboldt-Universität zu Berlin; christoph.breunig@hu-berlin.de

Abstract: We consider the problem of regression with selectively observed covariates. Identification of the regression function relies on instrumental variables that are independent of selection conditional on potential covariates. We propose a consistent two-step estimation procedure and derive its rate of convergence; also its pointwise asymptotic distribution is established. We demonstrate the usefulness of our method in survey data with non random missingness.

Keywords: Endogenous selection; Instrumental variable; Sieve estimation.

2.27 A RKHS-based proposal for variable selection in functional regression

J.R. Berrendero[1], B. Bueno-Larraz[1,*] and A. Cuevas[1]

[1] Universidad Autónoma de Madrid; joser.berrendero@uam.es, beatriz.bueno@uam.es, antonio.cuevas@uam.es

Abstract: Variable selection techniques have become a popular tool for dimension reduction with an easy interpretation. Here we propose a new variable selection methodology for regression problems, with a model-based motivation. We focus on the case of a functional predictor and a scalar response. We try to approximate the response variable in the subspace generated by the random variables of the stochastic process. Then, through the Loève's isometry [1], we can rewrite it as a functional optimization problem in a RKHS. The reproducing kernel of the RKHS associated with the process plays the role of a "delta" function, that selects and combines temporal instants of the process. In this context, under a sparsity assumption, we can find the more relevant points for the original regression problem.
Our proposed method is an iterative approximation to this optimization. This is an easy-to-interpret methodology which allows for easily adding extra information about the model. Some simulations and real data examples are given, including some comparisons with other recent proposals. Some results of rates and asymptotic convergence were already derived in [2] for a particular case of the method.

Keywords: Functional regression; Variable selection; RKHS; Loève's isometry.

References

[1] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Kluwer Academic, Boston.

[2] McKeague, I. W. and Sen, B. (2010). Fractals with point impact in functional linear regression. *Ann. Statis.*, **38(4)**, 2559–2586.

2.28 Sufficient Reductions in Regression and Classification with Exponential Family Predictors

Efstathia Bura

George Washington University

Abstract: A main objective of statistical inference is the reduction of data: variables are replaced by relatively few quantities (reductions) which adequately represent the relevant information contained in the original data. Sufficient dimension reduction methodology (a) identifies and (b) provides estimation algorithms of sufficient reductions in regressions/classifications with many predictors. The reductions are sufficient or exhaustive in the sense that they are all that is needed to model and predict the response(s) and result in reducing big data to "analyzable" size. The talk will focus on model-based sufficient reductions for regressions with predictors in the multivariate exponential family of distributions. This set-up includes regressions where predictors are all continuous, all categorical or mixtures of the two. The minimal sufficient reduction of the predictors and its maximum likelihood estimator are derived by modeling the conditional distribution of the predictors given the response. Whereas nearly all extant estimators of sufficient reductions are linear and only partly capture the sufficient reduction, our method identifies both the linear and the nonlinear components of the sufficient reduction. It also provides the exact form of the sufficient reduction, which is exhaustive, its maximum likelihood estimates via an iterative re-weighted least squares (IRLS) estimation algorithm and asymptotic tests for the dimension of the regression. A classification example with multivariate Bernoulli predictors, a central problem in the machine learning community, will be presented.

2.29 Adaptive minimax tests for large covariance matrices with incomplete data

C. Butucea and R. Zgheib

Univ. Paris-Est Marne-la-Vallée, France; cristina.butucea@u-pem.fr, rania.zgheib@u-pem.fr

Abstract: We observe n independent $p-$dimensional Gaussian vectors with missing coordinates, that is each value (which is assumed standardized) is observed with probability $a > 0$. We investigate the problem of minimax nonparametric testing that the high-dimensional covariance matrix Σ of the underlying Gaussian distribution is the identity matrix, using these partially observed vectors. Here, n and p tend to infinity and $a > 0$ tends to 0, asymptotically. We assume that the covariance matrix Σ belongs to a Sobolev-type ellipsoid with parameter $\alpha > 0$. When α is known, we give asymptotically minimax consistent test procedure and find the minimax separation rates $\tilde{\varphi}_{n,p} = (a^2 n\sqrt{p})^{-\frac{2\alpha}{4\alpha+1}}$, under some additional constraints on n, p and a. We show that, in the particular case of Toeplitz covariance matrices, the minimax separation rates are faster by a $\sqrt{p}$ factor, $\tilde{\phi}_{n,p} = (a^2 np)^{-\frac{2\alpha}{4\alpha+1}}$. We note how the "missingness" parameter a deteriorates the rates with respect to the case of fully observed vectors ($a = 1$). We also propose adaptive test procedures, that is free of the parameter α in some interval, and show that the loss of rate is $(\ln\ln(a^2 n\sqrt{p}))^{\alpha/(4\alpha+1)}$ and $(\ln\ln(a^2 np))^{\alpha/(4\alpha+1)}$ for Toeplitz covariance matrices, respectively.

Keywords: Adaptive test, Covariance matrices, Goodness-of-fit, Minimax separation rate, Missing data.

References

[1] Butucea, C. (2016). Adaptive tests for large covariance matrices with missing observations, arxiv:2016.04310.

2.30 Efficient Use of EMR for Discovery Research

Abhishek Charborty[1], Jessica Gronsbell[1] and Tianxi Cai[1*]

[1] Department of Biostatistics, Harvard T. H. Chan School of Public Health

Abstract: In clinical practice, patients with the same disease diagnosis often differ in outcomes and response to treatment. The ability to both classify and predict disease phenotypes would be a valuable asset in clinical decision-making. Large datasets containing both a wealth of clinical and experimental data now exist as a result of the increasing adoption of electronic medical records (EMR) linked with specimen bio-repositories. These datasets allow for data driven classification and prediction of sub-phenotypes and investigation of shared risk factors across a group of phenotypes. In this talk, I'll discuss various statistical and informatics methods that illustrate both the challenges and potential

opportunities that arise from analyzing EMR data. For example, obtaining validated phenotype information is a major bottleneck in EMR research, as it requires laborious medical record review. Thus gold standard labels are typically available only in a small training set nested in a large cohort. In contrast, data on the clinical predictors of the phenotype are often available on all subjects. To improve phenotype definition, we developed robust semi-supervised learning methods that can leverage such rich source of auxiliary information. These methods are illustrated with an EMR cohort of rheumatoid arthritis patients.

Keywords: Efficiency Gain; Electronic Medical Records; Model mis-specification; Prediction Performance; Semi-supervised Learning.

2.31 Nonparametric Inference for big-but-biased data

R. Cao

Faculty of Computer Science, Universidade da Coruña, Spain; rcao@udc.es

Abstract: [2] has recently warned about the risks of the sentence "with enough data, the numbers speak for themselves". Some of the problems raising from ignoring sampling bias in big data statistical analyses has been recently reported by [1]. The problem of nonparametric statistical inference in big data under the presence of sampling bias is considered in this work. The mean estimation problem is studied in this setup when the biasing weight function is known (unrealistic) as well as for known weight functions (realistic). The question of how big the sample size has to be to compensate the sampling bias in big data is considered. Some examples and simulations illustrate the performance of the nonparametric methods proposed in this work.

Keywords: Big data; kernel method; length bias; smoothing parameter.

References

[1] Cao, R. (2015). Inferencia estadística con datos de gran volumen. La Gaceta de la RSME, 18, 393-417.

[2] Crawford, K. (2013). The hidden biases in big data. Harvard Business Review 2013, april 1st. Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data

2.32 Confidence sets for matrix completion

A. Carpentier[1,*], O. Klopp[2], M. Löffler[3] and R. Nickl[3]

[1] Universität Potsdam; carpentier@math.uni-potsdam.de
[2] Université Paris Nanterre and CREST; olga.klopp@math.cnrs.fr
[3] University of Cambridge; ml718@cam.ac.uk, r.nickl@statslab.cam.ac.uk

Abstract: This presentation will be about adaptive and honest confidence sets for high dimensional, bounded and low rank matrix completion. Two design assumptions will be considered:

a) that the (noisy) entries of the matrix are sampled uniformly at random

b) that each (noisy) entry of the matrix has a given probability of being revealed.

If an additional information on the noise that is added to the entries, e.g. its variance, is not available, then one can prove that although adaptive and honest confidence sets exist in model a), they do not exist in model b). This highlight a fundamental difference between models a) and b), which does not exist in the case of optimal and adaptive estimation of the low rank matrix (where the optimal rates of estimation are the same up to logarithmic factors in both models).

Keywords: matrix completion ; uncertainty quantification ; confidence sets ; composite tests

2.33 The duality of light-heavy tails

Joan del Castillo[1,*] and Maria Padilla[1]

[1] Department of Mathematics, Universitat Autonoma de Barcelona, Barcelona, Spain; castillo@mat.uab.cat, mpadilla@mat.uab.cat.

Abstract: Extreme value theory has recently achieved great interest in several fields. In practical applications often a mixture of non-parametric and parametric methods is required. The Peak over Threshold method splits the sample into two parts: the body, where non-parametric estimators are used, and the tail, where the Pickands-Balkena-DeHaan theorem suggests using generalized Pareto distribution (GPD). It is especially important for applications to determine the threshold where the tail begins and to distinguish between polynomial and exponential tails. Hill-plot and mean excess (ME) plot are graphics methods identifying the threshold between the body and the tail of the distribution, but the selection is often a matter of subjective choice based on visual inspection. [1] introduced the residual coefficient of variation as a stochastic process showing some advantages over ME-plot: it does not depend on the scale parameter and detecting constant functions is easier than linear functions. However, the method only works when the extreme value index is smaller than 0.25. To fix this, some transformations that relate light-heavy tails are introduced. The methodology is updated with multiple threshold tests for a GPD and a threshold selection algorithm designed in a way that avoids subjectivity as much as possible. Finally, a main contribution is to extend the methodology based on moments to all distributions, even with no finite moments.

Keywords: Statistics of extremes; Heavy tails; High quantile estimation; Value at risk.

References

[1] Castillo, J., Daoudi, J. and Lockhart, R. (2014). Methods to distinguish between polynomial and exponential tails. *Scandinavian Journal of Statistics* , **41**, 382–393.

2.34 Nonparametric Instrumental Variable Estimation of Additive Models with Multiple Endogenous and Exogenous Components

Samuele Centorrino[1*] and Sorawoot Srisuma[2]

[1] Economics Department, Stony Brook University; samuele.centorrino@stonybrook.edu
[2] Department of Economics, University of Surrey; s.srisuma@surrey.ac.uk

Abstract: Models with multiple endogenous variables are often considered in applied economics, for instance in the estimation of demand or production functions. This paper provides conditions for the identification of a fully nonparametric regression model with multiple endogenous and exogenous variables under an additivity constraint. The exogenous components can be identified under standard conditions. The endogenous components are identified with an aid of instrumental variables. For the latter, we only require component-wise completeness condition as long as the ranges of the corresponding conditional expectation operators have empty intersection. We propose to estimate our model in two steps. The first is a sieve regularization estimator, which allows us to easily impose the additive structure on the model; and a second step kernel Tikhonov estimator, to derive the asymptotic properties of each additive component. We show that by appropriately controlling the divergence of the regularization parameter in the first step, we are able to obtain minimax rates of convergence of our nonparametric estimator. We conclude with an empirical illustration by estimating the production function using panel data from Chilean firms.

Keywords: Nonparametric; Endogeneity; Instruments; Additive Models; Regularization.

2.35 Adaptive estimation of a conditional distribution given a functional covariate

G. Chagny[1,*] and A. Roche[2]

[1] LMRS, Univ. Rouen; gaelle.chagny@univ-rouen.fr [2] CEREMADE, Univ. Paris Dauphine; angelina.roche@dauphine.fr

Abstract: We study the link between a real random variable of interest Y and a predictor X which is a functional random variable (typically a curve). Both regression and conditional cumulative distribution function estimation are considered. Starting from collection of kernel estimators introduced by [2], we propose fully data-driven bandwidth

selection rules (local or global) inspired both by [3] and by model selection. The selected estimators are shown to be adaptive and minimax optimal up to a logarithmic loss: nonasymptotic bounds and convergence rates are derived under various assumptions on the decay of the small ball probability of the functional variable. Most of the corresponding lower bounds are established. Numerical results illustrate the method. The choice of the semi-norm involved in the definition of the estimators, as well as open questions will be discussed.

Keywords: Functional data analysis; Minimax and adaptive estimation; Regression model.

References

[1] Chagny, G. and Roche, A. (2016) Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, **146**, 105–118.

[2] Ferraty, F. and Laksaci, A. and Vieu, P. (2006) Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, **9**, no. 1, 47–76.

[3] Goldenshluger, A. and Lepski, O. (2011) Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Annals of Statistics*, **39**, no. 3, 1608–1632.

2.36 Hybrid regularization of functional linear models

A. Chakraborty* and V. M. Panaretos

École Polytechnique Fédérale de Lausanne; vanchak@gmail.com, victor.panaretos@epfl.ch

Abstract: We consider the problem of estimating the slope function in a functional linear model with a scalar response and a functional covariate. This central problem of functional data analysis is well known to be ill-posed, thus requiring a regularized estimation procedure. The two most commonly used approaches are based on spectral truncation and Tikhonov regularization of the empirical covariance operator. In principle, Tikhonov regularization is the more canonical choice, as it is robust to eigenvalue ties while attaining the asymptotically optimal minimax rate of convergence in mildly ill-posed settings. In this talk, we discuss that, surprisingly, one can strictly improve upon the performance of the Tikhonov estimator while retaining its stability properties by combining it with a form of spectral truncation. Specifically, we construct a hybrid estimator that additively decomposes the functional covariate by projecting it onto two orthogonal subspaces defined via functional PCA; it then applies Tikhonov regularization to one component, while leaving the other component unregularized. We show that this hybrid estimator enjoys the same minimax optimal rates as the Tikhonov estimator, but that it strictly and uniformly improves upon it in a non-asymptotic sense, for all sufficiently large sample sizes. We discuss the performance of the hybrid estimator by means of a simulation and demonstrate that one can make considerable gains even in finite samples.

Keywords: Functional data analysis, Mean squared error, Principal component analysis, Spectral truncation, Tikhonov regularization.

2.37 Confidence intervals for contextual bandits

A. Chambaz[1*], and M. J. van der Laan[2]

[1] Modal'X, Université Paris Ouest Nanterre La Défense; achambaz@u-paris10.fr
[2] Division of Biostatistics, UC Berkeley; laan@berkeley.edu

Abstract: An operator can undertake one of two actions. Each action is rewarded randomly, from a conditional law given the context in which the action is carried out. The objective is to learn, by repeating the experiment parsimoniously, *(i)* the conditional law of the optimal action given the context and *(ii)* the mean reward under this degenerate law. In this framework of contextual bandits, we address *(i)* and *(ii)* from the angle of inference rather than that of regret minimization. Under mild assumptions, we obtain narrow confidence intervals on the mean reward but also on different cumulative regrets.

Keywords: Contextual Bandits; Maximal inequality; Targeted Minimum Loss Inference.

2.38 Change-point detection for locally dependent data

H. Chen

University of California, Davis; hxchen@ucdavis.edu

Abstract: Local dependence is common in multivariate and object data sequences. We consider the testing and estimation of change-points in such sequences. A new way of permutation, circular block permutation with a randomized starting point, is proposed and studied for a scan statistic utilizing graphs representing the similarity between observations. The proposed permutation approach could correctly address for local dependence and make it possible the theoretical treatments for the non-parametric graph-based scan statistic for locally dependent data. We derive accurate analytic approximations to the significance of graph-based scan statistics under the circular block permutation framework, facilitating its application to locally dependent multivariate or object data sequences.

Keywords: Local dependence; Circular block permutation; Change-point; Multivariate data; Object data.

2.39 Statistical Inference for Matrix-variate Gaussian Graphical Models and False Discovery Rate Control

Xi Chen[1] and Weidong Liu[2]

[1]Stern School of Business, New York University
[2] Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai

Abstract: Matrix-variate Gaussian graphical models (GGM) have been widely used for modelling matrix-variate data. Since the supports of sparse row and column precision matrices encode the conditional independence among rows and columns of the data, it is of great interest to conduct support recovery. A commonly used approach is the penalized log-likelihood method. However, due to the complicated structure of the precision matrices of matrix-variate GGMs, the log-likelihood is non-convex, which brings a great challenge for both computation and theoretical analysis. In this paper, we propose an alternative approach by formulating the support recovery problem into a multiple testing problem. A new test statistic is proposed and based on that, we further develop a method to control false discovery rate (FDR) asymptotically. Our method is computationally attractive since it only involves convex optimization. Theoretically, our method allows very weak conditions, i.e., even when the sample size is finite and the dimensions go to infinity, the asymptotic normality of the test statistics and FDR control can still be guaranteed. The finite sample performance of the proposed method is illustrated by both simulated and real data analysis.

2.40 Structured Sufficient Dimension Reduction and its Applications

F. Chiaromonte[1,2,*], Yang Liu[1], Bing Li[1]

[1] The Pennsylvania State University; fxc11@psu.edu, ywl5222@psu.edu, bxl9@psu.edu
[2] Sant′Anna School of Advanced Studies

Abstract: In this talk we describe Structured Ordinary Least Squares (sOLS), a method for Sufficient Dimension Reduction in regression problems where both the predictors and the statistical units present a group structure. We also introduce some recent generalization of sOLS to settings in which the response is binary, or the data is collected spatially – creating the need to correct for correlated observations. Finally, we illustrate the use of sOLS and its variants with applications to Genomics data, where we investigate features affecting the prevalence of non-coding functional elements and de novo mutations in the human genome, as well as Medicare Provider Utilization and Payment data, where we investigate the determinants of per-capita health care costs for the elderly population in the US. This is joint work with the groups of Kateryna Makova and Guido Cervone at The Pennsylvania State University.

Keywords: Sufficient Dimension Reduction; Group Structure; Genomics Data; Medicare Data.

2.41 Simultaneous change-point and factor analysis for high-dimensional time series

H. Cho[1,*], M. Barigozzi[2] and P. Fryzlewicz[2]

[1] School of Mathematics, University of Bristol, UK; haeran.cho@bristol.ac.uk
[2] Department of Statistics, London School of Economics, UK

Abstract: In this paper, we propose a method for simultaneously analysing the factor structure of the data and detecting (possibly) multiple change-points in high-dimensional time series.

Firstly, we introduce a piecewise stationary factor model that enables introducing and, consequently, detecting changes not only in loadings but also in factors and idiosyncratic component, which has not been explored in the existing literature. Next, it is shown that the common component estimated with an over-estimated factor number achieves consistency, which motivates our change-point detection methodology. Then, we propose to transform the data so that an existing panel data segmentation method [1] is applicable to the problem of detecting multiple change-points in the factor structure, and consistency of such an approach is established in terms of the total number and locations of estimated change-points. Empirical performance of the proposed method is investigated on simulated datasets as well as macroeconomic and financial time series.

Keywords: change-point analysis, factor analysis, high-dimensional time series, dimension reduction, nonstationary time series

References

[1] Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *In submission*.

2.42 Variable Selection in Heteroscedastic Single Index Quantile Regression

E. Christou[1,*] and M. G. Akritas[1]

[1] The Pennsylvania State University; exc277@psu.edu, mga@stat.psu.edu

Abstract: Quantile regression (QR) has become a popular method of data analysis, especially when the error term is heteroscedastic, due to its relevance in many scientific studies. The ubiquity of high dimensional data has led to a number of variable selection methods for linear/nonlinear QR models and, recently, for the single index quantile regression (SIQR) model. We propose a new algorithm for simultaneous variable selection and parameter estimation applicable also for heteroscedastic data. The proposed algorithm, which is non-iterative, consists of two steps. Step 1 performs an initial variable selection method. Step 2 uses the results of Step 1 to obtain better estimation of the conditional quantiles and, using them, to perform simultaneous variable selection and estimation of the parametric component of the SIQR model. It is shown that the initial variable selection method of Step 1 consistently estimates the relevant variables, and that the estimated parametric component derived in Step 2 satisfies the oracle property.

Keywords: Dimension reduction; Index model; Nadaraya-Watson estimator; Quantile regression; SCAD penalty.

2.43 Bootstrap uniform central limit theorems for Harris recurrent Markov chains

Gabriela Ciołek

Univ. Science and Technology, Krakow, Poland and Modal'X, Univ. Paris Ouest, France, gabrielaciolek@gmail.com

Abstract: The main objective of this talk is to present bootstrap uniform functional central limit theorem for Harris recurrent Markov chains over uniformly bounded classes of functions. We show that the result can be generalized also to the unbounded case. To avoid some complicated mixing conditions, we make use of the well-known regeneration properties of Markov chains. We show that in the atomic case the proof of the bootstrap uniform central limit theorem for Markov chains for functions dominated by a function in L^2 space proposed by [2] can be significantly simplified. Regenerative properties of Markov chains can be applied in order to extend some concepts in robust statistics from i.i.d. to a Markovian setting. [1] have defined an influence function and Fréchet differentiability on the torus what allowed to extend the notion of robustness from single observations to the blocks of data instead. In this talk, we present bootstrap uniform central limit theorems for Fréchet differentiable functionals in a Markovian case.

Keywords: Bootstrap; Empirical processes indexed by classes of functions; Markov chains; Regenerative processes; Fréchet differentiability.

References

[1] Bertail, P., Clémençon, S. (2006). Regenerative block bootstrap for Markov chains. *Bernoulli*,**12**, 689–712.

[2] Radulović, D. (2004). Renewal type bootstrap for Markov chains. *Test*, **13**, 147–192.

2.44 Nonparametric methods for change-point detection in parametric models

G. Ciuperca

Université Lyon 1, Institut Camille Jordan, France; Gabriela.Ciuperca@univ-lyon1.fr

Abstract: We consider a posteriori and in real time change-point models. The parametric regression functions of the each phase can be nonlinear or linear, and moreover, in the linear case, the number of the explanatory variables could be large. Theoretical results and simulations are presented for each model and nonparametric method. For a posteriori nonlinear change-point model, the results obtained by two nonparametric estimation techniques are given in the case when the change-point number is known. So, the quantile and empirical likelihood nonparametric methods are considered. If the number of the change-points is unknown, a consistent criteria is presented. When the change-point model is linear but with a large number of explanatory variables, then it would make the automatic selection of variables. The adaptive quantile LASSO method is then proposed and studied. In the other hand, we propose a nonparametric test based on the empirical likelihood, in order to test if the model changes. For detecting in real time a change in model, we consider two cases. For a nonlinear model, a hypothesis test based on weighted CUSUM of least squares residuals is constructed. For a linear model with large number of explanatory variables, we propose a CUSUM test statistic based on adaptive LASSO residuals.

Keywords: Change-point; Quantile model; Empirical likelihood; CUSUM; LASSO.

2.45 Laguerre estimation for k-monotone densities observed with noise

D. Belomestny[1], F. Comte[2,*] and V. Genon-Catalot[2]

[1] Duisburg-Essen University; denis.belomestny@uni-due.de

[2] Univ. Paris Descartes, MAP5, fabienne.comte@parisdescartes.fr, valentine.genon-catalot@parisdescartes.fr

Abstract: We study the models $Z_i = Y_i + V_i, Y_i = X_i U_i,\ i = 1, \ldots, n$ where the V_i's are nonnegative, *i.i.d.* with known density f_V, the U_i's are *i.i.d.* with $\beta(1, k)$ density, $k \geq 1$, the X_i's are *i.i.d.*, nonnegative with unknown density f. The sequences $(X_i), (U_i), (V_i)$ are independent. We aim at estimating f on $\mathbb{R}^+$ in the three cases of direct observations $(X_1, \ldots, X_n)$, observations $(Y_1, \ldots, Y_n)$, observations $(Z_1, \ldots, Z_n)$. We propose projection estimators using a Laguerre basis and give upper bounds on the $\mathbb{L}^2$-risks on specific Sobolev-Laguerre spaces. Lower bounds matching with the upper bounds are proved in the case of direct observation of X and in the case of observation of Y. A general data-driven procedure is described and proved to perform automatically the bias variance compromise. The method is illustrated on simulated data.

Keywords: Adaptive estimation; Lower bounds; Model selection; Multiplicative censoring; k-monotone densities; Projection estimator.

References

[1] Ehrenberg, A. C. S. (1982). Writing technical reports and papers. *The American Statistician*, **36**, 326–329.

[2] László Györfi and Michael Kohler and Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer. New-York.

[3] Lamport, L. (1986). *LaTeX A Document Preparation System.* Addison-Wesley. Boston.

2.46 Estimating the boundary measure of a set in the Euclidean space

Antonio Cuevas

Departamento de Matemáticas, Universidad Autónoma de Madrid; antonio.cuevas@uam.es

Abstract: We consider the problem of estimating several geometric quantities of interest for a compact set S in the Euclidean space from random information of different types. We especially focus on the problem of estimating the measure of ∂S, defined in terms of the Minkowski content. We will briefly review some previous contributions and then summarize our recent research on this topic, which is still work in progress. The assumptions of positive reach for S and the (closely related) assumption of polynomial volume will play an especially relevant role in our approach. This talk is a summary of joint work with **Catherine Aaron**, ***Alejandro Cholaquidis*** *and* ***Beatriz Pateiro-López****.*

Keywords: Minkowski content; Positive reach; Polynomial volume.

2.47 Optimal robust estimation of the precision matrix based on row sparsity

S. Balmand[1] and A. Dalalyan[2,*]

[1] ENSG; samuel.balmand@ensg.eu
[2] ENSAE ParisTech, CREST; arnak.dalalyan@ensae.fr

Abstract: Multivariate Gaussian distribution is often used as a first approximation to the distribution of high-dimensional data. Determining the parameters of this distribution under various constraints is a widely studied problem in statistics, and is often considered as a prototype for testing new algorithms or theoretical frameworks. In this paper, we develop a nonasymptotic approach to the problem of estimating the parameters of a multivariate Gaussian distribution when data are corrupted by outliers. We propose an estimator-efficiently computable by solving a convex program-that robustly estimates the population mean and the population covariance matrix even when the sample contains a significant proportion of outliers. In the case where the dimension p of the data points is of smaller order than the sample size, our estimator of the corruption matrix is provably rate optimal simultaneously for the entry-wise l_1-norm, the Frobenius norm and the mixed l_2/l_1 norm. Furthermore, this optimality is achieved by a penalized square-root-of-least-squares method with a universal tuning parameter (calibrating the strength of the penalization). These results are partly extended to the case where p is potentially larger than n, under the additional condition that the inverse covariance matrix is sparse.

Keywords: Precision matrix; group-lasso; minimax rate.

2.48 A Rank-Sum Test for Clustered Data when the Number of Subjects in a Group within a Cluster is Informative

Sandipan Dutta[1], and Somnath Datta[2,*]

[1] University of Louisville, Louisville, KY, USA; sandipan.dutta@louisville.edu
[2] University of Florida, Gainesville, FL 32610, U.S.A.; somnath.datta@ufl.edu

Abstract: The Wilcoxon rank-sum test is a popular nonparametric test for comparing two independent populations (groups). In recent years, there have been renewed attempts in extending the Wilcoxon rank sum test for clustered data, one of which [1] addresses the issue of informative cluster size, i.e., when the outcomes and the cluster size are correlated. We are faced with a situation where the group specific marginal distribution in a cluster depends on the number of observations in that group (i.e., the intra-cluster group size). We develop a novel extension of the rank-sum test for handling this situation. We compare the performance of our test with the Datta-Satten test, as well as the naive Wilcoxon rank sum test. Using a naturally occurring simulation model of informative intra-cluster group size, we show that only our test maintains the correct size. We also compare our test with a classical signed rank test based on averages of the outcome values in each group paired by the cluster membership. While this test maintains the size, it has lower power than our test. Extensions to multiple group comparisons and the case of clusters not having samples from all groups are also discussed. We apply our test to determine whether there are differences in the attachment loss between the upper and lower teeth and between mesial and buccal sites of periodontal patients.

Keywords: Correlated data; Dental data; Nonparametric tests; Wilcoxon rank-sum test; Within-cluster resampling

References

[1] Datta, S. and Satten, G. A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association*, **100**, 908–915.

2.49 Differential Network Analysis with Multiply Imputed Lipidomic Data

Maiju Kujala[1], Jaakko Nevalainen[2], and Susmita Datta[3,*]

[1] University of Turku, Turku, Finland; mekuja@utu.fi
[2] University of Tampere, FI-33014 Tampere, Finland; jaakko.nevalainen@uta.fi
[3]University of Florida, Gainesville, FL 32610, U.S.A.; susmita.datta@ufl.edu

Abstract: The importance of lipids for cell function and health has been widely recognized, e.g., a disorder in the lipid composition of cells has been related to atherosclerosis caused cardiovascular disease (CVD). Lipidomics analyses are characterized by large yet not a huge number of mutually correlated variables measured and their associations to outcomes are potentially of a complex nature. Differential network analysis provides a formal statistical method capable of inferential analysis to examine differences in network structures of the lipids under two biological conditions. It also guides us to identify potential relationships requiring further biological investigation. We provide a recipe to conduct permutation test on association scores resulted from partial least square regression with multiple imputed lipidomic data from the LUdwigshafen RIsk and Cardiovascular Health (LURIC) study, particularly paying attention to the left-censored missing values typical for a wide range of data sets in life sciences. Left-censored missing values are low-level concentrations that are known to exist somewhere between zero and a lower limit of quantification. To make full use of the LURIC data with the missing values, we utilize state of the art multiple imputation techniques and propose solutions to the challenges that incomplete data sets bring to differential network analysis. The customized network analysis helps us to understand the complexities of the underlying biological processes by identifying lipids and lipid classes that interact with each other, and by recognizing the most important differentially expressed lipids between two subgroups of coronary artery disease (CAD) patients, the patients that had a fatal CVD event and the ones who remained stable during two year follow-up.

Keywords: Differential connectivity; Lipidomics; Bootstrap; Imputation

2.50 Ruin probability for correlated Brownian motions

K. Dębicki[1], E. Hashorva[2], L. Ji[2,*] and T. Rolski[1]

[1] Mathematical Institute, University of Wrocław, Wrocław, Poland; debicki@math.uni.wroc.pl, rolski@math.uni.wroc.pl
[2] University of Lausanne, Lausanne, Switzerland; Enkelejd.Hashorva@unil.ch, Lanpeng.Ji@unil.ch

Abstract: Let $\boldsymbol{B}(t) = (B_1(t), \dots, B_d(t))^{'}, t \geq 0$ be a standard d-dimensional Brownian motion with independent coordinates. For a nonsingular matrix A and vectors $\boldsymbol{\alpha} > \mathbf{0}$, $\boldsymbol{\mu}$ with $\max_{1\leq i\leq d} \mu_i > 0$, we focus exact asymptotics of ruin probability

$$\mathbb{P}\left(\exists\ t \geq 0 : \boldsymbol{A}\boldsymbol{X}(t) - \boldsymbol{\mu} t > \boldsymbol{\alpha} u\right),$$

as $u \to \infty$. Additionally, we analyze properties of multidimensional counterparts of Pickands and Piterbarg constants that appear in the derived asymptotics.

Keywords: Exact asymptotics; Supremum distribution; Ruin probability.

2.51 Nonparametric trend in extreme value indices

L. de Haan[1,*] and C. Zhou[2,1]

[1] Erasmus University Rotterdam; ldehaan@ese.eur.nl, zhou@ese.eur.nl
[2] Bank of The Netherlands; c.zhou@dnb.nl

Abstract: Let $\{X_i\}_{i=1,2,\cdots,n}$ be independent random variables. Assume that the distribution function of each X_i, F_i, is in the domain of attraction of some extreme value distribution. The extreme value index corresponding to F_i is denoted as $\gamma(i/n)$, where $\gamma(s)$ is a continuous positive function defined on $[0, 1]$. Under appropriate conditions on these distribution functions we can estimate $\gamma(s)$ locally. A global estimator of the integrated $\gamma(s)$-function can also be established. This allows for accurate estimation and testing regarding the $\gamma(s)$ function. The asymptotic normality for both local and global estimators are shown. The main tool in the proof is the extension of the tail empirical process: we derive asymptotic properties of the moments of the error terms in tail empirical processes.

Keywords: Non-stationarity; Tail empirical process.

2.52 Nonparametric covariate-adjusted regression

A. Delaigle[1,*], P. Hall[1] and W. Zhou[2]

[1] School of Mathematics and Statistics, University of Melbourne and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS); A.Delaigle@ms.unimelb.edu.au,
[2]Department of Operations Research and Financial Engineering, Princeton University; wenxinz@princeton.edu

Abstract: We consider nonparametric estimation of a regression curve when the data are observed with multiplicative distortion which depends on an observed confounding variable. This problem was oiginally studied in the parametric context by [1]. We suggest several estimators, ranging from a relatively simple one that relies on restrictive assumptions usually made in the literature, to a sophisticated piecewise approach that involves reconstructing a smooth curve from an estimator of a constant multiple of its absolute value, and which can be applied in much more general scenarios. We show that, although our nonparametric estimators are constructed from predictors of the unobserved undistorted data, they have the same first order asymptotic properties as the standard estimators that could be computed if the undistorted data were available. We illustrate the good numerical performance of our methods on both simulated and real datasets. This talk is based on the paper by [2].

Keywords: discontinuities; local linear estimator; multiplicative distortion; nonparametric smoothing; predictors

References

[1] Şentürk, D. and Müller, H.-G. (2005). Covariate-adjusted regression. *Biometrika* **92** 75–89.

[2] Delaigle, A., Hall, P. and Zhou, W. (2016). Nonparametric covariate-adjusted regression. *Annals of Statistics*, to appear.

2.53 Nonparametric Tests for Conditional Symmetry

Miguel A. Delgado[1,*], Xiaojun Song[2]

[1] Universidad Carlos III de Madrid; delgado@est-econ.uc3m.es
[2] Peking University; sxj@gsm.pku.edu.cn

Abstract: We propose tests for symmetry of the conditional distribution of a time series process about a nonparametric regression function. The test statistic is the integrated squared difference between the restricted and unrestricted estimators of the joint characteristic function of nonparametric errors and explanatory variables with respect to a given weighting function, whose critical values are estimated with the assistance of a bootstrap technique. The test is sensitive to local alternatives converging to the null at the parametric rate $T^{-1/2}$. We investigate the finite sample performance of the test by means of Monte Carlo experiments and an empirical application to testing whether losses are more likely than gains in financial markets given the relevant information

Keywords: Conditional symmetry; Nonparametric testing; Smoothing; Time series data.

2.54 Nonparametric clustering of functional spatially dependent data

L. Delsol[1,*] and C. Louchet[1]

[1] MAPMO, University of Orléans; laurent.delsol@univ-orleans.fr, cecile.louchet@univ-orleans.fr

Abstract: It is nowadays quite common to face with functional and spatially dependent data (hyperspectral images, image time series, spatial time series, ...). An hyperspectral image, for instance, is a sample of functional data (a curve is associated to each pixel) presenting some spatial interactions coming form the geometric shape of the scene. The classification of such data into several homogeneous groups has to take into account these dependencies and take benefit from them. A recent approach, introduced by [2], combines functional kernel smoothing methods (see [1]) with standard bayesian methodologies - based on Gibbs-Markov random fields to mimic spatial interactions - to segment (i.e. to cluster into several groups) hyperspectral images. This new method depends on two parameters: h (the smoothing parameter fo the kernel estimate) and β (the grabularity parameter of the Potts random field used as prior). The aim of this talk is both to discuss the way these parameter should be chosen in practice (extending the idea of [3]) and explain how the methodology may be extended to deal with other functinal and spatially dependent data samples.

Keywords: Clustering; Functional data; Potts random field; Kernel smoothing; Segmentation.

References

[1] S. Dabo-Niang (2004). Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. *Applied Mathematics Letters*, **17** (4), pp. 381-386.

[2] L. Delsol, C. Louchet (2013). Segmentation of hyperspectral images from functional kernel density estimation. *International workshop on functional and operatorial statistics, Jun 2014, Stresa, Italy*. Societa Editrice Escalupio, pp.101-105, 2014.

[3] M. Pereyra, N. Dobigeon, H. Batatia and J. Y. Tourneret (2013). Estimating the Granularity Coefficient of a Potts-Markov Random Field Within a Markov Chain Monte Carlo Algorithm, *IEEE Transactions on Image Processing*, **vol. 22**, no. 6, pp. 2385-2397.

2.55 Detecting long-range dependence in non-stationary time series

H. Dette [1,*],P. Preuss [1] and K. Sen[1]

[1] Ruhr-Universität Bochum, Germany; holger.dett@rub.de

Abstract: An important problem in time series analysis is the discrimination between non-stationarity and long-range dependence. Most of the literature considers the problem of testing specific parametric hypotheses of non-stationarity (such as a change in the mean) against long-range dependent stationary alternatives. In this paper we suggest a simple approach, which can be used to test the null-hypothesis of a general non-stationary short-memory against the alternative of a non-stationary long-memory process. The test procedure works in the spectral domain and uses a sequence of approximating tvFARIMA models to estimate the time varying long-range dependence parameter. We prove uniform consistency of this estimate and asymptotic normality of an averaged version. These results yield a simple test (based on the quantiles of the standard normal distribution), and it is demonstrated in a simulation study that - despite of its semi-parametric nature - the new test outperforms the currently available methods, which are constructed to discriminate between specific parametric hypotheses of non-stationarity short- and stationarity long-range dependence.

Keywords: long-memory, non-stationary processes, goodness-of-fit tests, integrated periodogram, locally stationary process.

References

[1] Ehrenberg, A. C. S. (1982). Writing technical reports and papers. *The American Statistician*, **36**, 326–329.

[2] László Györfi and Michael Kohler and Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer. New-York.

[3] Lamport, L. (1986). *LaTeX A Document Preparation System*. Addison-Wesley. Boston.

2.56 Doubly Robust Survival Trees

Liqun Diao[1], Jon Steingrimsson[2], Annette Molinaro[3] and Robert Strawderman[4]

[1] University of Waterloo; liqun.diao@uwaterloo.ca
[2] Johns Hopkins University
[3] University of California, San Francisco
[4] University of Rochester

Abstract: Estimating a patient's mortality risk is important in making treatment decisions. Survival trees are a useful tool and employ recursive partitioning to separate patients into different risk groups. Existing "loss based" recursive partitioning procedures that would be used in the absence of censoring have previously been extended to the setting of right censored outcomes using inverse probability censoring weighted estimators of loss functions. We propose new "doubly robust" extensions of these loss estimators motivated by semiparametric efficiency theory for missing data that better utilize available data. Simulations and a data analysis demonstrate strong performance of the doubly robust survival trees compared to previously used methods.

Keywords: CART; Censored Data; Loss estimation; Inverse Probability of Censoring Weighted Estimation; Semi-parametric Estimation.

2.57 Optimal detection of weak principal components in high-dimensional data

Edgar Dobriban

Stanford University

Abstract: Principal component analysis is a widely used method for dimension reduction. In high dimensional data, the "signal" eigenvalues corresponding to weak principal components (PCs) do not necessarily separate from the bulk of the "noise" eigenvalues. In this setting, it is not possible to decide based on the largest eigenvalue alone whether or not there are "signal" PCs in the data. In this talk we explore this phenomenon in a general semiparametric model that captures the shape of eigenvalue distributions often seen in applications. We show how to construct statistical tests to detect principal components, based on all eigenvalues. We also explain how recent computational advances in random matrix theory enable the efficient implementation of our methods. The talk is based on two papers: arxiv.org/abs/1602.06896 [arxiv.org] and arxiv.org/abs/1507.01649 [arxiv.org].

2.58 Bootstrapping the Distribution of Tail Processes

R. A. Davis[1], H. Drees[2,*], J. Segers[3], M. Warchoł[3]

[1] Columbia University; rdavis@stat.columbia.edu
[2] University of Hamburg; drees@math.uni-hamburg.de
[3] Université catholique de Louvain; johan.segers@uclouvain.be, michal.warchol@uclouvain.be

Abstract: The extreme value dependence structure of a regularly varying stationary time series is captured by the so-called tail process, i.e. the limit of the standardized time series given that the observation at time 0 exceeds an increasingly high threshold. Using results from [1], one may prove asymptotic normality for estimators of the distribution of this tail process; cf. [2]. However, as the covariance function of the limiting Gaussian process is intricately influenced by the unknown dependence structure, the limit result cannot be directly used for the construction of confidence regions. Therefore, we propose a multiplier block bootstrap approach and show its consistency. Moreover, for popular models of financial time series, we analyze the finite sample performance of the resulting confidence regions.

Keywords: Extreme value dependence; Multiplier block bootstrap; Tail process.

References

[1] H. Drees and H. Rootzén (2010). Limit Theorems for Empirical Processes of Cluster Functionals. *The Annals of Statistics*, **38**, 2145–2186.

[2] H. Drees and J. Segers and M. Warchoł (2015). Statistics for Tail Processes of Markov Chains. *Extremes*, **18**, 369–402.

2.59 Generalized Seasonal Tapered Block Bootstrap

A. Dudek

IRMAR, Univ. Rennes 2, France; anna.dudek@univ-rennes2.fr

Abstract: Seasonality appears naturally in economics, vibroacoustics, mechanics, hydrology and many other fields. Periodicity is often present not only in the mean but also in the covariance function. Thus, to build statistical models periodically correlated (PC) processes are used. The purpose of the talk will be to present a new block bootstrap method designed for periodic time series, which is called the Generalized Seasonal Tapered Block Bootstrap (GSTBB). Consistency of the GSTBB for parameters associated with PC time series is shown; these are the overall mean, seasonal means and Fourier coefficients of the autocovariance function. Consequently, the construction of bootstrap pointwise and simultaneous confidence intervals for such parameters is possible.
Joint work with E. Paparoditis and D. Politis
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 655394.

Keywords: Block bootstrap; Confidence intervals; Consistency; Seasonal means; Autocovariance function.

References

[1] Dudek, A. E., Paparoditis, E. and Politis, D. Generalized Seasonal Tapered Block Bootstrap - submitted.

2.60 Monotone single index model

F. Balabdaoui[1], C. Durot[2,*] and H. Jankowski[3]

[1] Université Paris-Dauphine, Paris, France; fadoua.balabdaoui@gmail.com
[2] Université Paris Ouest Nanterre La Défense, Nanterre, France; cecile.durot@gmail.com
[3] Department of Mathematics and Statistics, York University, Toronto, Canada; hkj@mathstat.yorku.ca

Abstract: We consider single index models with monotone ridge function when the response variables have a distribution belonging to an exponential family. Under very mild assumptions, the index and the ridge function are identifiable, and both Maximum-likelihood and Least-squares estimator exist. They are characterized in terms of the least concave majorant of a cumulative sum diagram which is built from the data rearranged in an appropriate order. The MLE is consistent in a Hellinger sense. Rate of convergence and asymptotic behavior are studied (work in progress).

Keywords: Shape constraint, Single index model

References

[1] Ehrenberg, A. C. S. (1982). Writing technical reports and papers. *The American Statistician*, **36**, 326–329.

[2] László Györfi and Michael Kohler and Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer. New-York.

[3] Lamport, L. (1986). *LaTeX A Document Preparation System.* Addison-Wesley. Boston.

2.61 Spatial Functional Modeling of Weather Change Impact on Corn Yield in Kansas

J. Du[1,*], Z. Hao[1] and Z. Zhu[2]

[1] Kansas State University; dujuan@ksu.edu, haozheng@math.ksu.edu
[2] Iowa State University; zhuz@iastate.edu

Abstract: To study the impact of weather change on corn yield in Kansas and effectively utilize information available, we develop a spatial functional regression model accounting for the fact that weather data, such as temperatures and precipitation, are usually recorded daily or hourly as opposed to the crop yield data obtained on yearly basis. The underlying space/space-time autocorrelation structure is also explored. To deal with the collinearity among multiple covariates, the nonlinear functional relationship between temperature and yield is derived and incorporated into the proposed spatial functional model. Finally, the characteristics of the estimated temporally varying parameter function as well as the confidence band based on spline is investigated to show the influential factor and critical period of weather change affecting crop yield during growing season.

Keywords: B-splines; Functional linear model; Spatial covariance function.

2.62 On Perfect Classification for Gaussian Processes

Juan A. Cuesta-Albertos[1] and Subhajit Dutta[2,*]

[1] Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain; juan.cuesta@unican.es
[2] Department of Mathematics and Statistics, IIT Kanpur, India; duttas@iitk.ac.in

Abstract: We study the problem of discriminating J (≥ 2) Gaussian processes by analyzing the behavior of the underlying probability measures in an infinite-dimensional space. Motivated by singularity of a certain class of Gaussian measures, we first propose a data based transformation for the training data. For a J class classification problem, this transformation induces complete separation among the associated Gaussian processes. The misclassification probability of a componentwise classifier when applied on this transformed data asymptotically converges to zero. In finite samples,

the empirical classifier is constructed and related theoretical properties are studied. Good performance of the proposed methodology is demonstrated using simulated as well as benchmark data sets when compared with some parametric and nonparametric classifiers for such functional data.

Keywords: Bayes' risk; Consistent in probability; Cross-validation; Difference in covariance operators; Hajek and Feldman property; Mahalanobis' distances

References

[1] Ehrenberg, A. C. S. (1982). Writing technical reports and papers. *The American Statistician*, **36**, 326–329.

[2] László Györfi and Michael Kohler and Adam Krzyżak and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer. New-York.

[3] Lamport, L. (1986). *LaTeX A Document Preparation System.* Addison-Wesley. Boston.

2.63 Data-adaptive estimation of time-varying spectral densities

A. van Delft[1] and M. Eichler[1,*]

[1] Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands;
a.vandelft@maastrichtuniversity.nl, m.eichler@maastrichtuniversity.nl

Abstract: This paper introduces a data-adaptive approach for spectral density estimation of non-stationary processes. Estimation of time-dependent spectra commonly proceeds by means of local kernel smoothing. The performance of these non-parametric estimators depends however crucially on the smoothing bandwidths that need to be specified in both time and frequency direction. The objective of this paper is to construct local spectral density estimates where the respective smoothing kernels are iteratively adapted to the data at hand. The main idea, inspired by the concept of propagation-separation developed in Polzehl (2006), is to describe the largest local vicinity of every design point in the time-frequency plane over which smoothing is justified by the data. Our method circumvents the problem of optimal bandwidth selection in the strict sense without imposing additional assumptions. The procedure permits full flexibility for the degree of smoothing and automatically adjusts for structural breaks in the time-dependent spectrum.

Keywords: Local stationary processes; time-varying spectral density; data-adaptive kernel estimation.

References

[1] Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probability Theory and Related Fields*, **135**, 335-362.

2.64 A continuous updating weighted least squares estimator of tail dependence in high dimensions

J.H.J. Einmahl[1,*], A. Kiriliouk[2] and J. Segers[2]

[1] Tilburg University, Department of Econometrics & OR and CentER; j.h.j.einmahl@uvt.nl
[2] Univ. catholique de Louvain, Institut de Statistique; anna.kiriliouk@uclouvain.be, johan.segers@uclouvain.be

Abstract: Likelihood-based procedures are a common way to estimate tail dependence parameters. They are not applicable, however, in non-differentiable models such as those arising from recent max-linear structural equation models. Moreover, they can be hard to compute in higher dimensions. An adaptive weighted least-squares procedure matching nonparametric estimates of the stable tail dependence function with the corresponding values of a parametrically specified proposal yields a novel minimum-distance estimator. The estimator is easy to calculate and applies to a wide range of sampling schemes and tail dependence models. In large samples, it is asymptotically normal with an explicit and estimable covariance matrix. The minimum distance obtained forms the basis of a goodness-of-fit statistic whose asymptotic distribution is chi-square. Extensive Monte Carlo simulations confirm the excellent finite-sample performance of the estimator and demonstrate that it is a strong competitor to currently available methods. The estimator is then applied to disentangle sources of tail dependence in European stock markets.

Keywords: Brown–Resnick process; Extremal coefficient; Max-linear model; Multivariate extremes; Stable tail dependence function.

2.65 Hidden Markov chain change point estimation

ROBERT J. ELLIOTT AND SEBASTIAN ELLIOTT

University of Adelaide

Abstract: A hidden Markov model is considered where the dynamics of the hidden process change at a random 'change point'. In principle this gives rise to a non-linear filter but closed form recursive estimates are obtained for the conditional distribution of the hidden process and of the change point.

2.66 Nonparametric Tolerance Tubes for Functional Data

Y. Fan[1*] and R. Liu[1]

[1] Department of Statistics and Biostatistics, Rutgers University, USA; yifan@stat.rutgers.edu,rliu@stat.rutgers.edu

Abstract: Tolerance intervals and tolerance regions are important tools for statistical quality control and process monitoring of univariate and multivariate data, respectively. The goal of this paper is to generalize the tolerance intervals/regions to tolerance tubes in the infinite dimensional setting for functional data. In addition to the generalizations of the commonly accepted definitions of the tolerance level of $\beta-$content or $\beta-$expectation, we introduce a modification of $\beta-$expectation tolerance tube by coupling it with an exempt level α. The latter loosens the definition of $\beta-$expectation tolerance tube by allowing α (usually pre-set by domain experts) portion of each functional be exempt from the requirement. More specifically, for a sample of n functional data, a $\beta-$expectation tolerance tube with exempt level α is expected to contain $n\beta$ functionals in such a way that at least $(1-\alpha)\times 100\%$ portion of each functional is contained within the limit of the tube.

Those proposed tolerance tubes are completely nonparametric and thus broadly applicable. We investigate their theoretical justifications and properties. We also show that the exempt $\beta-$expectation tolerance tube is particularly useful in the setting where occasional short term aberrations of the functional data are deemed acceptable if those aberrations do not cause substantive deviation from the norm. This desirable property is elaborated and illustrated further with both simulations and real applications in continuous monitoring of blood glucose level in diabetes patients as well as of aviation risk pattern during aircraft landing operations.

Keywords: Functional data analysis; tolerance tubes; data depth; exempt level

2.67 Optimization-driven Supervised Dimension Reduction

G. Felici[1,*] F. Chiaromonte[2,3] and Yang Liu[2]

[1] Consiglio Nazionale delle Ricerche – IASI; giovanni.felici@iasi.cnr.it
[2] The Pennsylvania State University; fxc11@psu.edu, ywl5222@psu.edu
[3] Sant′Anna School of Advanced Studies

Abstract: As very high dimensional data sets become more common, statistical techniques for regularization, feature selection, and dimension reduction have become critical for contemporary regression applications. We start this talk presenting an approach that, leveraging tools from Mixed Integer Optimization, allows one to very effectively and flexibly select discrete features in classification problems [1]. Next, we describe how this approach can be extended to regression problems comprising continuous features and responses, and draw some connections with penalization-based techniques such as the LASSO and Elastic Nets [2**?**]. We then show how our approach can be combined with Sufficient Dimension Reduction (SDR). The recent development of group-wise techniques [4**?**] allows SDR to be applied to data characterized by group structures in terms of subpopulations and/or predictor domains, but their effectiveness is still limited when the number of predictors is very large compared to the number of available observations. Our feature selection approach can be effectively used as a preliminary filter prior to the application of SDR, or to construct groups of predictors based on available auxiliary information when the groups are not known in advance.

Keywords: Supervised Dimension Reduction; Feature Selection; Regression; Mixed Integer Programming

References

[1] P. Bertolazzi, G. Felici, P. Festa, G. Fiscon, and E. Weitschek (2015). Integer Programming models for Feature Selection: new extensions and a randomized solution algorithm. *European Journal of Operational Research*, **250**, 389–399.

[2] R. Tibshirani (1995). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.***58**, 267-288.

[3] H. Zou and T Hastie (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 301-320.

[4] L. Li, B. Li, and X (2010). Zhua. Groupwise Dimension Reduction. *Journal of the American Statistical Association* **105 491**, 1188-1201.

[5] Z. Guo, L. Li, W., and B. Li (2015). Groupwise Dimension Reduction with Envelope Methods. *Journal of the American Statistical Association*, **110 512**, 1515-1527.

2.68 Comparing ordered trees: Current status and open questions

A. Feragen

Department of Computer Science, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Abstract: Tree-structured data are abundant in anatomy, where they represent delivery systems for blood, air, signals, etc. As opposed to phylogenetic trees, where a significant statistical literature has built up over recent years, far less is known about the statistical and geometric tools available for analyzing general anatomical trees. Whereas phylogenetic trees have a pre-fixed set of leaf labels, which enable the construction of a phylogenetic tree-space with polynomial time shortest path computations, most anatomical trees do not have known leaves or even natural labels. This complicates both the formulation of a proper mathematical framework for anatomical trees, as well as the computational complexity of standard algorithms given an attempt to formulate such a framework. Some anatomical trees grow on the surface of organs, which leads to a well-defined planar ordering of their branches. This may lead to a simplification of algorithms such as shortest-paths, which may, in turn, simplify the development of scalable statistical algorithms. In this talk we shall survey the current status of geometric spaces of planar trees, and discuss open problems.

Keywords: Tree-space statistics; Planar trees; Anatomical trees.

2.69 Convolution models on netorks

J-P Florens

Toulouse School of Economics

Abstract: The paper considers linear models between functional variables. If Y and Z are elements of an Hilbert space $\mathcal{H}$, general linear models take the form $Y = \Pi Z + U$ where Π is an operator from $\mathcal{H}$ to $\mathcal{H}$. In order to reduce the dimension of the problem we want consider cases where ΠZ reduces to $Z * a$ where a is in $\mathcal{H}$ and $*$ is a convolution operator. The main question is to define this convolution and this will be done through the introduction of a Laplacian operator on $\mathcal{H}$. The spectral family of this Laplacian may be used to define a Fourier analysis on $\mathcal{H}$ and then a convolution operator in the spirit of previous works of P. Vandergheynst and co-authors. Finite (large) networks or manifolds give examples of this situation. In econometric application endogeneity questions are treated and the model is estimated under a convolution instrumental assumption.

2.70 Goodness-of-fit tests for log-volatility GARCH models

C. Francq[1,*], O. Wintenberger[2] and J-M. Zakoïan[1]

[1] CREST and Univ. of Lille, France; christian.francq@univ-lille3.fr, zakoian@ensae.fr
[2] Universities of Paris 6 and Copenhaguen, LSTA Paris, France; olivier.wintenberger@upmc.fr

Abstract: This paper studies goodness of fit tests and specification tests for an extension of the log-GARCH model which is stable by scaling. A Lagrange-Multiplier test is derived for testing the null assumption of extended log-GARCH against more general formulations including the Exponential GARCH (EGARCH). The null assumption of an EGARCH is also tested. Portmanteau goodness-of-fit tests are developed for the extended log-GARCH. Simulations illustrating the theoretical results and an application to real financial data are proposed.

Keywords: EGARCH; LM tests; Invertibility of time series models; log-GARCH; Portmanteau tests

2.71 On the residual-based bootstrap for functional autoregressions

J. Franke[1,*] and E. Nyarige[1]

[1] Dept. of Mathematics, University of Kaiserslautern; franke@mathematik.uni-kl.de, nyarige@mathematik.uni-kl.de

Abstract: Functional autoregressions (FAR), also known as autoregressive Hilbertian processes, of order 1 are among the most popular models for time series of functional data in a Hilbert space $\mathcal{H}$, e.g. of curves in some L^2-space :

$$X_{t+1} = \Psi(X_t) + \epsilon_{t+1},$$

where the innovations ϵ_t are i.i.d. random variables in $\mathcal{H}$. FARs are frequently used for forecasting or tests, e.g. for changes in the structure of the data generating process, compare, e.g., [3]. In many practical applications, resampling methods are used for calculating approximate prediction intervals or critical values for tests. However, the theoretical basis for those methods is still rather incomplete. We consider the residual-based, also called naive, bootstrap where the bootstrap data are generated by

$$X^*_{t+1} = \hat{\Psi}_n(X^*_t) + \epsilon^*_{t+1},$$

where $\hat{\Psi}_n$ is an estimate of Ψ and $\epsilon^*_1, \ldots, \epsilon^*_n$ are i.i.d., conditional on the original data, with distribution given by the empirical distribution of the centered sample residuals $X_t - \hat{\Psi}_n(X_{t-1}), t = 1, \ldots, n$. This resampling method is quite popular in the scalar and multivariate case, and, there, it forms the starting point for the widely applicable autoregressive sieve bootstrap, compare [5]. For functional data, this kind of resampling has already been investigated for functional linear regression models by [2], but for autoregressions, the situation is more complicated such that we have to follow a different strategy for proving the asymptotic validity of the residual-based bootstrap which uses ideas from the scalar autoregressive case as in [4]. In this talk we present first results, starting from the convergence of the empirical distribution of the sample residuals to the true distribution of the innovations which can be shown using the strong consistency result of [1] for the estimate $\hat{\Psi}_n$ of the autoregressive operator Ψ. This result is the key to proving the approximability of the distributions of the sample mean and the sample covariance and lag-1 autocovariance operators by means of the residual-based bootstrap.

Keywords: Bootstrap; Functional data; Autoregression; Time series

References

[1] Bosq D. (2010). *Linear Processes in Function Spaces.* Springer. Berlin-Heidelberg-New York.

[2] González-Manteiga, W. and Martínez-Calvo, A. (2011). Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference*, **141**, 453–461.

[3] Horváth, L. and Kokoszka P. (2010). *Inference for Functional Data with Applications.* Springer. Berlin-Heidelberg-New York.

[4] Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving average models. *Journal of Time Series Analysis*, **13**, 297–317.

[5] Kreiss, J.-P. and Paparoditis, E. (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, **40**, 357–378.

2.72 High dimensional model selection with an adaptive ridge procedure for L_0 regularization

F. Frommlet[1]

[1] Medical University Vienna, Section of Medical Statistics; Florian.Frommlet@meduniwien.ac.at

Abstract: Penalized selection criteria like AIC or BIC are among the most popular methods for variable selection. Their theoretical properties have been studied intensively and are well understood in case of a moderate number of variables. However, these criteria do not work well in a high-dimensional setting under the assumption of sparsity. We will introduce different modifications of AIC and BIC [1**?**] which will allow to control the family wise error rate or false discovery rate, respectively, in terms of including false positive regressors in the model. After briefly discussing theoretical properties of these modified criteria [3] we will discuss the difficult task of model search. We will focus on the recently introduced adaptive ridge approach [4**?**], where iteratively weighted ridge problems are solved whose weights are updated in such a way that the procedure converges towards selection with L0 penalties.

Keywords: Model selection; High dimensional; Information Criteria; Sparsity; Regularization

References

[1] Bogdan, M., Ghosh, J.K., Zak-Szatkowska, M.(2008). Selecting explanatory variables with the modified version of Bayesian Information Criterion. *Quality and Reliability Engineering International*, 24, 627 - 641.

[2] Frommlet, F., Ruhaltinger, F., Twarog, P., Bogdan, M.(2012). Modified versions of Bayesian Information Criterion for genome-wide association studies. *CSDA*, 56, 1038 - 1051.

[3] Frommlet, F., Bogdan, M. (2013). Some optimality properties of FDR controlling rules under sparsity. *Electronic Journal of Statistics*, Vol. 7, No. 0, 1328-1368.

[4] Rippe RC, Meulman JJ, Eilers PH (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS ONE*, 7: e38230.

[5] Frommlet, F., Nuel, G. (2016). An adaptive Ridge procedure for L_0 regularization. *PLoS ONE*, In Print.

2.73 Bootstrap and permutation independence tests for point processes, with applications in neuroscience

M. Albert[1], Y. Bouret[2], M. Fromont[3,*] and P. Reynaud-Bouret[4]

[1] Univ. Grenoble Alpes, UMR 5216, France; melisande.albert@gipsa-lab.grenoble-inp.fr
[2] Univ. Nice Sophia Antipolis, CNRS, LPMC, UMR 7336, France; yann.bouret@unice.fr
[3] Univ. Rennes 2, IRMAR,France; magalie.fromont@univ-rennes2.fr
[4] Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, France; reynaudb@unice.fr

Abstract: With a view to a better understanding of the neural code, detecting the spikes synchronization phenomenon appears as a fundamental issue in neuroscience, treated by many authors (see for instance [2], [1], [3]). Since no parametric model for spike trains has been commonly accepted by neuroscientists, we consider this issue as a nonparametric problem of testing independence between two point processes. We propose new tests based on bootstrap or permutation approaches. Without any constraining assumption on the distribution of the observed processes, we obtain general consistency results with respect to Wasserstein's metric for both approaches. These results allow us to prove that our tests are asymptotically of the prescribed size and consistent against any reasonable alternative, the permutation test having the further advantage to be exactly of the prescribed level. An experimental study is then performed to illustrate this theoretical study, and to compare the performance of the present tests with existing ones in the neuroscientific literature. Finally, in order to detect the precise locations of synchronization, the permutation test is integrated in a multiple testing procedure, which is applied to simulated and real data.

Keywords: Independence test; U-statistic; Point process; Bootstrap; Permutation

References

[1] Grün, S., Diesmann, M., and Aertsen, A. M. (2010). *Analysis of parallel spike trains,* chapter *Unitary Events analysis*. Springer Series in Computational Neuroscience.

[2] Pipa, G. and Grün, S. (2003). Non-parametric significance estimation of joint-spike events by shuffling and resampling. *Neurocomputing*, **52–54**, 31–37.

[3] Tuleau-Malot, C., Rouis, A., Grammont, F., and Reynaud-Bouret, P. (2014). Multiple tests based on a Gaussian approximation of the Unitary Events method. *Neural Computation*, **26(7)**.

2.74 Tail-greedy bottom-up data decompositions and fast multiple change-point detection

P.Fryzlewicz[1,*]

[1] Department of Statistics, London School of Economics, UK; p.fryzlewicz@lse.ac.uk

Abstract: We propose a 'tail-greedy', bottom-up transform for one-dimensional data, which results in a nonlinear but conditionally orthonormal, multiscale decomposition of the data with respect to an adaptively chosen Unbalanced Haar wavelet basis. The 'tail-greediness' of the decomposition algorithm, whereby multiple greedy steps are taken in a single pass through the data, both enables fast computation and makes the algorithm applicable in the problem of consistent estimation of the number and locations of multiple change-points in data. The resulting agglomerative change-point detection method avoids the disadvantages of the classical divisive binary segmentation, and offers very good practical performance. Details appear in [1].

Keywords: Tail-greediness; Bottom-up methods; Multiscale methods, Multiple change-point detection, Thresholding.

References

[1] Fryzlewicz, P. (2016). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Preprint.*

2.75 Inference and resampling for non-stationary, long memory and heavy tails time series

J. Leśkow[1], E. Gajecka-Mirek[2,*]

[1] Institute of Mathematics, Cracow University of Technology, Poland; jleskow@pk.edu.pl,
[2] Institute of Economic, State Higher Vocational School, Nowy Sącz, Poland; egajecka@gmail.com

Abstract: Statistical inference based on asymptotic distributions in the case of dependent data is very often ineffective. In the last three decades there was a significant development of the resampling methods. Using these methods, one can efficiently approximate the sampling distributions of statistics and estimators.
A problem that needs to be solved during such study is the consistency of the proposed procedure, that is its ability to capture quantiles of the unknown sampling distribution of the statistic at hand. In our research, we are motivated by the weak dependence structure, defined by P. Doukhan [1].
In the talk a time series model with three specific features: long memory, heavy tails and a periodic structure will be considered.

Using weak dependence conditions we show the consistency theorem for the resampling method. Specifically, we study the subsampling for the estimator of the mean function will be presented. Additionally, the real data example will be given.

Keywords: Long range dependence; Heavy tails; Weak dependence; Periodicity; Subsampling.

References

[1] Dedecker J., Doukhan P., Lang G., León J. , Louhichi S., Prieur C. (2008). *Weak Dependence*. L. N. 190 in S., Springer-Verlag

[2] Doukhan P., Prohl S., Robert C. Y. (2011). Subsampling weakly dependent times series and application to extremes. *TEST*, **20**, 487–490.

[3] Jach A., McElroy T., Politis D. N. (2012). Subsampling inference for the mean of heavy-tailed long memory time series. *J. Time Ser. Anal.*, **33**, 96–111.

2.76 Depth-based nonparametric tests for homogeneity of functional data

G. Geenens[1,*]

[1] School of Mathematics and Statistics, UNSW Australia, Sydney (Australia); ggeenens@unsw.edu.au

Abstract: In this work we study some tests for the homogeneity between two independent samples of functional data. The null hypothesis of 'homogeneity' here means that the latent stochastic processes which generated the two samples are actually identically distributed. Because it is generally not possible to define a probability density for functional data, it seems natural to opt for nonparametric procedures in this setting. Making use of recent developments on functional depths, we adapt some Kolmogorov-Smirnov- and Cramer-von-Mises-type of criteria to the functional context. Exact p-values for the test can be obtained via permutations, or, in case of too large samples, a bootstrap algorithm is easily implemented. A simulation study illustrates how powerful the devised methodology is, and some real data examples are analysed. Finally, the extension to the case of more than two samples is discussed.

Keywords: FDA; Functional Depth; Two-sample test; Homogeneity; Permutation tests.

2.77 A nonparametric copula approach for clustered right-censored event time data

C. Geerdens[1,*], P. Janssen[1] and N. Veraverbeke[1]

[1] Center for Statistics, Universiteit Hasselt, Hasselt, Belgium; candida.geerdens@uhasselt.be, paul.janssen@uhasselt.be, noel.veraverbeke@uhasselt.be

Abstract: In survival analysis interest is in the time until a predefined event (e.g., the time to blindness). Often, this event time is right-censored for some items in the study sample, i.e., only a lower time bound for the event is observed (e.g., due to the end of the study period). A further complexity can be the grouping of data (e.g., the time to blindness is registered for both left and right eye). Since clustered items share common traits, their event times are associated. Copulas provide a popular tool to describe the association in grouped time-to-event data. In a data setting where it is less evident to predetermine a parametric copula, such as the one of right-censored data, one may opt to apply a nonparametric copula. We define a new nonparametric copula estimator for the joint survival function of grouped right-censored event time data. In here, we consider two right-censoring schemes: univariate censoring and copula censoring. For the new nonparametric copula estimator, we establish the consistency and we assess the finite sample performance in various data settings via a simulation study. Focus is on the overall performance and the behavior in the extremal points of the unit square. A comparison with the recent nonparametric copula estimator of Gribkova and Lopez (2015) is given.

Keywords: Nonparametric estimator; Right-censored data; Copula

References

[1] Geerdens, C., Janssen P. and Veraverbeke, N. (2015). Large sample properties of nonparametric copula estimators under bivariate censoring. *Statistics*, doi: 10.1080/02331888.2015.1119149.

[2] Gribkova, S. and Lopez, O. (2015). Nonparametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics*, doi: 10.1111/sjos.12144.

2.78 Median-based Nonparametric Confidence Regions for the Central Orientation of Random Rotations

U. Genschel[1,*] and B. Stanfill[2]

[1] Department of Statistics, Iowa State University; ulrike@iastate.edu
[2] Pacific Northwest National Laboratory; bryan.stanfill@pnnl.gov

Abstract: Three-dimensional orientation data, with observations as 3×3 rotation matrices, have applications in many areas such as computer science, kinematics or materials sciences where it is often of interest to nonparametrically estimate a central orientation parameter $\boldsymbol{S}$ represented by a 3×3 rotation matrix. Although $SO(3)$ is bounded, data contamination still occurs when observations are sufficiently far from $\boldsymbol{S}$ adversely affecting estimation results of mean-based estimators of $\boldsymbol{S}$. The projected median (Fletcher et al. (2009) and Stanfill et al. (2013)) is an alternative, more robust estimator but its performance has only been evaluated empirically through simulation. We show that the projected median is a strongly consistent estimator for the central orientation $\boldsymbol{S}$ and is asymptotically normally distributed. Using these results we propose a large-sample normal theory-based method for setting confidence regions for $\boldsymbol{S}$ in addition to a bootstrap-based approach for approximating the sampling distribution of the projected median statistic and calibrating nonparametric confidence regions for $\boldsymbol{S}$. In the presence of contamination we show that both confidence regions for the central orientation based on the median achieve a smaller size and closer to nominal coverage rates compared to those based on the projected arithmetic mean.

Keywords: Orientation Data; Pivotal Bootstrap; Projected Arithmetic Mean; Projected Arithmetic Mean.

References

[1] Fletcher, P., Venkatasubramanian, S. and Joshi, S. (2009). The Geometric Median on Riemannian Manifolds with Application to Robust Atlas Estimation. *NeuroImage*, **45**(1), S143–S152.

[2] Stanfill, B., Genschel, U. and Hofmann, H. (2013). Point Estimation of the Central Orientation of Random Rotations. *Technometrics*, **55**(4), 524–535.

2.79 Nonparametric lack-of-fit test of nonlinear regression in presence of heteroscedastic variances

M. Gharaibeh[1] and H. Wang[2,*]

[1] Department of Mathematics, Al al-Bayt University, Mafraq 25113, Jordan; mhmd78@ksu.edu
[2] Department of Statistics, Kansas State University, Manhattan, KS 66506; hwang@ksu.edu

Abstract: In this paper, a nonparametric lack-of-fit test of nonlinear regression in presence of heteroscedastic variances is proposed. We consider regression models with a discrete or continuous response variable without distributional assumptions so that the test is widely applicable. The test statistic is developed using a k-nearest neighbor augmentation defined through the ranks of the predictor variable. The asymptotic distribution of the test statistic is derived under the null and local alternatives for the case of using fixed number of nearest neighbors. The parametric standardizing rate is achieved for the asymptotic distribution of the proposed test statistic. This allows the proposed test to have faster convergence rate than most of nonparametric methods. Numerical studies show that the proposed test has good power to detect both low and high frequency alternatives even for moderate sample size. The proposed test is applied to an engineering data example.

Keywords: Lack-of-fit; Hypothesis testing; k-nearest neighbor.

2.80 Some exact distribution-free one sample tests for high dimension, low sample size data

M. Biswas, M. Mukhopadhyay and A. K. Ghosh*

Abstract: Several rank-based tests for the multivariate one-sample problem are available in the literature. But, unlike univariate rank-based tests, most of these multivariate tests are not distribution-free. Moreover, many of them are not applicable when the dimension of the data exceeds the sample size. We develop and investigate some distribution-free

tests for the one-sample location problem, which can be conveniently used in high dimension low sample size (HDLSS) situations. Under some appropriate regularity conditions, we prove the consistency of these tests when the sample size remains fixed and the dimension grows to infinity. Some simulated and real data sets are analyzed to compare their performance with some popular one-sample tests.

Keywords:

2.81 Oracle convergence rates for Bayesian density regression in high dimensional spaces

W. Shen[1] and S. Ghoshal[2] *

[1] University of California-Irvine; swn1989@gmail.com, steven.smith@mail.com
[2] North Carolina State University; ghoshal@stat.ncsu.edu

Abstract: Density regression provides a completely flexible nonparametric approach to study the dependence of the distribution of a response variable Y on a p-dimensional covariate X. The complexity of the problem is equivalent to estimating the joint density of Y and X, which is a $(p+1)$-dimensional nonparametric estimation problem. The optimal convergence rate of convergence for estimating α-smooth functions is $n^{-\alpha/(2\alpha+p+1)}$, which is incredibly slow in high dimension. However, in high dimensional spaces, usually relations are sparse, meaning that the distribution of Y may actually depend on a much fewer, d number of covariates. The oracle procedure using the additional knowledge of these predictors, can construct an estimator that will converge at the much faster rate $n^{-\alpha/(2\alpha+d+1)}$. We develop a nonparametric Bayesian procedure to achieve the oracle rate up to a logarithmic factor even when the dimension p can be exponentially large compared with the available sample size n. The prior is constructed by expanding the density function in the basis of tensor product of B-splines, and putting appropriate priors on both the number of terms and the coefficients, as well as on the subsets of predictors that enter the model. We also show that the posterior mean can be computed without resorting to MCMC methods if we choose a Dirichlet distribution on the coefficients of the basis expansion.

Keywords: Density regression; B-splines; High Dimension; Oracle rate; Posterior convergence.

2.82 On kernel smoothing with Gaussian subordinated spatial data

Sucharita Ghosh[1]

[1] Swiss Federal Research Institute WSL, Birmensdorf, Switzerland; rita.ghosh@wsl.ch

Abstract: We consider a nonparametric regression model with spatial observations and the Priestley-Chao kernel estimator for the regression surface. Suppose that the centered observations are Gaussian subordinated via a location dependent possibly non-linear transformation of a latent Gaussian random field. Using kernels that have absolutely integrable characteristic functions, we discuss some consistency properties and propose a direct estimator for the variance of the mean surface estimator that relies on a locally stationary type property of the errors. An excerpt from a Total Column Ozone data set (source: NASA) is used to motivate the problem.

Keywords: Smoothing; Long memory; Characteristic function.

2.83 Testing Mean Stability of Heteroskedastic Time Series

V. Dalla[1], L. Giraitis[2,*] and PCB. Phillips[3]

[1] National and Kapodistrian University of Athens; violetta.dalla@econ.uoa.gr
[2] Queen Mary, University of London; L.Giraitis@qmul.ac.uk
[3] Yale University, University of Auckland, University of Southampton, Singapore Management University; peter.phillips@yale.edu

Abstract: Time series models are often fitted to the data without preliminary checks for stability of the mean and variance, conditions that may not hold in much economic and financial data, particularly over long periods. Ignoring such shifts may result in fitting models with spurious dynamics that lead to unsupported and controversial conclusions about time dependence, causality, and the effects of unanticipated shocks. In spite of what may seem as obvious differences between a time series of independent variates with changing variance and a stationary conditionally heteroskedastic (GARCH) process, such processes may be hard to distinguish in applied work using basic time series diagnostic tools. We develop and study some practical and easily implemented statistical procedures to test the mean and variance

stability of uncorrelated and serially dependent time series. Application of the new methods to analyze the volatility properties of stock market returns leads to some unexpected surprising findings concerning the advantages of modeling time varying changes in unconditional variance.

Keywords: Heteroskedasticity; KPSS test; Mean stability; Variance stability; VS test.

2.84 Estimation of the functional Weibull tail-coefficient

S. Girard[1,*] and L. Gardes[2]

[1] Inria Grenoble Rhône-Alpes & LJK, France; Stephane.Girard@inria.fr
[2] Université de Strasbourg & CNRS, IRMA, UMR 7501, Strasbourg, France; gardes@unistra.fr

Abstract: We present a nonparametric family of estimators for the tail index of a Weibull tail-distribution when functional covariate is available. Our estimators are based on a kernel estimator of extreme conditional quantiles, extending a previous work [1] to the infinite dimensional case. Asymptotic normality of the estimators is proved under mild regularity conditions. Their finite sample performances are illustrated both on simulated and real data. We refer to [2] for further details.

Keywords: Conditional Weibull tail-coefficient; Extreme quantiles; Nonparametric estimation.

References

[1] Daouia, A., Gardes, L., Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, **19**, 2557–2589.

[2] Gardes, L., Girard, S. (2016). On the estimation of the functional Weibull tail-coefficient, *Journal of Multivariate Analysis*, to appear, http://dx.doi.org/10.1016/j.jmva.2015.05.007

2.85 Classification methods for Hilbert data based on surrogate density

Aldo Goia

Univ. da Novara, Italy

Abstract: We study classification approaches for Hilbert random curves resting on the use of a surrogate of the probability density which is defined, in a distribution-free mixture context, from an asymptotic factorization of the small-ball probability. The latter is rigorously established exploiting the Karhunen-Loéve expansion whose basis turns out to be the optimal one in controlling the approximation errors. That surrogate density is estimated by a kernel approach from the principal components of the data. The remaining part of the work focuses on the illustration of the classification algorithms and the computational implications, with particular attention to the tuning of parameters involved. Some asymptotic results are sketched. Applications on simulated and real datasets show how the proposed methods work.

2.86 Optimal inference in the sparse additive model

Karl Gregory, Enno Mammen, Martin Wahl

Abstract: We consider the construction of pointwise confidence intervals for a single function in the additive nonparametric regression model, in which the conditional mean of the response is the sum of a large number of covariate effects of unspecified form. We allow the number of covariates to grow along with the sample size while assuming that only a small but growing number of the covariates have an influence on the response. Estimation in this setting is well-studied, but there exists almost no machinery for inference, as estimators typically involve Lasso penalization and have very complicated distributions. We introduce an estimator which is asymptotically normal by adapting "desparsified Lasso" techniques recently introduced in the linear regression setting to the nonparametric regression setting. Moreover, we develop a two-step presmoothing-resmoothing estimator which yields asymptotically optimal pointwise confidence intervals for a single function in the sense that our estimator achieves, asymptotically, up to first order terms, the same bias and variance as the oracle estimator, for which only the function of interest is unknown.

2.87 Kernel Statistical Tests for Random Processes

K. Chwialkowski,[1] D. Sejdinovic,[2] and A. Gretton[1*]

[1] University College London; kacper.chwialkowski@gmail.com, arthur.gretton@gmail.com
[2] University of Oxford; dino.sejdinovic@gmail.com

Abstract: We propose a kernel approach to statistical testing for random processes, which we illustrate using two settings: a two-sample test (comparison of marginal distributions of random processes), and an independence test. Our test statistics are in both cases straightforward V-statistics, derived from embeddings of probability measures to a reproducing kernel Hilbert space (RKHS) [**?**]. We use both shuffling [**?**] and wild bootstrap [**?**] approaches to construct provably consistent tests, and demonstrate that these are successful for cases where a naive permutation-based bootstrap fails. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler.

Keywords: Hypothesis Test; Reproducing Kernel Hilbert Space; Two-Sample Test; Independence Test

2.88 Simulation based Bias Correction Methods for Complex Models

Stéphane Guerrier[1], Elise Dupuis-Lozeron[2], Yanyuan Ma[3] and Maria-Pia Victoria-Feser[2]

[1] Department of Statistics, University of Illinois at Urbana-Champaign, USA.
[2] Research Center for Statistics, University of Geneva, Switzerland.
[3] Department of Statistics, University of South Carolina, USA.

Abstract: Along the ever increasing data size and model complexity, an important challenge frequently encountered in constructing new estimators or in implementing a classical one such as the maximum likelihood estimator, is the computational aspect of the estimation procedure. To carry out estimation, approximate methods such as pseudo-likelihood functions or approximated estimating equations are increasingly used in practice as these methods are typically easier to implement numerically although they can lead to inconsistent and/or biased estimators. In this context, we extend and provide refinements on the known bias correction properties of two simulation based methods, respectively indirect inference and bootstrap, each with two alternatives. These results allow one to build a framework defining simulation based estimators that can be implemented for complex models. Indeed, based on a biased or even inconsistent estimator, several simulation based methods can be used to define new estimators that are both consistent and with reduced finite sample bias. This framework includes the classical method of indirect inference for bias correction without requiring specification of an auxiliary model. We demonstrate the equivalence between one version of the indirect inference and the iterative bootstrap, both correct sample biases up to the order n^{-3}. The iterative method can be thought of as a computationally efficient algorithm to solve the optimization problem of the indirect inference. Our results provide different tools to correct the asymptotic as well as finite sample biases of estimators and give insight on which method should be applied for the problem at hand. The usefulness of the proposed approach is illustrated with the estimation of robust income distributions and generalized linear latent variable models.

Keywords: Iterative bootstrap; Two-step estimators; Indirect inference; Robust statistics; Weighted maximum likelihood estimators; Generalized latent variable models.

2.89 High Dimensional Stochastic Regression with Latent Factors, Endogeneity and Nonlinearity

Jinyuan Chang[1], Bin Guo[2] and Qiwei Yao[3,*]

[1] Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia 3010; jinyuan.chang@unimelb.edu.au
[2] School of Economics, Sichuan University, Chengdu, China; guobinscu@scu.edu.cn
[3] Department of Statistics, London School of Economics, London, WC2A 2AE, U.K. and Guanghua School of Management, Peking University, Beijing, China; q.yao@lse.ac.uk

Abstract: We consider a multivariate time series model which represents a high dimensional vector process as a sum of three terms: a linear regression of some observed regressors, a linear combination of some latent and serially correlated factors, and a vector white noise. We investigate the inference without imposing stationary conditions on the target multivariate time series, the regressors and the underlying factors. Furthermore we deal with the the endogeneity that

there exist correlations between the observed regressors and the unobserved factors. We also consider the model with nonlinear regression term which can be approximated by a linear regression function with a large number of regressors. The convergence rates for the estimators of regression coefficients, the number of factors, factor loading space and factors are established under the settings when the dimension of time series and the number of regressors may both tend to infinity together with the sample size. The proposed method is illustrated with both simulated and real data examples.

Keywords: α-mixing, dimension reduction, instrument variables, nonstationarity, time series

2.90 Generalized Fiducial Inference for Non-Parametric Problems

Jan Hannig[1*], Yifan Cui[1]

[1] University of North Carolina at Chapel Hill jan.hannig@unc.edu, cuiy@live.unc.edu

Abstract: R. A. Fisher, the father of modern statistics, proposed the idea of fiducial inference during the first half of the 20th century. While his proposal led to interesting methods for quantifying uncertainty, other prominent statisticians of the time did not accept Fisher's approach as it became apparent that some of Fisher's bold claims about the properties of fiducial distribution did not hold up for multi-parameter problems. Beginning around the year 2000, the authors and collaborators started to re-investigate the idea of fiducial inference and discovered that Fisher's approach, when properly generalized, would open doors to solve many important and difficult inference problems. They termed their generalization of Fisher's idea as generalized fiducial inference (GFI). The main idea of GFI is to carefully transfer randomness from the data to the parameter space using an inverse of a data generating equation without the use of Bayes theorem. The resulting generalized fiducial distribution (GFD) can then be used for inference. After more than a decade of investigations, the authors and collaborators have developed a unifying theory for GFI, and provided GFI solutions to many challenging practical problems in different fields of science and industry. Overall, they have demonstrated that GFI is a valid, useful, and promising approach for conducting statistical inference. In this paper we apply the GFI paradigm to non-parametric function estimation setting. We obtain GFD for survival function in the right censoring case and show that it satisfies Bernstein-von Mises theorem. The application of GFI in other similar context will be also discussed.

Keywords: Generalized Fiducial Inference, Fusion Learning, Kaplan-Meyer Estimator, Confidence Distribution.

2.91 Nonparametric goodness of fit via cross-validation Bayes factors

Jeffrey D. Hart[1,*] and Taeryon Choi[2]

[1] Department of Statistics, Texas A&M University; hart@stat.tamu.edu
[2] Department of Statistics, Korea University; trchoi@gmail.com

Abstract: A nonparametric Bayes procedure is proposed for testing the fit of a parametric model for a distribution. Alternatives to the parametric model are kernel density estimates. Data splitting makes it possible to use kernel estimates for this purpose in a Bayesian setting. A kernel estimate indexed by bandwidth is computed from one part of the data, a training set, and then used as a model for the rest of the data, a validation set. A Bayes factor is calculated from the validation set by comparing the marginal for the kernel model with the marginal for the parametric model of interest. A simulation study is used to investigate how large the training set should be, and examples involving astronomy and wind data are provided. A theorem on Bayes consistency of the proposed test is also given.

Keywords: Bandwidth selection; Bayes consistency; Dirichlet processes; Kernel density estimates; Polya trees.

2.92 Ruin Probability & Ruin Time Approximation for γ-reflected Gaussian Risk Models with Tax

K. Dębicki[1], E. Hashorva[2] and P. Liu[2]

[1] Mathematical Institute, University of Wrocław,Poland; Krzysztof.Debicki@math.uni.wroc.pl
[2] University of Lausanne, Switzerland; enkelejd.hashorva@unil.ch, peng.liu@unil.ch

Abstract: In this talk we are concerned with approximations of the ruin probability and the ruin times (first and last one) in γ-reflected Gaussian risk models with tax. A particular instance is the fractional Brownian motion risk model with tax. Our solutions are motivated by a reformulation of the problem in terms of extremes of certain Gaussian random fields.

Keywords: Ruin probability; ruin time; Gaussian process; fractional Brownian motion; ruin time approximation.

2.93 Asymptotic Normality in Extreme Depth-base Quantile Region Estimation

Y. He

Dpt of Econometrics and Operations Research and Center for Economic Research, Tilburg University; y.he@tilburguniversity.edu

Abstract: Consider the small-probability multivariate quantile regions consisting of extremely outlying points with nearly zero data depth value. We extend the extreme-value-theory based estimation method proposed in [1] and [2] to incorporate general depth functions. Under weak regular variation conditions, both the consistency and asymptotic normality results are derived. A refined asymptotic normality result is established for half-space depth. The simulation study clearly demonstrates the good performance of our refined asymptotic approximation in finite samples. We use our method for risk management by applying it to financial data.

Keywords: Extreme value statistics; Multivariate quantile; Asymptotic normality; Half-space depth; Outlier detection.

References

[1] Cai, J.-J., Einmahl, J.H.J. and de Haan, L. (2011). Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, **39**, 1803-1826.

[2] He, Y. and Einmahl, J.H.J. (2016). Estimation of extreme depth-based quantile regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, DOI: 10.1111/rssb.12163.

2.94 Smooth backfitting of multiplicative structured hazards

M. Hiabu[1,*], E. Mammen[2], M.D. Martínez-Miranda[3] and J.P. Nielsen[1]

[1] City University London, UK; munir.hiabu.1@cass.city.ac.uk, jens.nielsen.1@city.ac.uk
[2] University of Heidelberg, Germany; mammen@math.uni-heidelberg.de
[4] University of Granada, Spain; mmiranda@ugr.es

Abstract: We introduce a smooth backfitting procedure which estimates the one-dimensional components of a multiplicative separable hazard. Given a local linear pilot estimator of the d-dimensional hazard, the backfitting algorithm is motivated form a least squares criterion and converges to the closest multiplicative function. We show that the one dimensional components are estimated with a one-dimensional convergence rate, and hence do not suffer from the curse of dimensionality. The setting is very similar to [1], but we have two significant improvements. First, our approach works without the use of higher order kernels. With them one can theoretically derive nearly $n^{-1/2}$–consistency (with growing order), but they fail to show good performance in practice. Second, the support of the multivariate hazard does not need to be rectangular. We provide an application to non-life insurance where the support is triangular and the structural assumptions on the underlying hazard allow an in-sample forecast of future claims.

Keywords: Smooth backfitting, Survival analysis, Structured models, Local linear kernel estimation, Hazard estimation

References

[1] Linton, O. B:, Nielsen, J. P., Vand de Geer, S. (2003).Estimating multiplicative and additive hazard functions by kernel methods. *The Annals of Statistics*, **31**,464-492.

2.95 Detection of periodicity in functional time series

S. Hörmann[1,*], P. Kokoszka[2] and G. Nisol[1]

[1,*] Université libre de Bruxelles (ULB); shormann@ulb.ac.be, gnisol@gmail.com
[3] Colorado State University; Piotr.Kokoszka@colostate.edu

Abstract: Periodicity is one of the most important characteristics of time series, and tests for periodicity go back to the very origins of the field. This talk is devoted to periodicity tests for time series of functions, which are often called functional time series (FTS's). Examples of FTS's include sequentially observed (e.g. daily) temperature or precipitation curves, pollution level curves, various daily curves derived from high frequency asset price data, bond yield curves, vehicle traffic curves and many others. Such data commonly exhibit periodic (e.g. weekly) patterns. Obtaining significant evidence of a potential periodic component is not only important for understanding the underlying data but also for further analyzing them. Many time series procedures are based on the assumption of stationarity, and hence a periodic pattern needs to be removed before using standard statistical tools. A common approach to inference for functional data is to project them onto a low dimensional basis system and to apply a suitable multivariate procedure to the vector of projections. In the context of testing for periodicity of FTS's such an approach is also possible, but it has the disadvantage of giving equal weight to all projections, whereas the inherent smoothness of functional data implies that projections on suitable systems, e.g. the functional principal components, are not equally important. A functional inferential procedure should emphasize most important directions in an infinite dimensional space in which the data live. We will hence compare projection based approaches with fully functional ones.

Keywords: Functional data; Periodicity; Time series; Fully functional tests

2.96 Confidence bands for a non-parametric regression model with Berkson-type errors-in-variables

H. Holzmann[1,*], N. Bissantz[2] and K. Proksch[3]

[1] Fachbereich Mathematik und Informatik, Philipps-Universität Marburg; holzmann@mathematik.uni-marburg.de
[2] Nicolai Bissantz, Fakultät für Mathematik, Ruhr-Universität Bochum; Nicolai.Bissantz@ruhr-uni-bochum.de
[3] Katharina Proksch, Institut für Mathematische Stochastik, Georg-August-Universität Göttingen; kproksc@uni-goettingen.de

Abstract: First, we briefly review some results on inference in non-parametric models involving deconvolution. Then, more specifically, we consider a measurement error model of Berkson type in a non-parametric regression context. In contrast to other authors we consider the case where the predictors are fixed up to a centered error, and propose an estimator that takes the error in the predictor into account. Further, we derive uniform confidence statements for the function of interest. The confidence statements are based on a Gaussian approximation and an anti-concentration inequality. In a special case we also derive the limit distribution of the maximal deviation of the estimator from its expectation. In a simulation study we investigate the performance of the uniform confidence sets in several settings and compare the results to those obtained when the errors in the variables are neglected.

Keywords: Anti-concentration; Berkson errors-in-variables, Confidence bands; non-parametric regression

2.97 Joint Modeling of Trajectory Data with Informative Dropout

H. Huang[1,*], Y. Shen[2] and Y. Guan[3]

[1] Center for Statistical Science, Peking University, Beijing, China; huanghui@math.pku.edu.cn
[2] Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA; yeshen@uga.edu
[3] Department of Management Science, University of Miami, Coral Gables, FL, USA; yguan@bus.miami.edu

Abstract: In substance use studies, multiple longitudinal trajectories of patients' behavior are usually recorded. These trajectories contain rich information about different patterns of drug-use and treatment effects, but conventional statistical methods may fail to capture them due to several reasons. First, longitudinal trajectories can be dense or sparse, or both; second, there can be lots of missingness in data, and in some cases they are non-ignorable, i.e. not missing at random; third, patients may lie about their actual use of substances, the data reliability should be taken into considerations. We propose a method to jointly model non-Gaussian longitudinal trajectories based on functional data analysis. Our approach assumes latent Gaussian processes behind each of the trajectories, and the correlations between trajectories only depend on the correlation between the Gaussian processes. In this general framework, we allow non-ignorable dropouts and some level of dishonesty in data. Our method is applied to an heroine addiction and treatment study.

Keywords: Functional data analysis; Joint model; Longitudinal trajectories; Non-ignorable dropouts; Latent Gaussian processes.

2.98 Sticky central limit theorems for the hyperbolic ice cream cone

Stephan Huckemann[1], Jonathan C. Mattingly[2], Ezra Miller[2], James Nolen[2]

[1] Universität Göttingen; huckeman@math.uni-goettingen.de
[2] Duke University; jonm@math.duke.edu, ezra@math.duke.edu, nolen@math.duke.edu

Abstract: We derive the limiting distribution of the barycenter of an i.i.d. sample of n random points on a planar cone with angular spread larger than 2π. There are three mutually exclusive possibilities:

(i) (fully sticky case) after a finite random time the barycenter is almost surely at the origin;

(ii) (partly sticky case) the limiting distribution comprises a point mass at the origin, an open sector of a Gaussian, and the projection of a Gaussian to the sectors' bounding rays; or

(iii) (nonsticky case) the barycenter stays away from the origin and the renormalized fluctuations have a fully supported limit distribution – usually Gaussian but not always. Curiously, a (fully) sticky central limit theorem can hold in the absence of any population mean at all. We conclude with an alternative, topological definition of stickiness that generalizes readily to measures on general metric spaces.

Keywords: Hyperbolic singularity; Stratified spaces; Frechét means; Non-positive curvature; Gaussian distribution

2.99 Nonparametric models for extreme risks with urn models

J. Hüsler

juerg.huesler@stat.unibe.ch

Abstract: Extreme risk is often modelled with parametric or semi-parametric models. Here we present nonparametric models for a large setup of risks where the underlying process is not stationary, but is changing with time. The approach is based on the flexible class of urn models. This is made possible for univariate risks as well as multivariate risks. These models can be treated theoretically assuming certain restrictions. If these assumptions are not holding one can always simulate the urn model to get estimates of the extreme risk probabilities.

Keywords: Nonparametric risks; Univariate risks; Multivariate risk; Urn models.

2.100 A Sequential Split-Conquer-Combine Approach for Gaussian Process Modeling using Confidence Distributions

C. Li, M. Xie* and Y. Hung

Department of Statistics and Biostatistics, Rutgers University

Abstract: Gaussian process (GP) models are widely used in the analysis of spatial data, computer experiments, and machine learning. However the computational issue that hinders GP from broader application is generally recognized, especially for massive data observed on irregular grids. In this talk, we introduce a sequential split-conquer-combine (SSCC) approach to tackle this problem. This SSCC approach can substantially reduce the computation at the same time provides an estimation result that is asymptotically equivalent to the one obtained from the entire data. Furthermore, the uncertainty of the proposed GP predictor is quantified using confidence distributions. We illustrate the proposed method by a data center example based on tens of thousands of computer experiments generated from a computational fluid dynamic simulator.

Keywords: Confidence distribution; Gaussian process model; Kriging; Uncertainty quantification

2.101 Change-point detection in multivariate setups

M.Hušková[1,*], S. Meintanis[2] and Z. Hlávka[1]

[1] Charles University in Prague; huskova@karlin.mff.cuni.cz, hlavka@karlin.mff.cuni.cz
[2] National and Kapodistrian University of Athens; simosmei@con.uoa.gr

Abstract: Break–detection procedures for vector observations, both under independence as well as under an underlying structural time series scenario are introduced and studied. The new methods involve L2–type criteria based on empirical characteristic functions. Asymptotic as well as Monte–Carlo results are presented. The new methods are also applied on time–series data from the financial sector.

Keywords: Change-point detection; Empirical characteristic functions; Vector autoregression; Tests; Simulations.

2.102 Bootstrap based uncertainty bands for prediction in functional kriging

M. Franco-Villoria[1] and R. Ignaccolo[1,*]

[1] Department of Economics and Statistics "Cognetti de Martiis", University of Torino; maria.francovilloria@unito.it, rosaria.ignaccolo@unito.it

Abstract: The increasing interest in spatially correlated functional data has led to the development of appropriate geostatistical techniques. Prediction of a curve at an unmonitored location can be obtained using a functional kriging with external drift model that takes into account the effect of exogenous variables (either scalar or functional). Nevertheless uncertainty evaluation for functional spatial prediction remains an open issue. We propose a semi-parametric bootstrap for spatially correlated functional data that allows to evaluate the uncertainty of a predicted curve. Prediction bands are obtained by ordering the bootstrapped predicted curves in two different ways according to band depth and L^2 distance. The performance of the proposed methodology is assessed via a simulation study. Moreover, the approach is illustrated on a well known data set of Canadian temperature and on a real data set of PM_{10} concentration in the Piemonte region, Italy. Based on the results it can be concluded that the method is computationally feasible and suitable for quantifying the uncertainty around a predicted curve.

Keywords: B-splines; Band depth; Functional data modelling; Generalized additive models; Trace-variogram.

References

[1] Cuevas, A. and Febrero, M. and Fraiman, R. (2006) On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis*, **51**, 1063–1074.

[2] Franco-Villoria, M. and Ignaccolo, R. (2015) Bootstrap based uncertainty bands for prediction in functional kriging. *http://arxiv.org/abs/1505.06966*

[3] Giraldo, R. and Delicado, P. and Mateu, J. (2011) Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, **18(3)**, 411–426.

[4] Ignaccolo, R. and Mateu, J. and Giraldo, R. (2014) Kriging with external drift for functional data for air quality monitoring. *SERRA*, **28**, 1171–1186.

[5] Iranpanah, N. and Mohammadzadeh, M. and Taylor, C. (2011) A comparison of block and semi-parametric bootstrap methods for variance estimation in spatial statistics. *Computational Statistics and Data Analysis*, **55**, 578–587.

[6] Lopez-Pintado, S. and Romo, J. (2009) On the concept of depth for Functional Data, *JASA*, **104:486**, 718–734.

2.103 Effective Dose estimation via Single Index Models

F. Balabdaoui[1], C. Durot[2] and H. Jankowski[3,*]

[1] Université Paris-Dauphine
[2] Université Paris Ouest
[3]York University, hkj@yorku.ca

Abstract: We consider the single index models assuming a monotone ridge function. We discuss the geometry arising from maximum likelihood estimation in this model, as well as asymptotics in a "fixed design" setting. Our asymptotic assumptions are motivated by two data sets: one studying the lethal dose in a decompression sickness study and the other studying the lethal dose in a cytotoxicity dataset.

Keywords: Single Index Model; Effective Dose; Shape Constraints; Isotonic Regression

2.104 Applications of the multivariate tail process for extremal inference

H. Drees[1] and A. Janßen[2,*]

[1] University of Hamburg; holger.drees@math.uni-hamburg.de
[2] University of Copenhagen; anja@math.ku.dk

Abstract: Multivariate regularly varying time series are a common tool for modelling the dynamics of heavy-tailed processes of dimension larger than one. Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary d-dimensional regularly varying process. The extremal behavior of this process can be described by the index $\alpha > 0$ of regular variation and the law of the so-called spectral tail process $(\Theta_t)_{t\in\mathbb{Z}}$, for which

$$\mathcal{L}\left(\frac{X_{-n}}{x},\dots,\frac{X_m}{x}\,\middle|\,\|X_0\|>x\right) \overset{w}{\Rightarrow} \mathcal{L}(Y\cdot\Theta_{-n},\dots,Y\cdot\Theta_m),\quad x\to\infty,$$

with a Pareto(α)-distributed random variable Y which is independent of $(\Theta_t)_{t\in\mathbb{Z}}$, cf. Basrak & Segers (2009). The spectral tail process satisfies a certain property which is sometimes called the "time change formula" that describes its behavior when shifted in time, cf. Basrak & Segers (2009). We are interested in estimating the law of Θ_t for $t\in\mathbb{Z}$ with a focus on the cases $t = 0, 1$. These two quantities are of interest in particular for Markov processes $(X_t)_{t\in\mathbb{Z}}$ where their joint distribution (together with the value of α) already determines the whole distribution of $(\Theta_t)_{t\in\mathbb{Z}}$, cf. Janßen and Segers (2014). By extending an idea used in Drees, Segers & Warchoł (2015) from the univariate to the multivariate case we show that it may be helpful for the estimation of Θ_1 to make use of the time change formula that gives us

$$P(\Theta_1\in A) = E\left(\mathbf{1}_A\left(\frac{\Theta_0}{\|\Theta_{-1}\|}\right)\|\Theta_{-1}\|^\alpha\right),\quad A\in\mathbb{B}^d \text{ with } \mathbf{0}\notin A,$$

and use an indirect estimator instead of a direct one. Furthermore, we try to detect independence of $\|\Theta_1\|$ and $\Theta_1/\|\Theta_1\|$ and explore the implications of this fact for the structure of the spectral tail process.

Keywords: Extreme values; Tail process; Multivariate regular variation; Multivariate time series.

References

[1] Basrak, B. and Segers, J.: Regularly varying multivariate time series, *Stochastic Processes and their Applications* **119**, 1055–1080 (2009)

[2] Drees, H., Segers, J. and Warchoł, M: Statistics for tail processes of Markov chains, *Extremes* **18**, 369–402 (2015)

[3] Janßen, A. and Segers, J.: Markov tail chains, *Journal of Applied Probability* **51**, 1133–1153 (2014)

2.105 Unobserved heterogeneity in semiparametric hazard models

G.J. van den Berg[1], L. Janys[2*], E. Mammen [3] and J. P. Nielsen[4]

[1] School of Economics, Finance and Management, University of Bristol; gjvdberg@xs4all.nl
[2]Institute for Financial Economics and Statistics, University of Bonn; ljanys@uni-bonn.de.org
[3] Institute for Applied Mathematics, Heidelberg University; mammen@math.uni-heidelberg.de
[4] Cass Business School, City University London; jens.nielsen.1@city.ac.uk

Abstract: We examine a new general class of hazard rate models for duration data, containing a parametric and a nonparametric component. Both can be a mix of a time effect and possibly time-dependent marker or covariate effects. A number of well-known models are special cases. In a counting process framework, a general profile likelihood estimator is developed and the parametric component of the model is shown to be asymptotically normal and efficient. The analysis improves on earlier results for special cases. Finite sample properties are investigated in simulations. The estimator is shown to work well under realistic empirical conditions. The estimator is applied to investigate the long-run relationship between birth weight and later-life mortality using data from the Uppsala Birth Cohort Study of individuals born in 1915-1929. The results suggest a relationship that is difficult to capture with simple parametric specifications. Moreover, its shape at higher birth weights differs across gender.

Keywords: Covariate effects; Kernel estimation; Semiparametric; Survival analysis; Unobserved Heterogeneity

2.106 Detecting non-simultaneous changes

D. Jarušková

Czech Technical University, Thakurova 7, CZ 166 29 Praha 6, Czech Republic, daniela.jaruskova@cvut.cz

Abstract: At time points $i = 1, \ldots, n$ we observe a sequence of independent two-dimensional vectors $\{\big(X_1(i), X_2(i)\big)\}$ such that $corr\big(X_1(i), X_2(i)\big) = \rho$ where ρ is known. The problem is to decide whether the mean of $\{X_1(i)\}$ and/or the mean of $\{X_2(i)\}$ has changed. It is supposed that if they change they need not to change at the same time. In the presented paper we suggest an off-line procedure for detecting a non-simultaneous change and we study its asymptotic behavior. If we know that both series have changed once, the goal of statistical inference is to estimate the change points (i.e. one change point in the first coordinate, the second change in the second coordinate) and provide their asymptotic distribution.

Keywords: Change point detection; Non-simultaneous changes; Estimates of change-points; Asymptotic disistribution.

References

[1] Jarušková, D. (2015). Detecting non-simultaneous changes in means of vectors. *Test*, **24**, 681–700.

[2] Bai J., Perron P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

[3] Hušková M. (1995). Estimators for epidemic alternatives. *Comment Math. Univ. Carolinae*, **36**, 279–291.

2.107 Bootstrapping INAR models

C. Jentsch[1,*] and C. W. Weiß[2]

[1] University of Mannheim, Department of Economics, L7, 3-5, 68131 Mannheim, Germany; cjentsch@mail.uni-mannheim.de
[2] Helmut Schmidt University, Department of Mathematics and Statistics, Postfach 700822, 22008 Hamburg, Germany; weissc@hsu-hh.de

Abstract: Integer-valued autoregressive (INAR) time series form a very useful class of processes suitable to model time series of counts. In the common formulation of [2], INAR models of order p share the autocorrelation structure with classical autoregressive time series. This fact allows to estimate the INAR coefficients e.g. by Yule-Walker estimators. However, contrary to the AR case, consistent estimation of the model coefficients turns out to be not sufficient to compute proper residuals to formulate a model-based bootstrap procedure. In this paper, we propose to use the semi-parametric estimator suggested by [1] to estimate jointly the INAR coefficients and the distribution of the innovations. Based on these estimates, we propose an INAR bootstrap scheme. We prove bootstrap consistency of our procedure for statistics belonging to the class of functions of generalized means under some mild regularity conditions. In an extensive simulation study, we provide numerical evidence of our theoretical findings and illustrate the superiority of the proposed INAR bootstrap over some obvious competitors.

Keywords: Time series of counts; Bootstrap consistency; Semiparametric estimation; Functions of generalized means

References

[1] Drost, F. C., van den Akker, R. and Werker, B. J. M. (2009). Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR(p) models. *Journal of the Royal Statistical Society, Series B* **71**, 467-485.

[2] Du, J.-G. and Li, Y. (1991). The integer valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis* **12**, 129-142.

2.108 Uniform change point tests in high dimension

M. Jirak[1]

[1] TU Braunschweig; m.jirak@tu-bs.de

Abstract: Consider d dependent change point tests, each based on a CUSUM-statistic. We provide an asymptotic theory that allows us to deal with the maximum over all test statistics as both the sample size n and d tend to infinity. We achieve this either by a consistent bootstrap or an appropriate limit distribution. This allows for the construction of simultaneous confidence bands for dependent change point tests, and explicitly allows us to determine the location of the change both in time and coordinates in high-dimensional time series. If the underlying data has sample size greater or equal n for each test, our conditions explicitly allow for the large d small n situation, that is, where $n/d \to 0$. The setup for the high-dimensional time series is based on a general weak dependence concept. The conditions are very flexible and include many popular multivariate linear and nonlinear models from the literature, such as ARMA, GARCH and related models. The construction of the tests is completely nonparametric, difficulties associated with parametric model selection, model fitting and parameter estimation are avoided. Among other things, the limit distribution for $\max_{1\leq h\leq d}\sup_{0\leq t\leq 1}|\mathcal{W}_{t,h}-t\mathcal{W}_{1,h}|$ is established, where $\{\mathcal{W}_{t,h}\}_{1\leq h\leq d}$ denotes a sequence of dependent Brownian motions. As an application, we analyze all S&P 500 companies over a period of one year.

Keywords: Change point analysis; Weakly dependent high-dimensional time series.

2.109 Targeted learning with big data

P. Bertail[1,2], A. Chambaz[1] and E. Joly[1]

[1] Modal'X; patrice.bertail@gmail.com, achambaz@u-paris10.fr, emilien.joly@u-paris10.fr
[2] CREST

Abstract: The objective of the talk is to construct an asymptotic confidence interval for a statistical parameter $\psi_0 = \psi(P_0)$ of a probability law P_0 based on a sample $O_1, \ldots, O_N$ of size N of observations drawn from P_0. When the function ψ has differentiable properties (in the sense of Hadamard differentiability), a technique called *targeted learning* is, very often, a better choice than the empirical mean based estimator $\psi(P_N)$. When N is very large, the computation of the targeted learning estimator is computationally demanding and may even be impossible. To handle this limitation, a subsampling approach, inspired by [1], is chosen to reduce the sample into another sample $O'_1, \ldots, O'_n$ of fixed size n with $n << N$. We will present a general asymptotic theorem in the context $n \to \infty$ and $n/N \to 0$ ("big data"). The dependence of the resampling procedure in the asymptotic behavior of the estimator will be discussed.

Keywords: Big data; Targeted Learning

References

[1] Bertail, P. and Chautru, E. and Clémençon, S. (2013). Empirical processes in survey sampling. *Preprint*, **https://hal.archives-ouvertes.fr/hal-00989585**.

[2] van der Laan, M. J. and Rose, S. (2011). Targeted learning. *Springer*.

2.110 Estimating a monotone density at zero

G. Jongbloed*, F.H. van der Meulen and L. Pang

Institute of Applied Mathematics, Delft University of Technology, The Netherlands; G.Jongbloed@tudelft.nl, F.H.vanderMeulen@tudelft.nl, L.Pang@tudelft.nl

Abstract: The problem of estimating a decreasing density has attracted quite some attention in the literature. It is encountered in models as the current durations model introduced in [1], but also in other fields of application. The (asymptotic) behavior of the Maximum Likelihood estimator (Grenander estimator) is well understood. For arguments $x > 0$, this estimator is $n^{-1/3}$-consistent under weak conditions. At zero, the estimator is inconsistent. This 'one point' where the estimator misbehaves may not look serious, but in most applications of the model, it is exactly this value that determines interesting quantities. Starting with [2], various approaches have been suggested to estimate the density at zero. In this presentation, some of the models where the problem of estimating a monotone density occurs will be

described, illuminating the relevance of the estimation at zero. A review of the possible approaches to the problem will be given and new thoughts on the matter shared.

Keywords: Spiking; Nonparametric estimation; Penalization.

References

[1] Keiding, N., Højberg Hansen, O.K., Sørensen, D.N. and Slama, R. (2012). The current duration approach to estimating time to pregnancy. *Scandinavian Journal of Statistics*, **39**, 185–204.

[2] Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Statistica Sinica*, **3**, 501–515.

2.111 Rank tests and estimates in measurement error models

J. Jurečková

Charles University in Prague, Faculty of Mathematics and Physics,Prague, Czech Republic; jurecko@karlin.mff.cuni.cz

Abstract: We consider a semiparametric partially linear model where a response is linearly regressed to a set of observable covariates and further depends on some possibly unobservable latent variable (or on its unknown function), which links the model to a semiparametric one. The regressors are either deterministic or random and affected by additive random measurement errors. Our primary goal is the rank inference on the regression parameter in this situation, namely the rank tests and rank estimates. The advantage of rank and signed rank procedures in the measurement errors models was discovered recently in [1] and in [3], among others; the latter made a detailed analysis of rank procedures in the linear model with a nonlinear nuisance regressor and under various kinds of measurement errors. Namely the rank tests can be recommended in this situation: It is shown in [1] that the critical region of the rank test for regression is insensitive to measurement errors in regressors under very general conditions; the errors affect only the power of the test. However, as it was shown in [2], the R-estimator of slope parameter in linear model is biased, because its distribution depends on the power function of the pertaining test. On the other hand, the local asymptotic bias of R-estimator's neither depends on the chosen rank test score-generating functions nor on the unknown distribution of the model errors. It depends only on value of slope parameter vector and on the covariance matrix of the measurement error distribution.

Keywords: Local asymptotic bias; Measurement error; Partially linear model; Rank estimate; Rank test.

References

[1] Jurečková, J., Picek, J. and Saleh, A.K.Md E. (2010). Rank tests and regression rank scores tests in measurement error models. *Computational Statistics and Data Analysis* **54**, 3108–3120.

[2] Jurečková, J., Koul, H. L., Navrátil, R. and Picek, J.(2014). Behavior of R-estimators under measurement errors. *Bernoulli* (in print). arXiv:1411.3609

[3] Sen, P. K. and Jurečková, J. and Picek, J. (2013). Rank tests for corrupted linear models. *Journal of the Indian Statistical Association*, **51**, 201–229.

2.112 Smooth Instrumental Variables: Unifying Inference when the Intensity of Persistence is Uncertain

Ioannis Kasparis and Peter C.B. Phillips

University of Cyprus
Yale University, University of Auckland

Abstract: A novel IV method is proposed that results in regression estimators that have standard limit distributions (normal or mixed normal) irrespective of the stationarity/persistence properties of the data. As a result t-tests and F-tests based on the proposed IV estimator have classic asymptotic distributions (normal and chi-square respectively). Our framework allows for a wide range of regression covariates that can be stationary processes possibly exhibiting long memory, nonstationary long memory processes as well as nearly integrated and mildly integrated processes possibly

driven by long memory innovations, and can accommodate nonlinear regression functions. The proposed instruments are based on nonlinear filtering of the covariates and the implementation of the procedure is very easy. In particular, the non-linear filtering is determined by kernel type of functionals that involve some bandwidth term. The resultant estimator attains the OLS convergence rate under stationarity and the OLS hyper consistency rate less a slowing varying factor under nonstationarity.

2.113 Nonparametric methods for doubly robust estimation of continuous treatment effects

E.H. Kennedy[1]∗ *et al.*

[1] University of Pennsylvania; kennedye@mail.med.upenn.edu

Abstract: Continuous treatments (e.g., doses) arise often in practice, but standard causal effect estimators are limited: they either employ parametric models for the effect curve, or else do not allow for doubly robust covariate adjustment. Double robustness allows one of two nuisance estimators to be misspecified, and is important for protecting against model misspecification as well as reducing sensitivity to the curse of dimensionality. In this work we develop a novel approach for causal dose-response curve estimation that is doubly robust without requiring any parametric assumptions, and which naturally incorporates general off-the-shelf machine learning. We derive asymptotic properties for a kernel-based version of our approach and propose a method for data-driven bandwidth selection. The methods are illustrated via simulation and in a study of the effect of hospital nurse staffing on excess readmissions penalties.

Keywords: Causal inference; Dose-response; Efficient influence function; Kernel smoothing; Semiparametric estimation.

2.114 Models and statistics for projective shape analysis

J. Kent[1]

[1] Department of Statistics, University of Leeds, UK; j.t.kent@leeds.ac.uk

Abstract: The projective shape of a geometric object contains the information that is invariant under projective transformations. The main application is to camera images, where the choice of projective transformation, or "pose", corresponds to the camera view of the object. The simplest example of a projective shape is the cross ratio for a 1d scene of four collinear points. A canonical choice of pose for a 1d/2d scene of landmark points ensures that a circular/spherical film image is "Tyler standardized", as studied in [1]. Tyler standardization is a powerful tool in Procrustes analysis to facilitate the comparison of projective shapes. This talk will describe two recent further developments. It will be shown how Tyler standardization can be used (a) to embed projective shape space into a suitable Euclidean space, and (b) to minimize the effects of measurement error.

Keywords: Cross ratio; Procrustes analysis; Tyler standardization.

References

[1] Kent, J. T. and Mardia, K. V. (2012). A geometric approach to projective shape and the cross ratio. *Biometrika*, **99**, 833–849.

2.115 Estimating conditional momen restriction models under measurement error with unknown distribution

Y. Kitamura[1,*], and T. Otsu[2]

[1] Yale University; yuichi.kitamura@yale.edu
[2] LSE; t.osu@lse.edu

Abstract: This paper is concerned with estimation of general nonlinear conditional moment restriction models in the presence of measurement error. We avoid a priori knowledge of the measurement error distribution. In particular, we utilize repeated measurements and nonparametric blind deconvolution techniques to develop new approaches that

effectively deal with measurement error of general form. The new method is straightforward to implement, and easily adapted to situations where only a subset of covariates are known to be mismeasured. Moreover, measurement error on both covariates and dependent variables can be treated. It is shown that we can achieve the parametric rate of convergence, and moreover, asymptotic normality results are obtained under nonparametric blind deconvolution. The class of moment functions and distribution functions are general and applicable to many economic applications.

Keywords: Measurement error; Deconvolution.

2.116 Oracle inequalities for network models and sparse graphon estimation

O. Klopp[1], Alexandre B. Tsybakov[2] and Nicolas Verzelen[1]

[1] MODAL'X and INRA; kloppolga@math.cnrs.fr, nicolas.verzelen@supagro.inra.fr
[2] ENSAE, UMR CNRS 9194; alexandre.tsybakov@ensae.fr

Abstract: Inhomogeneous random graph models encompass many network models such as stochastic block models and latent position models. We consider the problem of statistical estimation of the matrix of connection probabilities based on the observations of the adjacency matrix of the network. Taking the stochastic block model as an approximation, we construct estimators of network connection probabilities – the ordinary block constant least squares estimator, and its restricted version. We show that they satisfy oracle inequalities with respect to the block constant oracle. As a consequence, we derive optimal rates of estimation of the probability matrix. Our results cover the important setting of sparse networks. Another consequence consists in establishing upper bounds on the minimax risks for graphon estimation in the L_2 norm when the probability matrix is sampled according to a graphon model. These bounds include an additional term accounting for the "agnostic" error induced by the variability of the latent unobserved variables of the graphon model. In this setting, the optimal rates are influenced not only by the bias and variance components as in usual nonparametric problems but also include the third component, which is the agnostic error. The results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).

Keywords: Inhomogeneous random graph; Networks; Oracle inequality; Sparse graphon.

2.117 Image Reconstruction from Poisson Data

C. König[1,2], A. Munk[1,2,3] and F. Werner[2,3]

[1] Department of Mathematics and Computer Science, Georg-August-University of Goettingen, Germany;
claudia-juliane.koenig@mathematik.uni-goettingen.de, munk@math.uni-goettingen.de
[2] Inverse Problems in Biophysics Group, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany;
Frank.Werner@mpibpc.mpg.de
[3] Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Germany

Abstract: Photonic imaging amounts to an inverse (deconvolution) problem with Poisson data. To recover the image we discuss and analyze a particular class of constraint variational regularization estimators (cf. [2], [3]) which can be viewed as a multiscale generalization of the Dantzig selector ([4]). The side constraint combines local log-likelihood ratio tests in a multiscale fashion. Statistical inference for the resulting reconstruction requires to control the family wise error of this constraint. Its asymptotic distributional behaviour is well understood in many cases, e.g. for gaussian regression, cf. [1]. The Poisson model turns out to be substantially different which will be discussed in this talk. A limit theorem over a range of scales and probabilistic bounds will be presented which can be used to provide the resulting reconstructions with a smoothness guarantee, i.e. confidence statements on its smoothness measured in the underlying variational functional, e.g. TV norm. Algorithmic issues are briefly addressed and performance of the method is demonstrated on experimental data from nanoscale superresolution cell microscopy.

Keywords: Statistical multiscale analysis; Poisson limit theorem; Variational regularization; Photonic imaging

References

[1] Dümbgen, L. and Spokoiny, V.G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29**, 124–152.

[2] Frick, K. and Marnitz, P. and Munk, A. (2012). Statistical multiresolution Dantzig estimation in imaging: fundamental concepts and algorithmic framework. *Electronic journal of statistics*, **6**, 231–268.

[3] Frick, K. and Marnitz, P. and Munk, A. (2013). Statistical multiresolution estimation for variational imaging: With an application in Poisson-biophotonics. *Journal of Mathematical Imaging and Vision*, **46(3)**, 370–387.

[4] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

2.118 On the mean-square convergence of estimators of distribution functionals with using additional information

Yu. Dmitriev[1] and G. Koshkin[2,*]

[1] Dpt Applied Mathematics and Cybernetics, Tomsk State University, 36, Lenin, 634050 Tomsk, Russia; dmit@mail.tsu.ru
[2] Dpt Applied Mathematics and Cybernetics, Tomsk State University, 36, Lenin, 634050 Tomsk, Russia; kgm@mail.tsu.ru

Abstract: A class of nonparametric estimators of the main functional of distribution constructed with using additional information is proposed. It is shown that the use of the additional information as the knowledge of other distribution functionals in estimation of the main functional can often provide the mean square error smaller than that of estimators constructed without such additional information. For example, the mathematical expectation of a random variable can be taken as the main functional and the value of its variance can be used as the additional information. The asymptotic normality of the proposed estimators are proved and the main parts of their asymptotic mean square errors are found.

Keywords: Nonparametric estimator; Distribution functional; Additional information; Asymptotic normality; Mean square error.

2.119 On Risk Concentration

B. Das[1] and M. Kratz[2,*]

[1] SUTD Singapore; bikram@sutd.edu.sg
[2] ESSEC Business School, CREAR; kratz@essec.edu

Abstract: We study the behavior of extreme quantiles of the finite sum of heavy-tailed random variables, under multivariate second order regular variation condition. Looking at the literature, asymptotic (for high threshold) results have been obtained, one one hand when assuming (asymptotic) independence and second order regularly varying conditions on the variables, on the other hand when considering specific copula structures. We show that many models used in practice come under the purview of our assumption and provide a few examples. Moreover this ties up related results available in the literature under a broad umbrella. We deduce asymptotic risk concentration results.

Keywords: Aggregation; Multivariate regular variation; Risk measure.

References

[1] B. Basrak, R. Davis, and T. Mikosch. A characterization of multivariate regular variation. *The Annals of Applied Probability*, **12**, 908-920.

[2] L. De Haan, A. Ferreira. *Extreme Value Theory: An Introduction*. Springer-Verlag, New-York.

[3] T. Mao, and T. Hu. Second-order properties of risk concentrations without the condition of asymptotic smoothness. *Extremes*, **16**, 383-405.

[4] S.Resnick. *Heavy Tail Phenomena: Probabilistic and Statistical Modeling*. Springer-Verlag. New York.

2.120 Estimation of the expected shortfall given an extreme component under conditional extreme value model

Rafał Kulik

University of Ottawa

Abstract: For two risks, X, and Y, the Marginal Expected Shortfall (MES) is defined as $\mathrm{E}[Y \mid X > F_X^{\leftarrow}(1-p)]$, where F_X is the distribution function of X and p is small. MES is an important factor when measuring the systemic risk of financial institutions. In this paper we establish asymptotic normality of an estimator of MES on assuming that (X, Y) follow a Conditional Extreme Value (CEV) model. The theoretical findings are supported by simulation studies. Our procedure is applied to some financial data. This is a joint work with a PhD student, Zhigang Tong.

Keywords:

2.121 Goodness-of-fit test for noisy directional data

Claire Lacour collaborative work with Thanh Mai Pham Ngoc

Abstract: In astrophysics, the source of UHECR (Ultra High Energy Cosmic Rays) remains unclear, and we have at our disposal only few observations. To discriminate different assumptions, physicists are investigating whether these rays are distributed uniformly in space. We model the problem by a random variable X on the sphere S^2 spoiled by a random rotation R, so we only observe variable $Z = R(X)$. From the observation of a sample $Z_1, \ldots, Z_n$, we want to test whether the distribution of X is the uniform law. As alternative hypothesis H1, it is assumed that the distance between the density f of X and the uniform density is u_n and that f has regularity s. We explain how to build an adaptive test statistic using spherical harmonics. We give a lower bound and an upper bound for the separation rate u_n, which depends on s and the regularity of the density of R. We also present some simulations.

2.122 Necessary and Sufficient conditions for variable selection consistency of the LASSO in high dimensions

S.N. Lahiri

North Carolina State University

Abstract: In this talk, we consider conditions for variable selection consistency of the LASSO in a high dimensional regression model in the "large p, small n" set up, where p denotes the dimension of the regression model and n denotes the sample size. The main results of the paper give necessary and sufficient conditions for the same, potentially allowing p to grow arbitrarily fast as a function of n. These conditions require both upper and lower bounds on the growth rate of the penalty parameter. In addition to the *Irreprsentable Condition* (IRC) of Zhao and Yu (2006), a new condition (called the *Upper Irreprsentable Condition* or UIRC) is introduced. It is shown that under fairly general conditions, the LASSO with a single choice of the penalty parameter cannot achieve variable selection consistency and root-n consistency.

2.123 Instrumental Regression via Spline Smoothing

Abdelaati Daouia[1] and Pascal Lavergne[2,*]

[1] Toulouse School of Economics, University of Toulouse Capitole; pascal.lavergne@ut-capitole.fr
[2] Toulouse School of Economics, University of Toulouse Capitole; abdelaati.daouia@tse-fr.eu

Abstract: We propose a new nonparametric approach to estimating regression functions in the presence of endogeneity with instrumental variables. In the presence of instrumental variables the relation that identifies the regression function defines an ill-posed inverse problem. We suggest to solve this problem indirectly as a penalized minimization problem with Sobolev penalization. The solution is a natural spline approximation to the unknown regression function. The estimator generalizes the usual spline smoother to the instrumental variables setup. The practical choice of the penalization parameter is investigated.

Keywords: Spline Smoothing; Endogeneity; Instrumental Variable.

2.124 Some new ideas in nonparametric estimation

Oleg Lepski

Univ. Marseille, France

Abstract: In the framework of an abstract statistical model we discuss how to use the solution of one estimation problem (*Problem A*) in order to construct an estimator in another, completely different, *Problem B*. As a solution of *Problem A* we understand a data-driven selection from a given family of estimators $\mathbf{A}(H) = \left\{\widehat{A}_h, h \in H\right\}$ and establishing for the selected estimator so-called oracle inequality. If $\hat{h} \in H$ is the selected parameter and $\mathbf{B}(H) = \left\{\widehat{B}_h, h \in H\right\}$ is an estimator's collection built in *Problem B* we suggest to use the estimator $\widehat{B}_{\hat{h}}$. We present very general selection rule led to selector $\hat{h}$ and find conditions under which the estimator $\widehat{B}_{\hat{h}}$ is reasonable. Our approach is illustrated by several examples related to adaptive estimation.

2.125 Resampling techniques for nonstationary time series

J. Leśkow

Institute of Mathematics, Cracow University of Technology, Poland; jleskow@pk.edu.pl,

Abstract: The talk will be dedicated to recent research on resampling techniques available for nonstationary time series that exhibit periodic or almost periodic structure. Such time series are extremely popular in many applications, ranging from signal processing to wheel bearing fault detections to energy markets. Numerous methodological resuls that are recently published can be divided into two general groups. First group represents the case when the period of the underlying time series is known. The second group corresponds to the unknown period and in such cases we are using the approach based on almost periodic functions. Several techniques will be presented, among them: GSBB, MBB and subsampling. Fundamental issues regarding proofs of consistency will be established together with applications.

Keywords: Periodically and almost periodically correlated time series, consistency of resampling, cyclostationary signals.

References

[1] Napolitano, A. (2012) *Generalizations of Cyclostationary Signal Processing*, Wiley, IEEE Press.

[2] Dehay, D., Dudek, A. and Leśkow, J. (2014), Subsampling for continuous-time almost periodically correlated processes, *Journ. Stat. Plan. Inf*, **150**, 142 - 158.

2.126 Goodness-of-fit testing and nonparametric estimation in count time series

A. Leucht[1,*], M. H. Neumann[2] and F. Fokianos[3]

[1] Technische Universität Braunschweig; a.leucht@tu-bs.de
[2] Friedrich-Schiller-Universität Jena; michael.neumann@uni-jena.de
[3] University of Cyprus, fokianos@ucy.ac.cy

Abstract: So-called INGARCH processes have become popular count time series models.
First, we present a goodness-of-fit test for the special case of Poisson autoregression models. We use a Cramér-von Mises type statistic which can be approximated by a V-statistic. Since the underlying process is not mixing in general, we cannot use any existing result on the asymptotics of these statistics and have to develop new theory. Since the limiting distribution of the test statistic is complicated, we use a model-based bootstrap method to derive critical values.
In the second part of the talk, we address the problem of estimating an isotonic conditional mean function for count time series models with exogenous variables. A popular estimator in isotonic regression is the isotonic least squares estimator which has the advantage that no smoothing parameter has to be chosen. It is known that this estimator attains the optimal rate of convergence in dimension $d = 1$. However, in higher dimensions only consistency has been proven so far. We present a slightly modified estimator for the multivariate case and derive its rate of convergence.

Keywords: Bootstrap; Goodness-of-fit testing; INGARCH models; Isotonic regression.

2.127 Nonlinear sufficient dimension reduction for functional data

B. Li[1,*] and J. Song[1]

[1] Department of Statistics, The Pennsylvania State University; bxl9@psu.edu, junsong@psu.edu

Abstract: We propose a general theory and the estimation procedures for nonlinear sufficient dimension reduction where both the predictor and the response may be random functions. The relation between the response and predictor can be arbitrary and the sets of observed time points can vary from subject to subject. The functional and nonlinear nature of the problem leads to construction of two functional spaces: the first representing the functional data, assumed to be a Hilbert space, and the second characterizing nonlinearity, assumed to be a reproducing kernel Hilbert space. A particularly attractive feature of our construction is that the two spaces are nested, in the sense that the kernel for the second space is determined by the inner product of the first. We propose two estimators for this general dimension reduction problem, and establish the consistency and convergence rate for one of them. These asymptotic results are flexible enough to accommodate both fully and partially observed functional data. We investigate the performances of our estimators by simulations, and applied them to data sets about speech recognition and handwritten symbols.

Keywords: Convergence rate; linear operator; reproducing kernel Hilbert space; Sliced Average Variance Estimator; Sliced Inverse Regression.

2.128 Estimation of high quantiles for quantile autoregression model

D. Li[1,*] and J.H. Wang[2]

[1] Fudan University, China; deyuanli@fudan.edu.cn
[2] George Washington University, USA; judywang@gwu.edu

Abstract: Quantile autoregression (QAR) model was presented by Koenker and Xiao (2006), in which the autoregressive coefficients can be expressed as monotone functions of a single, scalar random variable. The statistical properties of the proposed models and associated estimators were studied and the limiting distributions of the autoregression quantiles processes were derived. In this paper, we use extreme value theory to estimate the autoregressive coefficients at tails and the high quantiles of QAR model. We establish the asymptotic properties of the proposed estimators and demonstrate through the simulation studies that the proposed methods enjoy higher accuracy than the conventional quantile autoregression estimates. A real application is also included.

Keywords: Quantile autoregression; High quantile; Extreme value theory.

References

[1] Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, **101**, 980–1006.

[2] Wang, H.J., Li, D. and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, **107**, 1453–11464.

2.129 A semiparametrically efficient estimator of the time-varying effects for survival data with time-dependent treatmen

Huazhen Lin

Center of Statistical Research, School of Statistical, Southwestern University of Finance and Economics

Abstract: The timing of time-dependent treatment—e.g., when to perform kidney transplantation—is an important factor for evaluating treatment efficacy. A naive comparison between the treatment and nontreatment groups, while ignoring the timing of treatment, typically yields results that might biasedly favor the treatment group, as only patients who survive long enough will get treated. On the other hand, studying the effect of time-dependent treatment is often complex, as it involves modeling treatment history and accounting for the possible time-varying nature of the treatment effect. We propose a varying-coefficient Cox model that investigates the efficacy of time-dependent treatment by utilizing a global partial likelihood, which renders appealing statistical properties, including consistency, asymptotic normality

and semiparametric efficiency. Extensive simulations verify the finite sample performance, and we apply the proposed method to study the efficacy of kidney transplantation for end-stage renal disease patients in the U.S. Scientific Registry of Transplant Recipients (SRTR).
Joint work with Zhe Fei, Yi Li

2.130 Association Analysis via Spearman RV (SRV) and Kernel Spearman RV (KSRV)

J. Lin[1], M. G. Akritas[1]

[1] Department of Statistics, Pennsylvania State University, USA; jul268@psu.edu, mga@stat.psu.edu

Abstract: [1] proposed the RV coefficient, a multivariate generalization of Pearson correlation, which can be used to test the independence of two random vectors. [2] proposed a kernelized RV (KRV), which is zero in population level if and only if the two random vectors are independent. In this presentation, the Spearman RV and kernelized Spearman RV are proposed, in both population and empirical cases. The null distribution of the test statistics is studied. Simulation results comparing SRV with RV and KSRV with KRV are reported, and a real date application to association analysis between multiple traits and multiple genetic variants are presented.

Keywords: RV coefficients; Association analysis; Reproducing kernel Hilbert space; Permutation tests.

References

[1] Escouffier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

[2] Zhan, X. (2016). Measuring and testing dependence by kernelized RV coefficient (preprint).

2.131 Statistical Inference for Matrix-variate Gaussian Graphical Models and False Discovery Rate Control

Xi Chen[1] and Weidong Liu[2]

[1] Stern School of Business, New York University. Email: xchen3@stern.nyu.edu
[2] Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University. Email: weidongl@sjtu.edu.cn

Abstract: Matrix-variate Gaussian graphical models (GGM) have been widely used for modelling matrix-variate data. Since the supports of sparse row and column precision matrices encode the conditional independence among rows and columns of the data, it is of great interest to conduct support recovery. A commonly used approach is the penalized log-likelihood method. However, due to the complicated structure of the precision matrices of matrix-variate GGMs, the log-likelihood is non-convex, which brings challenges for both computation and theoretical analysis. In this paper, we propose an alternative approach by formulating the support recovery problem into a multiple testing problem. A new test statistic is developed and based on that, we use the popular Benjamini and Hochberg's procedure to control false discovery rate (FDR) asymptotically. Our method is computationally attractive since it only involves convex optimization. Theoretically, our method allows very weak conditions, i.e., even when the sample size is finite and the dimensions go to infinity, the asymptotic normality of the test statistics and FDR control can still be guaranteed. The finite sample performance of the proposed method is illustrated by both simulated and real data analysis.

Keywords: Gaussian graphical model; false discovery rate; matrix-variate data.

2.132 Contiguity in High Dimensions

R. Lockhart[1,*], J. Taylor[2], R. Tibshirani[3], and R. Tibshriani[2]

[1] Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A 1S6, CANADA, lockhart@sfu.ca;
[2] Stanford University;
[3] Carnegie Mellon University.

Abstract: In parametric statistical models of fixed finite dimension contiguity results are usually obtained by considering a sequence of alternatives to some null model. This is effective because the null can be approached in a limited number of ways. Non-parametric models are often recast as infinite dimensional models but contiguity works by focussing on alternative sequences heavily loaded onto a fixed finite number of dimensions. In high dimensional regression, however, we propose to treat all directions of departure equally, provided they are aligned with the axes determined by the covariates. The result is that we study signed permutation limits for high dimensional quadratic forms to identify the relevant contiguity neighbourhoods. Results are sparse, so far.

Keywords: Sparse models; Invariance; Combinatorial limit theorems.

2.133 Robust nonparametric methods for analyzing imaging data based on a multivariate volume depth

Sara Lopez-Pintado[1,*] and Julia Wrobel[2]

[1] Columbia University; sl2929@cumc.columbia.edu
[2] Columbia University; jw3134@cumc.columbia.edu

Abstract: Research in many disciplines stands on the analysis of complex data sets of signals and images. For example, in clinical neuroscience large collections of brain images from different subjects are obtained by either functional magnetic resonance (fMRI) or positron emission tomography (PET) to study variations in different neurophysiological states or modifications during psychiatric disorders. Developing new robust statistical tools to analyze these rich data sets is needed. In the applications mentioned above the basic unit of observation can be considered as a general function which is defined in a subset of either the real line or a higher dimensional space, taking values in a univariate or multivariate space. Here we propose a definition of depth defined for general functions which we call multivariate modified volume depth (MMVD). Robust non-parametric statistics is particularly relevant in this setting since usually few assumptions can be made regarding the data generating process and difficult-to-detect outliers may be present in the data. The proposed depth will be used as a building block for developing robust statistics for general complex functions, such as images. We develop several nonparametric permutation tests for comparing location and dispersion of two groups of images based on MMVD. In a simulation study we show the robustness of these tests in the presence of outliers. These statistical tools will be applied to detect whether there are differences in the brain structure or function between healthy individuals and patients with specific mental disorders.

Keywords: Data depth; Image data; Permutation tests.

2.134 Nonparametric estimation of the baseline distribution in the Cox model under monotonicity constraints

H.P. Lopuhaä[1,*], E. Musta[1] and G.F. Nane[1]

[1] Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands; h.p.lopuhaa@tudelft.nl, e.musta@tudelft.nl, g.f.nane@tudelft.nl

Abstract: We investigate non-parametric estimation of a monotone baseline hazard and a decreasing baseline density within the Cox model. We derive the non-parametric maximum likelihood estimator for a non-decreasing baseline hazard function and also consider a Grenander type estimator, defined as the left-hand slope of the greatest convex minorant of the Breslow estimator. We demonstrate that the two estimators are strongly consistent and asymptotically equivalent and derive their common non-Gaussian limit distribution at a fixed point. Furthermore, we consider a Grenander type estimator for a non-increasing baseline density, defined as the left-hand slope of the least concave majorant of an estimator of the baseline cumulative distribution function, derived from the Breslow estimator. Finally, we discuss some recent results on kernel smoothed Grenander-type estimators for a monotone hazard rate and a monotone density in the presence of randomly right censored data, and extensions thereof for the Cox model. This leads to a faster rate of convergence and a Gaussian limit distribution.

Keywords: Cox model; Isotonic estimation; Grenander estimator; Smoothing.

2.135 Generalised partial autocorrelations and the mutual information between past and future

A. Luati[1,*] and T. Proietti[2]

[1] University of Bologna, Department of Statistics; alessandra.luati@unibo.it
[2] University of Rome Tor Vergata, Department of Economics and Finance; tommaso.proietti@uniroma2.it

Abstract: The paper introduces the generalised partial autocorrelation (GPAC) coefficients of a stationary stochastic process. The latter are related to the generalised autocovariances, the inverse Fourier transform coefficients of a power transformation of the spectral density function. By interpreting the generalised partial autocorrelations as the partial autocorrelation coefficients of an auxiliary process, we derive their properties and relate them to essential features of the original process.

Non parametric estimation of the GPAC coefficients is based on a method of moment estimator developed in Proietti and Luati (2015). An alternative estimation strategy is based on a parameterisation suggested by Barndorff-Nielsen and Schou (1973) and on Whittle likelihood. We also prove that the GPAC coefficients can be used to estimate the mutual information between the past and the future of a time series.

Keywords: Generalised autocovariance function; Mutual information; Periodogram.

References

[1] Barndorff-Nielsen, O. E. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, **3**, 408–419.

[2] Proietti, T. and Luati, A. (2015). The generalised autocovariance function. *Journal of Econometrics*, **187**, 245–257.

2.136 Adaptive Sampling for Subgroup Analyses

Alexander R. Luedtke[1] and Antoine Chambaz[2]

[1] Division of Biostatistics, Univ. of California, Berkeley
[2] Univ. Paris Ouest Nanterre

Abstract: Consider a population of patients partitioned into strata based on baseline covariates, each stratum covering a small proportion p, say 5 respect to the population's fixed covariate distribution. We investigate the effect of a binary treatment in an adaptive trial setting where the sampling distribution for covariates is determined by investigators. We wish to estimate the largest conditional average treatment effect within unions of m strata covering a proportion mp of the population, say 10 while sampling from the corresponding optimal union of m strata as often as possible. The multi-armed bandit literature studies this problem in terms of regret. We study it in terms of inference for the conditional average treatment effect in the optimal union of m strata. From our perspective, an optimal design should satisfy two conditions when the optimal union is unique. First, the resulting estimator should have the same asymptotic variance as the semiparametric efficient estimator in a trial where one only samples from the optimal union of strata. Second, the proportion of samples belonging to the suboptimal covariate strata should decay at the optimal rate given in the multi-armed bandit literature. Defining optimality is less straightforward when the optimal union is non-unique, but we will show why we expect massive variance gains over i.i.d. sampling regardless of the optimal union's uniqueness.

Keywords: Adaptive Sampling

2.137 Improved nonparametric long run variance estimation in stochastic regression models with slow consistency rate

Tassos Magdalinos

University of Southampton, UK

Abstract: The paper considers non parametric estimation of the spectral density at zero frequency of an unobserved stationary time series that is approximated by regression residuals, when the associated regression estimators do not achieve the standard $\sqrt{n}$-consistency rate. It is well known that a reduced consistency rate k_n (satisfying $k_n = o(\sqrt{n})$

) of the regression estimates gives rise to additional bias effects that result to less accurate estimation of the spectrum. The paper employs a bias correction that improves the consistency rate of standard kernel estimators with bandwidth parameter M_n from M_n/k_n to $(M_n/k_n)^3$. Bias-corrected estimators are obtained as a solution to a matrix quadratic (Riccati) equation. An application to cointegrating regression with slowly persistent regressors is included.

2.138 Spline backfitted kernel estimation of an additive model for censored data.

S. Maistre[1,*], A. El Ghouch[1] and I. Van Keilegom[1]

[1] Université Catholique de Louvain.

Abstract: Nonparametric additive models have been studied widely since the seminal work of Hastie and Tibshirani (1986). They provide a good compromise between parametric and fully nonparametric modelings. As each function of interest is one-dimensional, it can be pictured and therefore interpreted easily by practicians.

Different techniques aiming to avoid the curse of dimensionality have been proposed, including penalized regression splines and kernel backfitting. A relatively recent method combines the ideas of these two methods, namely Spline Backfitted Kernel (SBK) estimation. We propose to adapt it when the response variable is right censored. This method includes two steps : firstly, estimate the additive model using splines for which the number of knots leads to undersmoothing ; secondly, use the previous estimates to perform a univariate kernel smoothing to estimate each function of interest. We show that when we use synthetic data, the asymptotic results of this procedure are similar to those of the uncensored case.

Keywords: Nonparametric additive model; Right-censoring; Synthetic data.

References

[1] Hastie, T. and Tibshirani, R. (1986). Generalized additive models *with discussion. Statist. Sci.,* ***1(3)****, 297–318.*

[2] Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. Ann. Statist., ***9(6)****, 1276–1288.*

[3] Wang, J. and Yang, L. (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. Ann. Inst. Statist. Math., ***61(3)****, 663–690.*

2.139 Targeted solutions to linear ill-posed problems: a generalization of mollification

P. Maréchal[1,*] and S.K. Mischra[2]

[1]Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France; pr.marechal@gmail.com
[2]Department of Mathematics, Banaras Hindu University, Varanasi 221005, India; bhu.skmishra@gmail.com

Abstract: The use of mollifiers for the regularization of linear inverse problems finds its roots in the late 80's and early 90's. Two approaches have developed independently: the well known approximate inverses on the one hand, based on duality in Hilbert spaces, and Fourier synthesis on the other hand, which belongs to variational methods. Both approaches have in common that, prior to any technical choice, a target object is clearly defined in terms of the unknown true object: the initial ill-posed problem is replaced by that of recovering a smoothed version of the unknown object, smoothness being expressed in terms of convolution.

In order to obtain a general construction for the variational approach to mollification, it is necessary to find solution for an intertwining relationship between operators. In the favorable cases, the intertwining operator may be applied via the unbounded pseudoinverse of the model operator. Applying this unbounded operator may be performed in a stable manner by using the proximal point operator. Our aim here is twofold: we wish to review in a clear and concise manner the mollification approaches to ill-posed problems; and we propose a general construction for the variational approach which, in addition to extending the realm of applicability, will also offer a lot of flexibility in the choice of the target object.

Keywords: Ill-posed problems; Regularization; Intertwining operators, Proximal point algorithm.

References

[1] Bonnefond, X. and Maréchal, P. (2011). A variational approach to the inversion of some compact operators. *Pacific Journal of Optimization*, **5(1)**, 97–110.

[2] Lannes, A., Roques, S. and Casanove, M.-J. (1987). Stabilized reconstruction in signal and image processing; Part I: partial deconvolution and spectral extrapolation with limited field. *Journal of Modern Optics*, **34**, 161–226.

[3] Louis, A.K. and Maass, P. (1990). A mollifer method for linear operator equations of the first kind. *Inverse Problems*, **6**, 427–440.

2.140 Light- and heavy-tailed density estimation by gamma and Gamma-Weibull kernels

L. Markovich[1]

[1] Institute of Control Sciences of Russian Academy of Sciences; kimo1@mail.ru

Abstract:

In [1**?**] we focus on the gamma kernel estimators of density and its derivatives on positive semi-axis by dependent data by univariate and multivariate samples. We introduce the gamma product kernel estimators for the multivariate joint probability density function (pdf) with the nonnegative support and its partial derivatives by the multivariate dependent data with a strong mixing. The asymptotical behavior of the estimates and the optimal bandwidths in the sense of minimal mean integrated squared error (MISE) are obtained. However, it is impossible to fit accurately the tail of the heavy-tailed density by pure gamma kernel. Motivated by a problem arising we construct the new kernel estimator as a combination of the asymmetric gamma and Weibull kernels, ss. Gamma-Weibull kernel. The gamma kernel is nonnegative and it changes the shape depending on the position on the semi-axis and possesses good boundary properties for a wide class of densities. Thus, we use it to estimate the pdf near the zero boundary. The Weibull kernel is based on the Weibull distribution which can be heavy-tailed and hence, we use it to estimate the tail of the unknown pdf. The theoretical asymptotic properties of the proposed density estimator like the bias and the variance are derived. We obtain the optimal bandwidth selection for the estimate as a minimum of the MISE. The optimal rate of convergence of the MISE for the density is found.

Keywords: Multivariate density estimation; (Light-) Heavy-tailed distribution; Gamma kernel; Weibull kernel

References

[1] Markovich, L.A. (2016). Nonparametric gamma kernel estimators of density derivatives on positive semi-axis by dependent data. *RevStat Statistical journal*, (in appear).

[2] Markovich, L.A. (2016). Nonparametric estimation of multivariate density and its derivative by dependent data using gamma kernels. *Under revision in Journal of Nonparametric statistics*, (arXiv:1410.2507).

2.141 Clustering of extremes in large-scale networks and its nonparametric analysis

N. Markovich[1]

[1] Institute of Control Sciences of Russian Academy of Sciences; markovic@ipu.rssi.ru, nat.markovich@gmail.com

Abstract:

Clusters of consecutive exceedances of stochastic processes over sufficiently high threshold or conglomerates of extremes play an important role in the operating of complex large-scale systems in numerous applications. In [1] and [3] asymptotic geometric-like distributions of cluster and inter-cluster sizes are derived, where the cluster is determined as a number of consecutive exceedances over threshold between two consecutive non-exceedances. Considering distribution of an underlying process, its quantiles of sufficiently high levels serve as thresholds and their levels as probabilities in the geometric-like models of cluster and inter-cluster sizes. Expectations of the (inter-)cluster sizes are derived for the same conditions, too. Theoretically, these asymptotic models are obtained for sufficiently large sample sizes, (inter-)cluster sizes and thresholds. Since the models are invariant regarding the distribution of the process under study, one can apply it to stochastic sequences with unknown distributions. Based on these results, we obtain asymptotically equivalent

distributions of other functions like cluster duration, return interval and first hitting time that depend on the cluster structure of the underlying process. We discuss applications of clusters of rare events to social networks and to the caching of the most popular documents transmitted through Internet in the short memory (cache) proposed in [2].

Keywords: Nonparametric estimation; Cluster of extremes; Extremal index; Distributions of cluster and inter-cluster sizes; Caching.

References

[1] Markovich, N.M. (2014). Modeling clusters of extreme values. *Extremes* **17(1)**, 97–125.

[2] Markovich, N.M. (2015). A cluster caching rule in next generation networks. *Proceedings of the 18th international scientific conference Distributed computer and communication networks: control, computation, communications (DCCN-2015)* 127–135 Moscow.

[3] Markovich, N.M. (2016). Clusters of extremes: modeling and inferences. (Working paper).

2.142 Multidimensional two-component Gaussian mixtures detection

B. Laurent[1], C. Marteau[2,*] and C. Maugis-Rabusseau[1]

[1] INSA de Toulouse - Institut de Mathématique de Toulouse; beatrice.laurent@insa-toulouse.fr, cathy.maugis@insa-toulouse.fr
[2] Université Lyon I - Institut Camille Jordan; marteau@math.univ-lyon1.fr

Abstract: Let $(X_1, \ldots, X_n)$ be a d-dimensional i.i.d sample from a distribution with density f. The problem of detection of a two-component mixture is considered. Our aim is to decide whether f is the density of a standard Gaussian random d-vector ($f = \phi_d$) against f is a two-component mixture: $f = (1-\varepsilon)\phi_d + \varepsilon\phi_d(. - \mu)$ where (ε, μ) are unknown parameters. Optimal separation conditions on ε, μ, n and the dimension d are established, allowing to separate both hypotheses with prescribed errors. Several testing procedures are proposed and two alternative subsets are considered.

Keywords: Gaussian mixtures; Non-asymptotic testing procedure; Order statistics; Separation rates.

2.143 Efficient estimation of partially linear models under nonparametric endogeneity

C. Martins-Filho[1*], F. Yao[2] and J. Zhang[1]

[1] University of Colorado at Boulder, Boulder, CO, USA and IFPRI, Washington, DC, USA; carlos.martins@colorado.edu
[2] West Virginia University, Morgantown, WV, USA; feng.yao@mail.wvu.edu
[3] Chinese University of Hong Kong, Shatin, NT, Hong Kong; jszhang@cuhk.edu.hk

Abstract: We consider estimation of parametric and nonparametric components of a multivariate partially linear regression model when the "endogenous" regressors appear in the nonparametric component. Our estimation method relies on the control function approach of [3], which is used to generate suitable theoretical orthogonality conditions based on instrument functions ([2]). Estimators are obtained in two stages. First, pilot estimators are constructed by solving local linearly weighted empirical versions of the orthogonality conditions. Second, the pilot estimators are used in a one-step backfitting procedure to obtain final versions of the estimators. We establish consistency and asymptotic normality of the proposed estimators and show that they are oracle efficient. In contrast with some competing estimators ([1]; [4]), ours are computationally simple requiring no iterative procedure for their calculation. A small Monte Carlo experiment sheds light on our estimators' finite sample properties.

Keywords: Endogeneity; Partially linear model; Semi parametric regression; Control functions.

References

[1] Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**, 1795–1843.

[2] Kim, W., Linton, O. B. and Hengartner, N. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**, 278–297.

[3] Newey, W. K., Powell, J. L. and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equation models. *Econometrica*, **67**, 565–603.

[4] Otsu, T. (2011). Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econometric Theory*, **27**, 8–46.

2.144 Eigenvalue-free risk bounds for functional PCA projectors

André Mas

Institut Montpelliérain Alexander Grothendieck; andre.mas@umontpellier.fr

Abstract: We focus on the PCA of a n-sample of Hilbert-valued data and prove several non asymptotic results related to the difference between estimated and true eigenprojectors. We derive first a lower bound. Then we give upper bounds for the mean square risk for single projectors as well as for projectors associated to the k first eigenvalues. The main point is that these rates do not depend on the spacings between eigenvalues or on their rate of decrease. These results hold for both one-dimensional projectors and for projectors associated to the k first eigenvalues. They may be applied in several directions where fPCA is at work especially inference in the functional linear regression model. This talk puts together results obtained with several other authors in different papers.

Keywords: Functional PCA; Random projectors; Perturbation theory; Functional linear regression; Minimax risk.

2.145 Functional single index model

F. Jiang, S, Baek, J. Cao, Y. Ma[1]

[1] University of South Carolina

Abstract: To study the relation between a univariate response and multiple functional covariates, we propose a functional single index model that is semiparametric. The parametric part of the model integrates the linear regression modeling for functional data and the sufficient dimension reduction structure. The nonparametric part of the model further allows the response-index dependence or the link function to be unspecified. We use B-splines to approximate the coefficient function in the functional linear regression model part and reduce the problem to a familiar dimension folding model. We develop a new method to handle the subsequent dimension folding model by using kernel regression in combination with semiparametric treatment. The new method does not impose any special requirement on the inner product between the covariate function and the B-spline bases, and allows efficient estimation of both the index vector and the B-spline coefficients. The estimation method is general and applicable to both continuous and discrete response variables. We further derive asymptotic properties of the class of methods for both the index vector and the coefficient function. We establish the semiparametric optimality, which has not been done before in a semiparametric model where both kernel and B-spline estimation are involved.

Keywords: Dimension Folding, Single index

2.146 Stein's method and convergence of empirical distributions in an interpretation of quantum mechanics

Ian W. McKeague

Department of Biostatistics, Columbia University, 722 West 168th Street, New York, NY 10032, USA; im2131@columbia.edu

Abstract: [1] recently proposed that quantum theory can be understood as the continuum limit of a deterministic theory in which there is a large, but finite, number of classical "worlds". A resulting Gaussian limit theorem for the empirical distribution of particle positions in the ground state of a harmonic oscillator, agreeing with quantum theory, was conjectured by these authors and proven by [2] using Stein's method. In this talk we discuss new connections between Stein's method and Many Interacting Worlds theory. In particular, we show that quantum position probability densities for higher energy levels beyond the ground state arise as distributional fixed points in a new generalization of Stein's method. These are then used to obtain a rate of distributional convergence for conjectured particle positions in the first energy level above the ground state to the (two-sided) Maxwell distribution. The talk is based on joint work with Erol Peköz and Yvik Swan.

Keywords: Interacting particle system; Distributional approximation.

References

[1] Hall, M. J. W., Deckert, D. A. and Wiseman, H. M. (2014). Quantum phenomena modeled by interactions between many classical worlds. *Phys. Rev. X* **4**, 041013.

[2] McKeague, I. W. and Levin, B. (2015). Convergence of empirical distributions in an interpretation of quantum mechanics. http://arxiv.org/abs/1412.1563, *Ann. Appl. Probab.* to appear.

2.147 Improving the Linear Process Bootstrap through better autocovariance matrix estimation

Timothy L. McMurry[1,*] and Dimitris N. Politis[2]

[1] University of Virginia; tmcmurry@virginia.edu
[2] University of California, San Diego; dpolitis@ucsd.edu

Abstract: [2] introduced several refinements to the banded and tapered estimate of the autocovariance matrix of a stationary processes under short range dependence assumptions [1]. The banded and tapered estimate, which leaves the main diagonals of the sample autocovariance matrix intact while gradually down-weighting off diagonal entries, has been shown to have good asymptotic performance but is not positive definite, which limits its finite sample applicability. The refinements correct the banded and tapered estimate to positivity while maintaining the estimated scale of the process, the banded structure, and the matrices' asymptotic properties. In this work, we revisit the linear process bootstrap (LPB) [1] and demonstrate how these refinements can further improve the LPB's finite sample performance.

Keywords: Autocovariance matrix; Linear process bootstrap; Stationary process; Time series.

References

[1] McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, **31**(6), 471–482.

[2] McMurry, T. L. and Politis, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, **9**(1), 753–788.

2.148 Fourier Criteria for Monitoring Strict Stationarity

Simos Meintanis

Department of Economics, National and Kapodistrian University of Athens, Athens, Greece and Unit for Business Mathematics and Informatics, North–West University, Potchefstroom, South Africa

Abstract: We consider criteria for monitoring strict stationarity of an arbitrary time series. The detectors are formulated on the basis of the empirical characteristic function of the observations and they are meant to monitor stationarity of a general order. The asymptotic null distribution of the detector statistics is studied but the procedures are actually implemented by using a resampling bootstrap scheme appropriate for data involving dependence.

Keywords: Empirical characteristic function; Stationarity; Change–point detection

2.149 Nonparametric approaches for estimating risk maps

P. García-Soidán[1] and R. Menezes[2,*]

[1] Dept. of Statistics and Operations Research, University of Vigo, Spain; pgarcia@uvigo.es
[2] Center of Mathematics, University of Minho, Portugal; rmenezes@math.uminho.pt

Abstract:

Assessment of environmental contamination is increasingly a concern in nowadays society. The maximum levels for pollutants are heavily regulated, being necessary to ensure compliance. Consequently, it becomes important to construct probability maps of the observation region, showing the complementary value of the distribution function of

the variable involved at regulatory thresholds. These are usually called risk maps in the environmental setting. In this work, two kernel-type estimators of the spatial distribution function are constructed, which depart from approximating the distribution at the sampled sites and then obtaining a weighted average of the resulting values, to derive a valid estimator at any random location. Consistency of both approaches is proved under rather general conditions, such as local stationarity and the existence of a number of derivatives of the distribution function. Unlike other alternatives, the new proposals provide non-decreasing functions and do not require a previous estimation of the indicator variogram or the trend function. However, appropriate bandwidths parameters are needed and selection of them in practice needs to be addressed. Numerical studies are carried out, aiming at comparing the current proposal with more usual methods, such as those based on the sill estimation or the indicator kriging, described in [4] or [3], respectively, and redesigned in [1]. Finally, the new proposal is applied to arsenic data from Portugal, so that pollution risk maps of the referred region are constructed. Moreover, accuracy maps of the probability estimates might be constructed based on bootstrap replicas [2].

Keywords: Spatial data; Distribution function; Kernel function; Stationarity

References

[1] García-Soidán, P., Menezes, R. (2012). Estimation of the spatial distribution through the kernel indicator variogram. *Environmetrics*, **23**, 535–548.

[2] García-Soidán, P., Menezes, R. and Rubiños O. (2014). Bootstrap approaches for spatial data. *Stochastic Environmental Research and Risk Assessment*, **28**, 1207–1219.

[3] Pierre Goovaerts (1997). *Geostatistics for natural resources evaluation.* Oxford University Press. New-York.

[4] Journel, A.G. (1983). Nonparametric estimation of spatial distribution. *Mathematical Geology*, **15** (3), 445–468.

2.150 Empirical Characteristic Function-based Inference for Locally Stationary Processes

C. Beering[1], C. Jentsch[2], A. Leucht[1] and M. Meyer[1,*]

[1] Inst. f. Math. Stochastik, Pockelsstr, Germany; c.beering@tu-bs.de, a.leucht@tu-bs.de, marco.meyer@tu-bs.de
[2] University of Mannheim, Germany; cjentsch@mail.uni-mannheim.de

Abstract: We propose a kernel-type estimator for the local characteristic function (local CF) of locally stationary processes. Under weak moment conditions, we prove joint asymptotic normality for local empirical characteristic functions (local ECF). Precisely, for processes having a (two-sided) time-varying MA(1) representation, we establish a central limit theorem under the assumption of finite absolute first moments of the process. Additionally, we prove process convergence of the local ECF. We apply our asymptotic results to parameter estimation of time-varying α-stable distributions.

Finally, we provide a simulation study on minimum distance estimation for α-stable distributions based on local ECF.

Keywords: Locally stationary processes; Local characteristic functions; Kernel estimation.

2.151 Intelligent Sampling From Large Databases

George Michailidis[1]

[1] Department of Statistics, University of Florida; gmichail@ufl.edu

Abstract: In this talk, we demonstrate how intelligent sampling from an *observed database* can lead to precise/efficient estimation of various population parameters by using only a vanishing fraction of the total mass of data. The key strategy is to subsample sparingly at stage one and come up with a pilot estimator of the parameter as well as a calibration of the uncertainty of this pilot estimate; then use this information to sample a small but informative subsample of points from the complementary part of the data-base (the part that did not appear in the first sample), and prescribe an updated estimated based on this second stage subsample. We show that this strategy preserves the optimal convergence rate for estimating threshold/boundary type parameters in a number of different problems, and therefore should find use in the analysis of big data problems where the data-size may be astronomical and the computational algorithms for estimation,

time-consuming. We demonstrate the efficacy of this method via a number of simulations. This is joint work with Zhiyuan Lu and Moulinath Banerjee

Keywords: Subsampling, fast estimation, optimal covergence rate

2.152 Identifying gene signatures in randomized controlled trials for treatment outcome using Cox regression

S. Michiels[1,*], N. Ternès[1] and S. Rotolo[1]

[1] CESP, Service de Biostatistique et d'Épidémiologie, Gustave Roussy, University Paris-Sud, Villejuif, France; stefan.michiels@gustaveroussy.fr

Abstract: With the revolution of the genomic and the targeted era in oncology, gene signatures are becoming increasingly important in clinical research and even in clinical practice. They are expected to provide valuable information for prognosis assessment and for therapeutic decision-making. We investigate approaches to develop a gene signature in Cox regression model using main effects (prognosis) or using treatment by gene interactions (treatment modifiers). For prognosis, we propose different extensions to the lasso penalty in Cox models to reduce the false positive rate. For treatment modifiers, we propose to apply a permutation procedure in a survival model that controls the family-wise error rate at a pre-specified level. A gene signature can be developed for predicting the treatment effects with a crossvalidation scheme. We present simulations under null and alternative scenarios, and illustrate the methods with gene expression data from early breast cancer.

Keywords: Survival analysis; Gene signatures; Lasso; Permutation test; High-dimensional

References

[1] Michiels S., Potthoff R.F. and George S. (2011). Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint *Statistics in Medicine*, **30**, 1502-18.

[2] Michiels S., Rotolo F. (2015) Evaluation of Clinical Utility and Validation of Gene Signatures in Clinical Trials In: Design and Analysis of Clinical Trials for Predictive Medicine Edited by: Matsui S, Buyse M, Simon R. Chapman and Hall/CRC isbn:9781466558151.

[3] Ternes N., Rotolo F., Michiels S. (2015). Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models *Statistics in Medicine*, Under revision.

[4] Ternes N., Rotolo F., Heinze G., Michiels S. (2015). Prediction of treatment benefit in high-dimensional cox models via gene signatures in randomized clinical trials *Trials*, **16**, Suppl 2: 086.

2.153 Local Polynomials for time-varying correlations: adaptivity versus positivity

Giovanni Motta

Columbia University; g.motta@stat.columbia.edu

Abstract: In this paper we propose a new nonparametric method to estimate the time-varying correlation between two non-stationary time series. Linear smoothers of the cross-products are based on the same bandwidth for both numerator (covariance) and denominator (variances). This approach guarantees two important properties: the estimated correlation is bounded between minus one and one, and the resulting correlation matrix is positive semi-definite. However, the use of one common bandwidth for both numerator and denominator appears to be restrictive, as the covariance and the variances are in general characterized by different degrees of smoothness. On the other hand, a kernel-type estimator based on different smoothing parameters for numerator and denominator has two drawbacks. First, the ratio between time-varying numerators and denominators is not necessarily bounded between minus one and one; as a consequence, the resulting correlation matrix is not necessarily positive semi-definite. Second, the estimated bandwidths that are optimal for estimating the covariance and the variances are not necessarily optimal for estimating the ratio. The estimator we propose in this paper is based on local smoothing of the sign of the cross-products, which does not require distinguishing between numerator and denominator. Our novel method can be used to estimate the time-varying AR coefficients and time-varying spectra of locally stationary time series.

Keywords: Local Stationarity ; Auto-correlation; Local Polynomials.

References

[1] Dahlhaus, R. (1996). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. *Athens Conference on Applied Probability and Time Series Analysis, Lecture Notes in Statist.*, **115**, 145–159.

[2] Dahlhaus, R. (2000). *A likelihood approximation for locally stationary processes. The Annals of Statistics*, **28**, 1762–1794.

[3] Politis, D. N. (2011). *Higher-order accurate, positive semidefinite estimation of large sample covariance and spectral density matrices. Econometric Theory*, **27**, 703–744.

2.154 Breaking the Barriers: New Developments in the Theory of Spectral Analysis of Big Graphs

Subhadeep (DEEP) Mukhopadhyay

Temple University, Dpt of Statistics, Philadelphia, U.S., `http://sites.temple.edu/deepstat/`

Abstract: Spectral graph theory is undoubtedly the most favored graph data analysis technique, both in theory and practice. It has emerged as a versatile tool for a wide variety of applications including data mining, web search, quantum computing, computer vision, image segmentation, and among others. However, the way in which spectral graph theory is currently taught and practiced is rather mechanical, consisting of a series of matrix calculations that at first glance seem to have very little to do with statistics, thus posing a serious limitation to our understanding of graph problems from a statistical perspective. Our work is motivated by the following question: How can we develop a general statistical foundation of "spectral heuristics" that avoids the cookbook mechanical approach? A unified method is proposed that permits frequency analysis of graphs from a *nonparametric* perspective by viewing it as function estimation problem. We show that the proposed formalism incorporates seemingly unrelated spectral modeling tools (e.g., Laplacian, modularity, regularized Laplacian, etc.) under a single general method, thus providing better fundamental understanding. It is the purpose of this work to bridge the gap between two spectral graph modeling cultures: Statistical theory (based on nonparametric function approximation and smoothing methods) and Algorithmic computing (based on matrix theory and numerical linear algebra based techniques) to provide transparent and complementary insight into graph problems.

Keywords: Nonparametric spectral graph analysis; Graph correlation density field; Spectral regularization; Orthogonal functions based spectral approximation; Transform coding of graphs; High-dimensional discrete data smoothing.

References

[1] Chung, F. R. (1997). *Spectral graph theory*, vol. 92. American Mathematical Soc.

[2] Galerkin, B. (1915). Series development for some cases of equilibrium of plates and beams (in Russian). *Wjestnik Ingenerow Petrograd*, **19** 897–908.

[3] Mukhopadhyay, S. (2015). Strength of connections in a random graph: Definition, characterization, and estimation. *arXiv:1412.1530*.

[4] Schmidt, E. (1907). Zur Theorie der linearen und nicht linearen Integralgleichungen Zweite Abhandlung. *Mathematische Annalen*, **64** 433–476.

[5] Wiener, N. (1930). Generalized harmonic analysis. *Acta mathematica*, **55** 117–258.

2.155 Efficiency transfer for regression models with missing responses

Ursula U. Müller[1]

[1] Texas A&M University, College Station, TX, 77843-3143, USA; uschi@stat.tamu.edu

Abstract: We consider independent observations on a random pair (X,Y), where the response Y is allowed to be missing at random but the covariate vector X is always observed. We demonstrate that characteristics of the conditional distribution of Y given X can be estimated efficiently using complete case analysis, i.e. one can simply omit incomplete cases and work with an appropriate efficient estimator which remains efficient. This means in particular that we do not have to use imputation or work with inverse probability weights. Those approaches will never be better (asymptotically) than the complete case method.

This efficiency transfer is a general result and holds true for all regression models for which the distribution of Y given X and the marginal distribution of X do not share common parameters. We apply it to the general homoscedastic semiparametric regression model. This includes models where the conditional expectation is modelled by a complex semiparametric regression function, as well as all basic models such as linear regression and nonparametric regression. We discuss estimation of various functionals of the conditional distribution, e.g. of regression parameters and of the error distribution.

This talk is based on joint work with Anton Schick, Binghamton University.

Keywords: Complete case analysis; Efficient influence function; Semiparametric regression; Transfer principle

References

[1] Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.*, 40, 3031-3049.

[2] Müller, U.U. and Schick, A. (2015). Efficiency transfer for regression models with responses missing at random. Preprint. `http://www.stat.tamu.edu/~uschi/research/et.pdf`

2.156 On the optimality of averaging in distributed learning

J. Rosenblatt[1] and B. Nadler[2,*]

[1] Ben Gurion University of the Negev, Beer-Sheva, Israel; johnros@bgu.ac.il
[2] Weizmann Institute of Science, Rehovot, Israel; boaz.nadler@weizmann.ac.il

Abstract: In various applications the data collected is so large that it does not fit a single machine or processing it on a single machine might be too slow. Motivated by the popularity of the map-reduce scheme, we study the statistical loss of distributed machine learning. In particular, we analyze the case where data is randomly distributed among m machines, each computes its own estimator, which is then sent to a central node for merging. We show that under suitable regularity conditions, averaging the m estimators is optimal and we quantify the associated incurred losses. These can be non-negligible either for highly non-linear inference problems, or for high dimensional ones where the number of parameters is comparable to the number of samples in each machine.

Keywords: Distributed learning; M-estimation; Empirical risk minimization; Big data; High order asymptotics.

References

[1] Jonathan Rosenblatt and Boaz Nadler (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference*, to appear.

2.157 Asymptotic Distributional Theory of Weighted Estimation for Nested Case-Control Design

B. Nan[1,*], R. Prentice[2] and T. Cai[3]

bnan@umich.edu and rprentic@whi.org and tcai@hsph.harvard.edu

Abstract: Existing literature shows the improved efficiency of the inverse probability weighted approach compared with the matched set partial likelihood method for the nested case-control study via numerical examples, but the inverse probability weighted estimator lacks a rigorous asymptotic theory. We consider the design where time-matched controls are sampled with replacement and provide a proof of the asymptotic theory of the weighted estimator using modern empirical process theory.

Keywords: Empirical process; Inverse probability weighting; Sampling with replacement; Z-estimation theory.

2.158 Applications of Vector Fields on Infinite Dimensional Manifolds to Model Checking and the Foundations of Statistics

F. Miller[1] and J. Neill[2,*]

[1] Department of Mathematics, Kansas State University; frhjmiller@yahoo.com
[2] Department of Statistics, Kansas State University; jwneill@k-state.edu

Abstract: The set of all probability densities with respect to a fixed measure has the structure of an infinite dimensional Banach manifold. We discuss the use of vector fields on such manifolds to Bayesian statistics and the likelihood principle, conditional densities and ancillary events, and testing the adequacy of parametric models with heterogeneous variance. This includes the development of a matching coalescent clustering methodology.

Keywords: Banach manifold; Vector fields; Model adequacy tests; Clustering; Statistical foundations.

2.159 Nonparametric boundary regression

H. Drees[1], N. Neumeyer[1,*] and L. Selk[1]

[1] Department of Mathematics, University of Hamburg, Bundesstr. 55, 20146 Hamburg, Germany; holger.drees@uni-hamburg.de, natalie.neumeyer@uni-hamburg.de, leonie.selk@uni-hamburg.de

Abstract: We consider regression models with one-sided error distribution. The convergence rate of nonparametric estimators for the boundary curve depends on the regularity of the error distribution in its end point. In irregular models faster rates than those obtained in nonparametric mean regression models are possible. In this talk we discuss estimation of the regression function and the error distribution as well as model specification tests.

Keywords: Goodness-of-fit testing; Local polynomial approximation; One-sided error distribution; Uniform rate of convergence; Residual empirical process

2.160 On False Discovery Rate thresholding for classification under sparsity

P. Neuvial[1], E. Roquain[2]

[1] Université d'Évry Val d'Essonne/UMR CNRS 8071/ENSIIE/USC INRA; pierre.neuvial@genopole.cnrs.fr
[2] LPMA, Université Pierre et Marie Curie/Paris 6; etienne.roquain@upmc.fr

Abstract: We study the properties of false discovery rate (FDR) thresholding, viewed as a classification procedure. The "0"-class (null) is assumed to have a known density while the "1"-class (alternative) is obtained from the "0"-class either by translation or by scaling. Furthermore, the "1"-class is assumed to have a small number of elements w.r.t. the "0"-class (sparsity). We focus on densities of the Subbotin family, including Gaussian and Laplace models. Non-asymptotic oracle inequalities are derived for the excess risk of FDR thresholding. These inequalities lead to explicit rates of convergence of the excess risk to zero, as the number m of items to be classified tends to infinity and in a regime where the power of the Bayes rule is away from 0 and 1. Moreover, these theoretical investigations suggest an explicit choice for the target level α_m of FDR thresholding, as a function of m. Our oracle inequalities show theoretically that the resulting FDR thresholding adapts to the unknown sparsity regime contained in the data. This property is illustrated with numerical experiments.

Keywords: False Discovery Rate; Sparsity; Classification; Multiple testing; Bayes' rule

References

[1] Neuvial, P. and Roquain, E. (2012) On False Discovery Rate thresholding for classification under sparsity. *Annals of Statistics*, **40**(5), 2572–2600.

2.161 A general notion of functional symmetry

A. Nieto-Reyes[1] and H. Battey[2]

[1] Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria; alicia.nieto@unican.es
[2] ORFE, Princeton University; hbattey@princeton.edu and Imperial College London; h.battey@imperial.ac.uk

Abstract: We propose a notion of symmetry, and a corresponding definition of median, for distributions on a general functional metric space and demonstrate the consistency and robustness of the sampling estimators. An important finding is that our notion of symmetry, although significantly more general, coincides theoretically with another very natural but practically infeasible notion of symmetry in the special case of $\mathbb{L}_2$. Simulation studies corroborate our theoretical findings and illustrate the ability of our symmetry notion to capture subtle topological asymmetries. The practical utility of our procedure is illustrated using handwriting data.

Keywords: Functional Data; Exploratory Data Analysis; Quantile Estimation; Robust Procedures

2.162 A smoothed bootstrap method for time series quantile regression

Karl Gregory[1], Soumendra Lahiri[2] and Dan Nordman[3,*]

[1] University of Mannheim, kgregory@mail.uni-mannheim.de
[2] North Carolina State University, snlahiri@ncsu.edu
[3] Iowa State University, dnordman@iastate.edu

Abstract: Quantile regression is helpful for characterizing a response distribution beyond conditional means. Because quantile regression estimators have complex limit distributions, several bootstraps methods have been proposed for independent data, which involve smoothing steps to improve bootstrap approximations. However, no smoothed bootstraps with similar enhancements presently exist for quantile regression with dependent data. In this talk, we consider a smooth tapered block bootstrap for approximating distributions of quantile regression estimators from time series. The bootstrap method employs two rounds of data smoothing in resampling (i.e., individual observations via kernel smoothing techniques and data blocks via smooth tapering). The theoretical development is complicated by the non-smooth objective function in quantile regression along with time dependence and the various smoothing layers in resampling. Despite this, the validity of the proposed bootstrap procedure is established under weak conditions and, as a special case, also broadens the (unsmoothed) moving blocks bootstrap [1]. We illustrate the smooth bootstrap through numerical studies and examples.

Keywords: Kernel smoothing; Moving blocks bootstrap; Tapering; Value at risk.

References

[1] Fitzenberger, B. (1997). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics*, **82**, 235–287.

2.163 Principal Components Analysis and Minimal Surfaces in Tree-Space

T. Nye[1,*], X. Tang[2], G. Weyenberg[3] and R. Yoshida[4]

[1] School of Mathematics and Statistics, Newcastle, UK ; tom.nye@ncl.ac.uk
[2] Mathematics and Computer Sciences, University of Bremen, Germany; xtang@uni-bremen.de
[3] University of Bristol, UK; gradysw@gmail.com
[4] Applied Statistics Laboratory, University of Kentucky, USA; ruriko.yoshida@uky.edu

Abstract: Samples of phylogenetic trees are difficult to summarize and visualize. Methods for constructing principal geodesics in the metric space of phylogenetic trees have been developed previously [2**?**]. However, generalizations of these have not been developed which construct analogs of higher-order principal components. We propose a tree-space analog of the second order principal component which consists of a certain minimal surface which minimizes the sum of squared projected distances of data points onto the surface. The minimial surfaces we consider correspond to the locus of the Fréchet mean of three fixed points in tree-space, where the points are weighted by a probability vector. Each surface is obtained as the probability vector varies over the simplex. Unlike the convex hull of three points, which can have dimension greater than 2 in tree-space, the locus of the Fréchet mean of three points is 2-dimensional, and hence

is a more natural candidate for dimensional reduction. We establish basic properties of these surfaces using results from [1] and go on to discuss algorithms for orthogonal projection. The concepts generalize in a natural way to analogs of third and higher order principal components.

Keywords: Principal components analysis; Phylogenetics; Geometry.

References

[1] Miller, E. and Owen, M. and Provan, J. S. (2011). Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, **68**, 51–91.

[2] Nye, T. M. W. (2011). Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, **39**, 2716–2739.

[3] Nye, T. M. W. (2014). An algorithm for constructing principal geodesics in phylogenetic treespace. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **11**, 304–315.

2.164 Estimation of Fisher linear discrimination subspace using a pair of scatter matrices

R. Sabolová[1], G. Van Bever[1] , F. Critchley[1] and H. Oja[2,*]

[1] The Open University; radka.sabolova@open.ac.uk, germain.van-bever@open.ac.uk, f.critchley@open.ac.uk
[2] University of Turku; hannu.oja@utu.fi

Abstract: It is a remarkable fact that, using any pair of scatter matrices, the Fisher linear discriminant subspace and its dimension can be recovered without knowing group membership of the observations.

[1] introduced invariant coordinate selection (ICS) as a method for exploring multivariate data, based on the eigendecomposition of one scatter matrix relative to another. The aforementioned paper showed that, under a mild multiplicity condition, a subset of these invariant coordinates span the Fisher subspace generated by any mixture of equally-shaped elliptically symmetric densities. It is natural, then, to wonder how the choice of scatter matrices influence the separation between the Fisher subspace (signal space) and its orthogonal complement (noise space). For any choice of scatter matrices satisfying mild assumptions, we find the limiting distribution of the estimated signal space with known and unknown group sizes and develop asymptotic and bootstrap tests for the dimension of the signal space. The efficiencies of estimates with some natural choices of scatter matrices are also compared via simulation studies. Finally, a real data example is given.

Keywords: Asymptotic normality; Dimension reduction; Invariant coordinate selection; Source separation; Test for subellipticity.

References

[1] Tyler, D. E. and Critchley, F. and Dümbgen, L. and Oja, H. (2009). Invariant coordinate selection. *J. Royal Statist. Soc. B*, **71**, 549–592.

2.165 Nonparametrics for networks

Olhede joint work with Pierre-Andre Maugis and Patrick Wolfe

Abstract: Relational data have become an important component of modern statistics. Networks, and weighted networks, are ubiquitous in modern applications such as disease dynamics, ecology, financial contagion, and neuroscience. The inference of networks is harder, in parts because the measure placed on the observables need to satisfy sets of permutation invariances, and most networks are very sparse, with most possible relations not present. This talk will explore how to best construct nonparametric summaries of such objects, in such as way that the underlying statistical model of the observations is well described, and any estimators computable with scalable algorithms.

2.166 Representations for the MSE in kernel regression estimation

P.E. Oliveira[1]

[1] CMUC, Department of Mathematics, University of Coimbra, Portugal; paulo@mat.uc.pt

Abstract: The choice of the bandwidth is well known to be a crucial one in kernel estimation. For finite dimensional data this problem has been extensively studied, with characterizations depending essentially on the dimension of the data. Assuming the absolute continuity means that the geometry and smoothness of the distribution is inherited from the Lebesgue measure. For infinite dimensional or functional data there is no analogue to the Lebesgue measure, so the geometry must be taken into account differently. We discuss representations of the mean square error for the kernel regression estimator, allowing the derivation of an optimal bandwidth choice, in a functional framework depending explicitly on the properties of the distribution, its roughness, and the geometry of the functional space, trying to relax the assumptions on the distributional properties and on the kernel function.

Keywords: Mean square error; Kernel estimation; Small-ball probabilities; Asymptotics.

References

[1] Frédéric Ferraty and Philippe Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice.* Springer, New York.

[2] Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric Regression on Functional Data: Inference and practical aspects. *Aus. N. Z. J. Stat.*, **49**, 267–286.

[3] Jacob, P. and Oliveira, P.E. (1997). Kernel estimators of general Radon-Nikodym derivatives. *Statistics*, **30**, 25–46.

[4] Karol, A. and Nazarov, A. (2014). Small ball probabilities for smooth Gaussian fields and tensor products of compact operators. *Math. Nachr.*, **287**, 595-–609.

[5] Li, W.V. and Shao, Q.M. (2001). Gaussian processes: inequalities, small-ball probabilities and applications, In *Handbook of Statistics*, **19**, 533–597, Elsevier.

[6] Masry, E. (2005) Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Process. Appl.*, **115**, 155–177.

[7] Oliveira, P.E. (2000). Mean square error for histograms when estimating Radon-Nikodym derivatives. *Portugal. Math.*, **57**, 1—16.

[8] Oliveira, P.E. (2005). Nonparametric density and regression estimation for functional data. Preprint, *Publ. Dep. Matemática Univ. Coimbra* **05-09**.

2.167 Convex Hulls in Tree Space

Anna Lubiw[1], Daniela Maftuleac[1] and Megan Owen[2,*]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Canada
[2] Dpt of Mathematics and Computer Science, Lehman College, City University of New York, USA;
megan.owen@lehman.cuny.edu

Abstract: Data generated in such areas as evolutionary biology and medical imaging are frequently tree-shaped, and thus non-Euclidean in nature. As a result, standard techniques for analyzing data in Euclidean spaces become inappropriate, and new methods must be used. One such framework is the space of metric trees constructed by [1]. This space is non-positively curved (hyperbolic), so there is a unique geodesic path (shortest path) between any two trees. Based on this property, a number of statistical methods in Euclidean space can be analogously defined for tree space. One such concept is that of convex hulls, which can be used for computing both quartiles of data and data points of maximal depth. We give an algorithm for computing convex hulls in the space of trees with 5 leaves. We also give examples of how some fundamental properties of Euclidean convex hulls fail to transfer over to tree space.

Keywords: Tree space; Phylogenetics; Convex hull; Non-positively curved; Geometric center.

References

[1] Louis J. Billera and Susan P. Holmes and Karen Vogtmann (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, **27**, 733-767.

2.168 Testing uniformity on high-dimensional spheres against symmetric and asymmetric spiked alternatives

Chr. Cutting, D. Paindaveine* and Th. Verdebout

[1] Université libre de Bruxelles, ECARES and Département de Mathématique, Brussels, Belgium; Christine.Cutting@ulb.ac.be, dpaindav@ulb.ac.be, tverdebo@ulb.ac.be

Abstract: We consider the problem of testing uniformity on high-dimensional unit spheres. We are primarily interested in non-null issues and focus on spiked alternatives. We show that such alternatives lead to two Local Asymptotic Normality (LAN) structures. The first one is for a fixed spike direction θ and allows to derive locally asymptotically optimal tests under specified θ. The second one relates to the unspecified-θ problem and allows to identify locally asymptotically optimal invariant tests. Interestingly, symmetric and asymmetric spiked alternatives lead to very different optimal tests, based on sample averages and sample covariance matrices, respectively. Most of our results allow the dimension p to go to infinity in an arbitrary way as a function of the sample size n. We perform Monte Carlo studies to illustrate our asymptotic results and we treat an application related to testing for sphericity in high dimensions.

Keywords: Directional statistics; High-dimensional statistics; Invariance; Local asymptotic normality; Tests of uniformity

2.169 Smooth Plus Rough Functional Principal Components

Victor M. Panaretos[1]

[1] Department of Mathematics, EPFL; victor.panaretos@epfl.ch

Abstract: Functional data analyses typically proceed by smoothing, followed by functional PCA. This paradigm implicitly assumes that any roughness is due to nuisance noise. Nevertheless, relevant functional features such as time-localised or short scale variations may indeed be rough. These will be confounded with the smooth components of variation by the smoothing/PCA steps, potentially distorting the parsimony and interpretability of the analysis. We will explore how both smooth and rough variations can be recovered on the basis of discretely observed functional data. Assuming that a functional datum arises as the sum of two uncorrelated stochastic processes, one smooth and one rough, we develop identifiability conditions for the estimation of the two corresponding covariance operators. These elucidate the interplay between rank, resolution, and scale. We construct nonlinear estimators of the smooth and rough covariance operators and their spectra via matrix completion, and establish their consistency and rates of convergence. We then use them to recover the smooth and rough constituents of each functional datum, effectively producing separate functional PCAs for smooth and rough variation. (Based on joint work with M.-H. Descary, EPFL).

Keywords: Covariance Operator; Functional Data Analysis; PCA; Matrix completion.

2.170 Sieve Bootstrap for Functional Time Series

Efstathios Paparoditis[1]

[1]University of Cyprus, Cyprus; stathisp@ucy.ac.cy

Abstract: A new bootstrap procedure for functional time series is proposed which exploits a basic representation of the time series of Fourier coefficients appearing in the Karhunen-Loéve expansion of the functional process. A double sieve-type bootstrap method is developed which generates functional pseudo-time series. The method uses a finite set of functional principal components to capture the essential driving parts of the infinite dimensional process and a finite order parametric process to mimic the temporal dependence structure of corresponding vector time series of Fourier coefficients. By allowing the number of functional principal components as well as the order of the model used to increase to infinity (at an appropriate rate) as the sample size increases, a basic bootstrap central limit theorem is established which shows validity of the bootstrap procedure proposed for finite Fourier transforms and spectral density estimators. Some numerical examples illustrate the finite sample performance of the new bootstrap methodology proposed.

Keywords: Bootstrap; Karhunen-Loéve expansion, Vector autoregression.

References

[1] Cerovecki, C. and S. Hoermann (2015). On the CLT for discrete Fourier Transforms of Functional Time Series. Preprint.

[2] Hörmann, S. and P. Kokoszka (2010). Weakly Dependent Functional Data. The Annals of Statistics, 38, 1845-1884.

[3] Paparoditis, E. (2016). Sieve Bootstrap for Functional Time Series. Preprint

2.171 Using Surrogate Marker Information to Test for a Treatment Effect

L. Parast[1,*] and Lu Tian[2]

[1] RAND, Statistics Group, Santa Monica, CA, parast@rand.org
[2] Stanford University, Department of Health, Research and Policy, lutian@stanford.edu

Abstract: The use of surrogate markers to estimate and test for a treatment effect has been an area of popular research. Given the long follow-up periods that are often required for treatment or intervention studies, appropriate use of surrogate marker information has the potential to decrease required follow-up time. However, previous studies have shown that using inadequate markers or making inappropriate assumptions about the relationship between the primary outcome and the surrogate marker can lead to inaccurate conclusions regarding the treatment effect. Currently available methods for identifying, validating and using surrogate markers to test for a treatment effect tend to rely on restrictive model assumptions and/or focus on uncensored outcomes. We propose a novel nonparametric approach to quantify the proportion of treatment effect explained by surrogate marker information and to test for a treatment effect using surrogate marker information collected up to a landmark time in a censored time-to-event outcome setting. Our proposed approach accommodates a setting where individuals may experience the primary outcome before the surrogate marker is measured. Simulation studies demonstrate that the proposed procedures perform well in finite samples and illustrate the expected power loss associated with earlier assessment of a treatment effect. We illustrate our proposed procedure using data from the Diabetes Prevention Program.

Keywords: Survival; Kernel; Surrogate; Treatment effect

2.172 A simple and practical approach towards testing global restrictions on general functions

V. Patilea[1*] and J. Racine[2]

[1] CREST, Ensai, Campus de Ker-Lann, Bruz, France; valentin.patilea@ensai.fr
[2] Department of Economics, McMaster University, Hamilton, Ontario, Canada; racinej@mcmaster.ca

Abstract: We propose a simple bootstrap procedure for inference on vectors or functions in a general context that involves estimation only under the alternative, while constraints are imposed via choice of a suitable transformation of the unconstrained estimate. The procedure is quite general and applies directly to functions or derivatives defined by separable and non-separable regression models. It can be used with parametric, semi- and nonparametric estimators without modification. Potential applications include, but are not limited to, inequality inference on mean or quantile regression models where the bounds depend on the model's covariates, checking monotonicity, convexity, symmetry, homogeneity, for multivariate functions.

Keywords: Shape constraints; Moment inequalities; Bootstrap

2.173 Brand New Location Parameters on Manifolds

Vic Patrangenaru[1]

Department of Statistics, Florida State University; vic@stat.fsu.edu

Abstract: We introduce novel location parameters that reflect the nature of both a random object (r.o.) and its topology of the underlying object manifold $\mathcal{M}$. Assume the Fréchet function associated with a r.o. X on $\mathcal{M}$ is a Morse function (see Milnor(1963), p.50). The set of nondegenerate critical points of the Fréchet Morse function $\mathcal{F}_X$, with fixed index r is the *Fréchet mean set of index r* of X. If $\mathcal{M}$ has dimension m, the Fréchet mean set of index 0 is the Fréchet

antimean set (see Patrangenaru and Ellingson (2015), p.139). In case the Fréchet mean set of index r of X has one point, that point is called *Fréchet mean of index r of X*. Given a random sample of size n from the distribution Q associated with an r.o. on $\mathcal{M}$, we define the *Fréchet sample mean (set) of index r* to be the Fréchet mean (set) of index r of the empirical distribution $\hat{Q}_n$. If the manifold $(\mathcal{M}, \rho_0)$ is compact with respect to the chord distance ρ_0 associated with the embedding j of $\mathcal{M}$ in $\mathbb{R}^N$, we define the *extrinsic mean (set) of index r of Q* to be the Fréchet mean (set) of index r of Q associated with the distance ρ_0, and given a sample of size n from Q, its *extrinsic sample mean (set) of index r* is the extrinsic mean (set) of index r of the empirical distribution $\hat{Q}_n$. We discuss consistency and asymptotic distributions for extrinsic sample means of index r.

Keywords: Object space; Manifold; Fréchet-Morse function; Fréchet mean of index r; Asymptotic distribution.

References

[1] John Milnor (1963). *Morse Theory. Based on lecture notes by. M. Spivak and R. Wells.* Princeton, New Jersey. Princeton University.

[2] Victor Patrangenaru and Leif Ellingson (2015). *Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis*. CRC-Chapman & Hall.

2.174 Bootstrap confidence intervals in functional nonparametric regression under dependence

P. Raña[1*], G. Aneiros[1], J. Vilar[1] and P. Vieu[2]

[1] Departamento de Matemáticas, Universidade da Coruña; paula.rana@udc.es, ganeiros@udc.es, juan.vilar@udc.es
[2] Institut de Mathématiques, Université Paul Sabatier; philippe.vieu@math.univ-toulouse.fr

Abstract: This paper deals with the functional nonparametric regression with scalar response, in which the predictor is of functional nature and when the data are dependent. Two bootstrap procedures are proposed, one for homoscedastic data (the naive bootstrap) and the other for heteroscedastic data (the wild bootstrap) assuming α-mixing conditions on the sample. Asymptotic validity of the proposed procedures has been proved theoretically. This extends the asymptotic results in [2], established for independent samples, to the case of dependent ones using results obtained in [1] for the nonparametric regression estimator under α-mixing conditions. Based on that bootstrap procedures, pointwise confidence intervals for the regression function in the nonparametric model with functional predictor have been built. A simulation study shows promising results when finite sample sizes are used. An application to electricity demand data, from the Spanish Electricity Market, illustrates its usefulness in practice.

Keywords: Functional data; Bootstrap; Nonparametric regression; Confidence intervals; α-mixing.

References

[1] Delsol L. (2009). Advances on asymptotic normality in nonparametric functional time series analysis. *Statistics*, **43**, 13–33.

[2] Ferraty F., Van Keilegom I. and Vieu P. (2010). On the validity of the bootstrap in nonparametric functional regression. *Scandinavian Journal of Statistics*, **37**, 286–306.

[3] Raña P., Aneiros G., Vilar J. and Vieu P. (preprint). Bootstrap confidence intervals in functional nonparametric regression under dependence.

2.175 Oracle inequalities for time-dependent network models

M. Pensky[1,*]

[1] University of Central Florida; marianna.pensky@ucf.edu

Abstract: We consider a dynamic network represented by a sequence of snapshots of the network at discrete time steps. Under the assumption that the network exhibits some degree of regularity, we derive the oracle inequalities for the estimators of the connectivity matrix. In addition, we extend our results to the time-dependent graphon estimation.

Keywords: Dynamic network; Stochastic Block Model; Oracle inequalities.

2.176 Bayes adaptive procedures associated with operators

Dominique Picard

University Paris-Diderot and CNRS-LPMA

Abstract: In many practical problems, the statistical issue is very much connected to a functional operator. It is obviously the case in the so-called linear inverse problems, but also for data observed on geometrical objects, or high dimensional data with sparse geometric assumption, or data with a probabilistic structure involving an operator, as well as many more examples. The spectral decomposition of this operator plays an important role, gene- rally translating a deep and intrinsic structure which leads to a natural choice of estimates (covariance kernel, spectral clustering, inverse problems). Moreo- ver, the operator often induces a genuine regularization, leading to a regularity definition adapted to the structure of the data, which is fundamental in various situations : denoising, semi-supervised learning, classification... We consider here the problem of data with a geometrical structure such as directional data, or data defined on some specific manifolds such as graphs, trees, or matrices. We consider Gaussian a-priori measures. In particular, the problem of adaptation shows the need for adapting the a priori distribution to an harmonic analysis of the structure of the data, and in particular we associate the choice of the Gaussian measure with the Laplacian of the structure. We also investigate the problem from the more explicit angle of an a priori measure on 'manifold-wavelet ' coefficients. We extend the results of Ghosal, Ghosh and van der Vaart, on the concentration a posteriori measures, for the case of geometrical data.

References

[1] I. Castillo, and G. Kerkyacharian and D. Picard, Thomas bayes' walk on manifolds *Prob. Theory and Rel. Fields*, **158**, 665-710.

[2] S. Ghosal and J. Ghosh and A. van de Vaart, Convergence rates of posterio distributions, *Ann. Statist.*, **28**, 500-531.

2.177 Model-free prediction for stationary and nonstationary time series

Dimitris N. Politis[1]

[1] Department of Mathematics, University of California—San Diego, USA; dpolitis@ucsd.edu

Abstract: The Model-free Prediction Principle of Politis (2013, 2015) gives an alternative approach to inference as a whole, including point and interval prediction and estimation. Its application to time series analysis gives new insights to old problems, e.g., a novel approximation to the optimal linear one-step-ahead predictor of a stationary time series, but also helps address emerging issues involving nonstationary data. Examples of the latter include point predictors and prediction intervals for locally stationary time series, and volatility prediction for time-varying ARCH/GARCH processes.

Keywords: Optimal prediction; Prediction intervals.

References

[1] D.N. Politis (2013). Model-free model-fitting and predictive distributions. *Test*, **22**, 183–250.

[2] D.N. Politis (2015). *Model-Free Prediction and Regression: a Transformation-Based Approach to Inference.* Springer. New-York.

2.178 Change point detection in multivariate autoregression

Z. Prášková

Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic; praskova@karlin.mff.cuni.cz

Abstract: In the paper a sequential monitoring scheme is proposed to detect instability of parameters in a multivariate autoregressive process. The proposed monitoring procedure is based on the quasi-likelihood scores and the quasi-maximum likelihood estimators of the respective parameters computed from the training sample, and it is designed so that the sequential test has a small probability of a false alarm and asymptotic power one as the size of the training

sample is sufficiently large. The asymptotic distribution of the detector statistic is established under both the null hypothesis of no change and the alternative that a change occurs. An off-line modification of the score-based detection procedure is considered to detect change in a simple dynamic panel data model in case that both the number of panels and the number of observations are sufficiently large.

Keywords: Change point; Quasi-maximum likelihood; Score test; Autoregression

2.179 Multiscale Scanning in inverse problems - with applications to nanobiophotonics

A. Munk[1,2], K. Proksch[2,*] and F. Werner[1]

[1] Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Goettingen, Germany; Frank.Werner@mpibpc.mpg.de
[2] Institute for Mathematical Stochastics, University of Goettingen, Germany; munk@math.uni-goettingen.de, kproksc@uni-goettingen.de

Abstract: Scan statistics arise in scanning of space (or time) for clusters of events. The basic idea dates back to [1] who investigated the probability that n out of N data points, independently drawn from a $U(0,1)$-distribution, fall into a subinterval of fixed length p. From this starting point the idea was further elaborated and led to the conventional definition of a scan statistic as the maximal value of a suitable test statistic, evaluated at each element of a set of scanning windows.
In this talk we consider the following framework. Let H_1 and H_2 be Hilbert-Spaces, $T : H_1 \to H_2$ a linear, bounded operator and let $f \in H_1$. Suppose we observe Tf according to the following inverse regression model

$$Y_j = Tf(x_j) + \xi_j, \quad j = 1, \dots, N. \tag{1}$$

The points $x_j \in \mathbb{R}^d,\ j = 1, \dots, N$, are fixed design-points, $\xi_j,\ j = 1, \dots, N$, are independent, centered and possibly not identically distributed random variables. Let further $\mathcal{U} = \{\varphi_i\}_{i \in \{1,\dots,N\}} \subset H_1$ be a dictionary. As in the original context of scan statistics, but in the more general framework of model (1) and scanning functions given by the dictionary $\mathcal{U}$, we investigate the extremal distribution of the empirical coefficients $\langle \varphi_i, Y_i \rangle$ and use the results to construct confidence statements for the support of the function f in model (1) with respect to the dictionary $\mathcal{U}$. The case of moving widows of various sizes (scales) is included in this notion. As a particular application to nanobiophotonics, the results are used to derive uniform confidence statements for the positions of molecules in a given sample of which a blurred version is observed under noise.

Keywords: Multiscale inference; Scan statistics; Imaging.

References

[1] Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *J. Amer. Statist. Assoc.*, **60**, 532-538.

2.180 A Neighborhood-Assisted Test for High-Dimensional Means

S. Chen[1], J. Li[2] and Y. Qiu[3,*]

[1] Peking University and Iowa State University
[2] Kent State University
[3] University of Nebraska-Lincoln

Abstract: Although many tests have been proposed to remedy the classical Hotelling's T^2 test in high dimensional setting, those test statistics are constructed without incorporating data dependence by the sample covariance matrix S_n due to its noninvertibility. To incorporate advantageous effect of data dependence, we propose a novel Neighborhood-Assisted (NA) test with test statistic obtained by replacing S_n^{-1} in Hotelling's T^2 with the regularized estimator through banding the Cholesky factor. The NA test is robust to a wide range of dependence in the sense that its implementation does not rely on any structural assumption of the unknown covariance matrix. Simulations and case studies are given to demonstrate the performance of the proposed NA test.

Keywords: Asymptotic normality; Hotelling's T^2 test; High-dimensional data.

2.181 Individualized Variable Selection for Longitudinal Data

X. Tang and A. Qu

[1] University of Illinois at Urbana Champaign; xtang14@illinois.edu and anniequt@illinois.edu

Abstract: We propose a novel individualized variable selection method which performs coefficient estimation, subgroup identification and variable selection simultaneously. In contrast to traditional model selection approaches, an individualized regression model allows different individuals to have different relevant variables. The key component of the new approach is to construct a separation penalty which utilizes cross-subject information and assumes that within-group subjects share the same regression model. This allows us to borrow information from subjects within the same subgroup, and therefore improve the estimation efficiency and variable selection accuracy for each individual. Another advantage of the proposed approach is that it combines strength of homogeneity and heterogeneity in modeling and subgrouping, and therefore enhances the prediction power. We provide theoretical foundation in support of the proposed approach, and propose an effective algorithm to achieve an individualized variable selection. Simulations and an application to the HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.

Keywords: LASSO; Penalized variable selection; Personalized prediction; Separation penalty; Subgrouping.

2.182 Extension sampling designs for big networks - Application to Twitter

A. Rebecq[1]

[1] INSEE / Université Paris X; antoine.rebecq@insee.fr

Abstract: With the rise of big data, more and more attention is paid to statistical network analysis. However, exact computation of many statistics of interest (such as clustering or centrality) is of prohibitive cost for big graphs. A way to solve this problem is to use statistical estimators instead of exact ones. A broad literature on model-based estimation exists, but these estimates cannot be used for quick computation of statistics of interest. Therefore, design-based estimates relying on sampling methods were developed specifically for use on graph populations. Reference libraries for statistical network analysis now implement variations of sampling designs. In this paper, we test some sampling designs used by official statistics institutes to estimate quantities when characteristics of interest on the population can be modeled as networks. These sampling methods can be described as “extension" sampling designs. Unit selection happens in two phases: in the first phase, simple design such as Bernoulli sampling are used, and in the second phase, some units are selected among those that are somehow linked to the units in the first-phase sample. We test these methods on Twitter data because its size and structure typically match the structure of big social networks for which such methods would be very useful. Also, statistics on the Twitter graph are used in many papers in computer and social science.

Keywords: Sampling; Big data; Graphs

2.183 Closed form expressions for rescaled entropy rates of Markov chains

V. Girardin[1] and L. Lhote[2] and P. Regnault[3,*]

[1] Laboratoire de Mathématiques Nicolas Oresme, Univ. de Caen Normandie, FRANCE;
[2] GREYC, ENSICAEN, Univ. de Caen , FRANCE;
[3] Laboratoire de Mathématiques de Reims, EA 4535, FRANCE philippe.regnault@univ-reims.fr

Abstract: Usually, in information theory, the entropy rate – associated to any given entropy functional – of a stochastic process is defined as the limit of the marginal entropy normalized by time units. As established in [1], the limit is degenerated for most usual entropy functionals except Shannon and Rényi. Recently, [2] showed that rescaling the marginal entropy by a pertinent sequence leads to non trivial entropy and divergence rates for a large set of functionals – including Tsallis and Sharma-Mittal – and stochastic processes – including Markov chains.

The entropy rate of any ergodic Markov chain associated to Shannon entropy is well known to have explicit forms, as function of either the stationary distribution of the chain or the eigenvalue of a perturbation of the transition matrix. Under simple conditions on transition matrices, here we obtain closed form expressions for the rescaled entropy rates. They appear as direct extensions of the classical Shannon case, and are explicitly connected to the dynamics of the chain through extended notions of stationary distributions. Illustration is provided for classical entropy functionals and types of Markov chains.

Keywords: Entropy functionals; Entropy rates; Markov chains; Quasi-power property; Quasi-stationary distributions

References

[1] Ciuperca, G. and Girardin, V. and Lhote, L. (2011). Computation and estimation of Generalized entropy rates for denumerable Markov chains. *IEEE Transactions on Information Theory*, **57**(7), 4026–4034.

[2] Girardin, V. and Lhote, L. (2015). Rescaling Entropy and Divergence Rates. *IEEE Transactions on Information Theory*, **61**(11), 5868–5882.

2.184 Testing Separability of Functional Time Series

P. Constantinou[1], P. Kokoszka[2], and M. Reimherr[1,*]

[1] Penn State University; pzc140@psu.edu, mreimherr@psu.edu
[2] Colorado State University; Piotr.Kokoszka@colostate.edu

Abstract: In this work I will present a new statistical test for determining if a panel of functional time series is separable. Separability is property which can dramatically improve statistical efficiency while substantially reducing model complexity. In this context, separability means that the covariance structure factors into the product of two functions, one depending only on time and the other depending only on the coordinates of the panel. Separability is especially useful for functional data as it means that the functional principal components are the same for each member of the panel. However, in practice, such a strong assumption must be checked. Our test is based on functional norm differences and provides a very stable and powerful test. Our methodology will be illustrated via simulations and a financial application. Asymptotic theory will also be discussed.

Keywords: Functional data analysis; Separability; Hypothesis Testing

2.185 Optimal adaptation for early stopping in statistical inverse problems

Gilles Blanchard[1], Marc Hoffmann[2] and Markus Reiß[3,*]

[1] Universität Potsdam; gilles.blanchard@math.uni-potsdam.de
[2] Université Paris Dauphine; hoffmann@ceremade.dauphine.fr
[3] Humboldt-Universität zu Berlin; mreiss@math.hu-berlin.de

Abstract: For linear inverse problems $Y = \mathsf{A}\mu + \xi$, it is classical to recover the unknown signal μ by iterative regularisation methods $(\widehat{\mu}^{(m)}, m = 0, 1, \ldots)$ so that the weak or prediction error $\|\mathsf{A}(\widehat{\mu}^{(\tau)} - \mu)\|^2$ is controlled for some early stopping rule τ based on a discrepancy principle. In the context of statistical estimation with stochastic noise ξ, we study oracle adaptation in strong squared-error $E\big[\|\widehat{\mu}^{(\tau)} - \mu\|^2\big]$. We give precise lower bounds for estimation by early stopping and show how a transfer from weak prediction error to strong error is possible by establishing precise oracle adaptation bounds. Our results illustrate the powerful flexibility of the statistical oracle approach for solving linear inverse problems.

Keywords: Inverse problems; Early stopping; Discrepancy principle; Adaptive estimation; Oracle inequalities.

2.186 Continuous testing for Poisson process intensities

F. Picard[1], E. Roquain[2], A.-L. Fougères[3] and P. Reynaud-Bouret[4,*]

[1] LBBE, UMR CNRS 5558 Univ. Lyon 1, France;
[2] LPMA, Sorbonne Universités, UPMC Univ. Paris 6, Paris, France;
[3] Université de Lyon, CNRS, Université Lyon 1, Institut Camille Jordan, France;
[4] Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, Nice, France;

Abstract: Next Generation Sequencing technologies now allow the genome-wide mapping of binding events along genomes, like the binding of transcription factors for instance. More generally, the field of epigenetics is interested in the regulation of the genome by features that are spatially organized. One open question that remains is the comparison of spatially ordered features along the genome, between biological conditions. An example would be to compare the location of transcription factors between disease and healthy individuals. In Neuroscience, the comparison of two spike trains (eventually with repetitions due to different trials) corresponding to two different stimulus is also of importance to understand how and when the stimulus is affecting the apparition of action potentials. We propose here to model

the observations in both set-ups by two Poisson processes with unknown intensity functions on [0,1], and we restate the problem as the comparison of Poisson process intensites in continuous time. Contrary to global testing approaches that consist in testing whether the two intensities are equal on [0,1], we focus on a local testing strategy using scanning windows. Our method is based on kernel to build the test statistics, and on Monte-Carlo simulations to compute the p-value process. By using the continuous testing framework, we provide a procedure that controls the Family Wise Error Rate as well as the False Discovery Rate in continuous time. We illustrate our method on experimental data, and discuss its extensions in the general framework of testing for Poisson process intensities.

Keywords: Multiple testing; Poisson processes; Exact conditional distribution; Monte-Carlo

2.187 Local bandwidth selection via contrast minimization for locally stationary processes

R. Dahlhaus[1] and S. Richter[1]

[1] Heidelberg University; dahlhaus@statlab.uni-heidelberg.de, stefan.richter@iwr.uni-heidelberg.de

Abstract: In this talk local adaptive bandwidth selection for locally stationary processes is considered. We study locally stationary processes obtained by general Markov structured stationary processes via replacing the finite dimensional parameters by time-dependent parameter curves. For estimation of these curves, we propose nonparametric kernel-type maximum likelihood estimates depending on a smoothing parameter. Up to our knowledge, the theoretical behavior of a data adaptive local bandwidth choice method for such estimates has not been considered in the literature. In this talk, we investigate a local bandwidth selection procedure via contrast minimization. We prove that the corresponding estimators for the parameter curves achieve the asymptotically optimal minimax rate up to log factor which is characteristic for local procedures. Most of our conditions only concern the (usually well-known) corresponding stationary process which allows for an easy verification. The performance of the method is also studied in a simulation for some examples like tvAR, tvARCH and tvMA processes.

Keywords: Locally stationary processes; Bandwidth selection; Contrast minimization; Maximum Likelihood

2.188 Estimating the Price Impact of Trades in a High-Frequency Microstructure Model with Jumps

Eric Jondeau[1], Jérôme Lahaye[2], and Michael Rockinger[3,*]

[1] Affiliation of first and third author; eric.jondeau@unil.ch, michael.rockinger@unil.ch
[2] Affiliation of second author; jlahaye@fordham.edu

Abstract: We estimate a general microstructure model of the transitory and permanent impact of order flow on stock prices. Jumps are detected in both the transaction price (observation equation) and fundamental value (state equation). The model's parameters and variances are updated in real time. Prices can be altered by both the size and direction of trades, and the effects of buy-initiated and sell-initiated trades are different. We estimate this model using tick-by-tick data for 12 large-capitalization stocks traded on the Euronext-Paris Bourse. We find that, at tick frequency, the overnight return, the intraday jumps, and the continuous innovations represent approximately 7%, 8.5%, and 36.7% of the total variation of stock returns. The microstructure model explains on average 47.7% of the total variation. Once jumps are filtered and parameters are estimated in real time, we also find that the price impact of trades is symmetric on average. However, the price of highly liquid stocks with a large proportion of sell-initiated orders tends to be more sensitive to buy trades, whereas the price of less liquid stocks with a large proportion of buy-initiated orders tends to be more sensitive to sell trades.

Keywords: Microstructure model, jumps, noise, volatility, Kalman filter, particle filter.

2.189 Inference on social effects when the network is sparse and unknown

E. Gautier[1], C. Rose[2]

[1] Toulouse School of Economics; eric.gautier@tse-fr.eu
[2] Toulouse School of Economics; chris.rose@tse-fr.eu

Abstract: The paper considers models with social interactions when the network is unobserved and there are endogenous and exogenous social effects. The parameters are estimated by a convex program. The method does not require the knowledge of the variances of the errors nor a subgaussian assumption. It is possible to incorporate shape restrictions. The confidence sets are obtained by solving convex programs. Because the network is unknown, it is not known which exogenous variable does not have a direct effect and can be a valid excluded instrument for the endogenous variables having a direct effect. Therefore we extend the framework of Gautier and Tsybakov (2011, 2014) which handles unknown exclusion restrictions to systems of simultaneous equations. The confidence sets are robust to identification and can be infinite when there is not enough sparsity/exclusion restrictions, when, for some included endogenous regressor, all instruments are too weak or when the number of time periods is too small.

Keywords: Networks; Social effects; Model selection; Inference

2.190 Locally stationary Hawkes processes

F. Roueff

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay; roueff@telecom-paristech.fr

Abstract: We introduce non-stationary Hawkes processes which are defined similarly to standard Hawkes processes but with a time- (or space-)evolving base intensity and fertility function. The resulting process is inhomogeneous. However the usual conditions for the existence of a stationary Hawkes process are easily adapted to obtain a stable non-stationary model. A wildly non-stationary model cannot be consistently inferred, even from an infinite sample of data. Having in mind the statistical analysis of non-stationary Hawkes processes, we propose an approach inspired from locally stationary time series. We are thus interested in an asymptotic framework where the dimension of the observation windows tend to infinity while the time- (or space-)varying parameters are sampled from a function whose corresponding support remains unchanged. We show that under simple assumptions, the statistical properties of the locally stationary Hawkes process can be approximated by those of a stationary Hawkes process. In particular, this framework allows us to propose a time-frequency analysis of Hawkes processes with time varying parameters. This talk is based on [**?**]roueff-sachs-sansonnet2015.

Keywords: Hawkes processes; Locally stationary processes; Time-frequency analysis.

References

[1] François Roueff and Rainer von Sachs and Laure Sansonnet (2015). *Locally stationary Hawkes processes.* To appear in Stochastic Processes and their Applications.

2.191 Variable clustering, G-models and convex optimization

Florentina Bunea[1], Christophe Giraud[2], Martin Royer[1,2,*] and Nicolas Verzelen[3]

[1] Cornell University
[2] Université Paris-Sud
[3] INRA, UMR 729 MISTEA

Abstract: In the problem of variable clustering, we infer groups from the entries of a p-dimensional vector $X = (X_1, \ldots, X_p)$, with the understanding that variables within the same group must exhibit some kind of similarity. A family of probabilistic models – the G-models – has been introduced to define clusters of variables in Bunea et al. [1]. The separation between the thus defined clusters was measured via a new metric, CORD. [1] derive the size of the minimax CORD separation between clusters needed for exact cluster recovery, and show that it is of order $\sqrt{\log p/n}$. Their result still holds even when m, the size of the smallest cluster, grows linearly with p. It is a surprising result, as in other contexts, such as graph clustering with stochastic block models, the growth of m is known to be a helpful factor for the task of clustering. Motivated by this result, we show that the minimax rate for the cluster separation that is needed for exact recovery depends on the separation metric. In [2] and this talk, we investigate the influence of m on the minimax rate for exact recovery, when we replace the CORD metric by the commonly used within-between group correlation gap (GAP). We show that the minimax GAP lower bound for exact recovery is of order $\sqrt{\log p/mn}$. By analyzing an algorithm based on a semi-definite programming (SDP) relaxation of the MLE, we showcase the beneficial influence of an increasing minimal cluster size m in the problem of variable clustering. We use a number of spectral

clustering algorithms as theoretical and numerical benchmarks, and characterize the regimes for which our SDP-type algorithm has superior performance for exact cluster recovery.

Keywords: variable clustering; G-models; convex optimization

References

[1] F. Bunea, C. Giraud, X. Luo, *Minimax Optimal Variable Clustering in G-models via Cord*, arXiv:1508.01939.

[2] F. Bunea, C. Giraud, M. Royer and N. Verzelen (2016). *Near minimax optimal variable clustering via convex optimization* preprint.

2.192 A Nonparametric Approach to Change-point Detection in Dynamic Network

S. Roy∗, S.C. Olhede and P.J. Wolfe

Affiliation of all authors; sandipan.roy@ucl.ac.uk, s.olhede@ucl.ac.uk, p.wolfe@ucl.ac.uk

Abstract: Networks are used to model phenomena that describe relational structure among a group of entities. Many of the real world networks are dynamic in nature. Examples include human interaction network where interaction among groups of individuals undergo changes with time, biological networks such as time-course gene expression data, citation networks, collaboration networks etc. Typically, in many of these instances the underlying relational structure changes over time. Change-points represent a key feature for processes observed over time. Detecting change-points in temporal networks allows us to understand the variation in dependencies among individuals. We study a change-point detection problem in the context of a network that is observed over time. We model the link probabilities among the individuals using a non-parametric spatio-temporal function that extends the idea of non-parametric graphon ([1]) for static networks. In its simplest form we use a Tucker decomposition to separate out the space and the time component of the nonparametric function. We use a testing procedure based on the type of test statistic proposed by [2] to detect the change-points in the relational structure among the individuals in the network. Further, we study the theoretical properties of the test statistic and the numerical performance of the methodology is carried out for the synthetic data as well.

Keywords: Dynamic, Sparsity, Nonparametric graphon, Spatio-temporal, Tucker decomposition

References

[1] Olhede, S.C., and Wolfe, J.P (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences of the United States of America*, **41**, 14722-14727.

[2] La Fond, T., Neville, J., and Gallagher, B. (2014). Anomaly Detection in Dynamic Networks of Varying Size. *arXiv preprint arXiv:1411.3749*.

2.193 Multivariate extreme events : dimension reduction by feature clustering.

A. Sabourin, M. Chiapino, N. Goix, S. Clémençon[1]

[1]LTCI, CNRS, Télécom ParisTech, Univ. Paris-Saclay, Paris, France

Abstract: The dependence structure of multivariate extreme events of multivariate nature is a major concern for risk management. In a high dimensional context ($d > 50$), dimension reduction is a natural first step. However, analyzing the tails of a dataset requires specific approaches that standard algorithms such as PCA do not accommodate. One convenient characterization of extremal dependence is the *angular measure*, defined on the positive orthant of the $d-1$ dimensional hyper-sphere, which provides direct information about the probable 'directions' of extremes. Recent works [1**?**] have defined sparsity in multivariate extremes as the concentration of the angular measure on low dimensional subspheres, and have proposed an algorithm detecting such a pattern in cases where the latter is apparent. Given a dataset that exhibits no clear sparsity pattern, we propose an alternative clustering algorithm allowing to group together the features of the dataset that are 'dependent at extreme level', i.e. that are likely to take extreme values simultaneously. To bypass the computational issues that arise when it comes to dealing with possibly $O(2^d)$ groups

of features, our algorithm exploits the graphical structure stemming from the definition of the clusters, which reduces drastically the number of subgroups on which the estimation procedure has to be performed. Results on simulated and real data show that our method allows to recover a meaningful summary of the dependence structure of extremes in a reasonable amount of computational time.

Keywords: Multivariate extremes; Dimension reduction; Feature clustering; Floods; Anomaly detection

References

[1] Goix, N., Sabourin, A., Clémençon, S.: Sparsity in multivariate extremes with applications to anomaly detection (2015), arXiv preprint arXiv:1507.05899

[2] Goix, N., Sabourin, A., Clémençon, S.: Sparse representation of multivariate extremes with applications to anomaly ranking. In: To appear in the proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) (2016)

2.194 Gap safe rules: safe screening rules to speed-up sparse regression

E. Ndiaye [1], O. Fercoq [1], A. Gramfort [1] and J. Salmon[1,*]

[1] LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay,75013, Paris, France

Abstract: High dimensional regression benefits from sparsity promoting regularizations. Screening rules leverage the known sparsity of the solution by ignoring some variables in the optimization, hence speeding up solvers. When the procedure is proven not to discard features wrongly the rules are said to be "Safe". We will derive new safe rules for generalized linear models regularized with ℓ_1 and ℓ_1/ℓ_2 norms. GAP Safe rules can cope with any iterative solver and we illustrate their performance on coordinate descent for various applications (eg. multi-task Lasso, binary and multinomial logistic regression) demonstrating significant speed ups.

Keywords: Lasso; Duality gap; Sparse regression; Safe rules

2.195 Multivariate Subexponential Distributions and Their Applications

Gennady Samorodnitsky[1] and Julian Sun[2]

[1] Cornell University; gs18@cornell.edu
[2] Cornell University; ys598@cornell.edu

Abstract: We propose a new definition of a multivariate subexponential distribution. We compare this definition with the two existing notions of multivariate subexponentiality, and compute the asymptotic behaviour of the ruin probability in the context of an insurance portfolio, when multivariate subexponentiality holds. Previously such results were available only in the case of multivariate regularly varying claims.

Keywords: heavy tails; subexponential distribution; multivariate; insurance portfolio; ruin probability

2.196 Efficient multivariate entropy estimation with applications to testing shape constraints

T. B. Berrett[1], R. J. Samworth[1,*] and M. Yuan[2]

[1] Statistical Laboratory, University of Cambridge; t.berrett@statslab.cam.ac.uk, r.samworth@statslab.cam.ac.uk
[2] University of Madison-Wisconsin; m.yuan@stat.wisc.edu

Abstract: We study estimation of the differential entropy of a multivariate density. Detailed bias and variance calculations reveal that generalised nearest-neighbour based estimates are efficient in low-dimensional settings. These results are then applied to propose a new test for log-concavity.

Keywords: Efficiency; Entropy; Log-concavity; Nearest neighbours

2.197 Detection of dependence in a model of Poissonian interactions

L. Sansonnet[1,*] and C. Tuleau-Malot[2]

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France; laure.sansonnet@agroparistech.fr
[2] Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351 06100 Nice, France; malot@unice.fr

Abstract: This talk proposes a model of interactions between two point processes, ruled by a reproduction function h, which is considered as the intensity of a Poisson process. In particular, we focus on the context of neuroscience to detect possible interactions in the cerebral activity associated with two neurons. To provide a mathematical answer to this specific problem of neurobiologists, we address so the question of testing the nullity of the intensity h. We construct a multiple testing procedure obtained by the aggregation of single tests based on a wavelet thresholding method. This test has good theoretical properties: it is possible to guarantee the level but also the power under some assumptions and its uniform separation rate over weak Besov bodies is adaptive minimax. Then, some simulations are provided, showing the good practical behavior and the robustness of our testing procedure. This work is published in Statistics and Computing as [1].

Keywords: Adaptive tests; Poisson process; Uniform separation rate; Wavelets; Weak Besov bodies.

References

[1] Sansonnet, L. and Tuleau-Malot, C. (2015). A model of Poissonian interactions and detection of dependence. *Statistics and Computing*, **25**(2), 449–470.

2.198 Minimax Goodness-of-Fit Testing in Ill-Posed Inverse Problems with Partially Unknown Operators

C. Marteau[1] and T. Sapatinas[2,*]

[1]Université Lyon I–Claude Bernard; marteau@math.univ-lyon1.fr
[2] University of Cyprus; fanis@ucy.ac.cy

Abstract: We consider a Gaussian sequence model that contains ill-posed inverse problems as special cases. We assume that the associated operator is partially unknown in the sense that its singular functions are known and the corresponding singular values are unknown but observed with Gaussian noise. For the considered model, we study the minimax goodness-of-fit testing problem. Working with certain ellipsoids in the space of squared-summable sequences of real numbers, with a ball of positive radius removed, we obtain lower and upper bounds for the minimax separation radius in the non-asymptotic framework, i.e., for fixed values of the involved noise levels. Examples of mildly and severely ill-posed inverse problems with ellipsoids of ordinary-smooth and super-smooth sequences are examined in detail and minimax rates of goodness-of-fit testing are obtained for illustrative purposes.

Keywords: Gaussian sequence model; Ill-posed inverse problems; Minimax goodness-of-fit testing; Singular value decomposition.

References

[1] Marteau, C. and Sapatinas, T. (2015a). Minimax Goodness-of-Fit Testing in Ill-Posed Inverse Problems with Partially Unknown Operators. *arXiv:1503.08562 [math.ST]*.

[2] Marteau, C. and Sapatinas, T. (2015b). A Unified Treatment for Non-asymptotic and Asymptotic Approaches to Minimax Signal Detection. *Statistics Surveys*, **9**, 253-297.

2.199 Left truncation for clustered survival data

T. Scheike[1], F. Eriksson and T. Martinussen

[1] Department of Biostatistics, University of Copenhagen, ts@biostat.ku.dk

Abstract: Left truncation occurs frequently in survival studies and it is well known how to handle this when considering univariate survival times. However, when analyzing clustered survival data using semiparametric survival models there are very limited results on how to estimate for example dependence parameters as well as regression effects. Surprisingly, the existing methods only deal with special cases. In this paper we clarify different kinds of left truncation and suggest estimators for semiparametric survival models under specific kinds of left truncation. Small sample properties are investigated via simulation studies and the suggested estimators are used in a real application.

Keywords: Gamma frailty; Left truncation; Counting processes; Hazard model; Survival data; Time-varying effects.

2.200 A distribution function approach for signal reconstruction from ranking data

Michael G. Schimek[1,*] and Vendula Svendova[1]

[1] Medical University of Graz, IMI-RU 'Statistical Bioinformatics', Auenbruggerplatz 2/V, 8036 Graz, Austria; michael.schimek@medunigraz.at, vendula.svendova@medunigraz.at

Abstract: The analysis of Big Data is a challenging task that requires new statistical strategies. A promising strategy is the rank order representation of such data. We assume a set of distinct items of arbitrary size ordered by different ranking mechanisms, resulting in ranked lists. Recently, there have been proposals for the identification of those top ranked items that are characterized by high concordance in their rank positions ([2]; [1]). The thus obtained subset is rather small but highly informative compared to the total number of items. Let us further assume that the underlying signals or decision processes are unobserved. What we aim at is the signal reconstruction from informative ranked sublists. We assume a simple statistical model of the unobserved multiple measurements (i.e. those that have produced the rankings), consisting of a signal component and a normal error component. For the evaluation of the model based on the empirical matrix of ranks we apply indirect inference. A distribution function approach in combination with a numerical optimization technique allows us to estimate a generic (signal) parameter for each item. Under the empirical distribution function we can run a non-parametric Bootstrap to obtain the standard errors of these parameters. The behaviour of the proposed approach is studied by means of simulation.

Keywords: Bootstrap; distribution function; estimation; indirect inference; ranking data.

References

[1] Sampath, S. and Verducci, J. S. (2013). Detecting the end of agreement between two long ranked lists. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **6**, 458–471.

[2] Hall, P. and Schimek, M. G. (2012). Moderate deviation-based inference for random degeneration in paired rank lists. *Journal of the American Statistical Association*, **107**, 661–672.

2.201 Frequentist analysis of the posterior for high-dimensional models

J. Schmidt-Hieber[1]

[1] University of Leiden; schmidthieberaj@math.leidenuniv.nl

Abstract: Recently, various methods have been proposed for estimation and model selection in high-dimensional statistical settings. The most widely known procedure is the LASSO which can be interpreted as a maximum a posteriori probability estimate. Generalizing this, it seems natural to study high-dimensional statistics using the Bayesian method. In the first part of the talk, we summarise recent results concerning posterior shrinkage and model selection for spike-and-slab type priors. These methods are known to perform well theoretically but are hard to compute. The second part of the talk is devoted to priors which can be represented as scale mixtures of normals. This class includes for instance the horseshoe prior. We derive sharp conditions under which such priors are sparsity inducing and show some simulations. This is joint work with Stéphanie van der Pas (Leiden), JB Salomond (Paris Dauphine), Aad van der Vaart (Leiden) and Ismael Castillo (Paris VI).

Keywords: Nonparametric Bayes; high-dimensional statistics

References

[1] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics*, **43**, 1986–2018.

[2] van der Pas, S., Salomond, J.B. and Schmidt-Hieber (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics*, **10**, 976–1000.

2.202 Bagging random domain tessellations for Object Oriented Spatial Statistics

Piercesare Secchi[1]

[1] MOX, Department of Mathematics, Politecnico di Milano; piercesare.secchi@polimi.it

Abstract: Object Oriented Spatial Statistics (O2S2) adresses a variety of application-oriented statistical challenges where the atoms of the analysis are complex data points spatially distributed. The object oriented viewpoint consists in considering as building block of the analysis the whole data point, wheter it is a curve, a distribution or a positive definite matrix, regardless of its complexity. When data are observed over a spatial domain, an extra layer of complexity derives from the size, the shape or the texture of the domain, posing a challenge related to the impossibility, both theoretical and practical, of employing approaches based on global models for capturing spatial dependence. A powerful non-parametric line of action is represented by aggregating simpler and weaker object oriented analyses based on auxiliary data points for which spatial dependence is negligible. These new atoms of the analysis are generated by bootstrapping statistics which are functions of the original data points aggregated along an arrangement of neighbors generated by a random tessellation of the spatial domain. I will illustrate these ideas with a few examples where the target analysis is dimensional reduction, classification or prediction.

The talk is based on discussions and work developed at MOX, Department of Mathematics, Politecnico di Milano, with Alessandra Menafoglio, Simone Vantini and Valeria Vitelli (the latter is now at the Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo).

Keywords: Object Oriented Data Analysis; random tessellations; complex data; spatial dependence.

References

[1] A. Menafoglio and P. Secchi (2016). Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics, *Manuscript*.

[2] Secchi, P., Vantini, S., and Vitelli, V. (2012). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, **22**, 53–64.

[3] Secchi, P., Vantini, S., and Vitelli, V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan (with discussion). *Statistical Methods and Applications* , **24**(2), 279–300.

2.203 Multivariate Nonparametric Estimation of the Pickands Dependence Function using Bernstein Polynomials

G. Marcon[1], S. A. Padoan[1], P. Naveau[2], P. Muliere[1] and J. Segers[3,*]

[1] Bocconi University of Milan; giulia.marcon@phd.unibocconi.it, pietro.muliere@unibocconi.it, simone.padoan@unibocconi.it
[2] Laboratoire des Sciences du Climat et l'Environnement; naveau@lsce.ipsl.fr
[3] Université catholique de Louvain; johan.segers@uclouvain.be

Abstract: Many applications in risk analysis, especially in environmental sciences, require the estimation of the dependence among multivariate maxima. A way to do this is by inferring the Pickands dependence function of the underlying extreme-value copula. A nonparametric estimator is constructed as the sample equivalent of a multivariate extension of the madogram [1]. Shape constraints on the family of Pickands dependence functions are taken into account by means of a representation in terms of a specific type of Bernstein polynomials. The large-sample theory of the estimator is developed and its finite-sample performance is evaluated with a simulation study. The approach is illustrated by analyzing clusters consisting of seven weather stations that have recorded weekly maxima of hourly rainfall in France from 1993 to 2011.

Keywords: Bernstein polynomials; Extremal dependence; Extreme-value copula; Pickands dependence function.

References

[1] Naveau, P., A. Guillou, D. Cooley, and J. Diebolt (2009). Modelling pairwise dependence of maxima in space. *Biometrika*, **96**(1), 1–17.

2.204 Parameter Estimation of Multitype Branching Processes Using Probability Generating Functions

R. Senoussi[1] and A. Ouaari[2,*]

[1] senoussi@avignon.inra.fr and [2] ouaari@avignon.inra.fr

Abstract: Continuous time, homogeneous multitype branching processes are very appealing tools for modeling population dynamics. However for statistical purposes, very few is generally known concerning the analytic forms of the probability distributions of these Markov processes whereas more information are available on their probability generating functions. In this talk we develop a straight approach to deal with the estimation of parameters associated to the branching mechanisms. We first start describing the main properties of these processes and then explicit the leading partial differential equation associated to the the probability generating function of the offspring distribution. In a second part we focus on the inversion problem of the generating functions to carry out a maximum likelihood procedure to infer the parameters describing the infinitesimal generator of these processes. The inversion procedure is achieved either analytically or numerically using Cauchy formula for multivariable analytic functions. We finally illustrate the approach with several branching examples.

Keywords: Multi-type continuous-time branching processes;Probability generating function; Maximum-likelihood estimators; Linear partial differential equations; Cauchy's theorem.

References

[1] Becker N. (1977). Estimation for Discrete Time Branching Processes with Application to Epidemics. *Biometrics*, **33, No 3**, 515–522.

[2] Asmussen S. and Keiding N. (1978). Martingal Central Limit Theorems and Asymptotic Estimation Theory for Multitype Branching Processes. *Advances in Applied Prob.*, **10, No 1**, 109–129.

[3] González M., Minuesa C. and del Puerto I. (2016). Maximum likelihood estimation and expectation maximization algorithm for controlled branching processes. *Computational Statistics and Data Analysis*, **93**, 209–227.

[4] Athreya, K. B. and Ney P. E. (1972). *Branching Processes.* Springer. Berlin.

2.205 Global and local functional depths

C. Sguera[1,*], P. Galeano[2] and R. Lillo[2]

[1] Research Institute UC3M-Banco Santander on Financial Big Data, Universidad Carlos III de Madrid; csguera@est-econ.uc3m.es
[2] Department of Statistics, Universidad Carlos III de Madrid; pgaleano@est-econ.uc3m.es, lillo@est-econ.uc3m.es

Abstract: A functional data depth provides a center-outward ordering criterion which allows robust measures, such as the median, trimmed means, central regions or ranks, to be defined in the functional framework. A functional data depth can be global or local. With global depths, the degree of centrality of a curve x depends equally on the rest of the sample observations, while with local depths, the contribution of each observation in defining the degree of centrality of x decreases as the distance from x increases. We present a comparative analysis of the global and local approaches to the functional depth problem focusing on the "global" functional spatial depth (FSD) and its local version, the kernelized functional spatial depth (KFSD). Considering real applications and simulated data sets, we illustrate the differences between global and local depths, and when they should be expected.

Keywords: Functional data; Global depths; Local depths.

2.206 Efficient Mean Estimation with Nonignorable Nonresponse via Sufficient Dimension Reduction

P. Zhao, J. Shao* and L. Wang

Dpt of Statistics, Univ. of Wisconsin-Madison, USA; pyzhao@live.cn, shao@stat.wisc.edu, leiwang.stat@gmail.com

Abstract: We consider the estimation of a response variable subject to nonignorable nonresponse when the distribution of the variable and related covariates is unspecified and the nonresponse propensity is semiparametric. There exist some recently proposed methods based on estimation of the propensity, using kernel regression for the nonparametric component and a nonresponse instrument approach to estimate the parametric component. Two challenging issues have not been well addressed. The first one is how to apply nonparametric dimension reduction to produce an efficient kernel estimator or alleviate the curse of dimensionality. The second issue is how to find a covariate to serve as a nonparametric instrument in identifying and estimating the propensity. We propose to use the nonparametric sufficient dimension reduction (SDR) technique to overcome these two difficulties. In the presence of nonignorable nonresponse, we construct some suitable scores to apply SDR, and we find a way to select nonresponse instrument invariant to the parameter in the propensity. Asymptotic normality of the proposed estimators are established. We evaluate the performance of the proposed estimators in a Monte Carlo study and illustrate our method in an application to ACTG 175 data.

Keywords: Dimension reduction; Identifiability; Instrument; Nonparametric regression; Semiparametric propensity.

2.207 Semiparametric Bayesian Estimation and Comparison of Moment Condition Models

Siddhartha Chib[1], Minchul Shin[2] and Anna Simoni[3,*]

[1] Olin Business School, Washington University in St. Louis, chib@wustl.edu
[2] Department of Economics, University of Illinois, mincshin@illinois.edu
[3]CNRS and CREST-Ensae, simoni.anna@gmail.com

Abstract: In this paper we consider the problem of inference in statistical models characterized via moment restrictions and develop a semiparametric Bayes procedure for selecting valid and relevant moments. We cast the moment estimation problem in the Exponentially Tilted Empirical Likelihood (ETEL) framework introduced by [2]. Because the ETEL has a well-defined probabilistic interpretation and plays the role of a likelihood, a fully Bayesian framework can be developed. We show how the moment selection problem can be tackled on the basis of marginal likelihoods. These are computed exactly (up to simulation error) by [1]'s method. We show that our proposed marginal likelihood based moment selection procedure is consistent in the sense that it discards misspecified as well as irrelevant moment restrictions. As a byproduct, we prove that a posterior distribution obtained from the ETEL satisfies the Bernstein - von Mises theorem in misspecified moment models. The finite sample properties of our procedure are illustrated in simulation exercises in the settings of linear instrumental regression and quantile instrumental regression.

Keywords: Exponential tilting empirical likelihood, moment selection, Bayesian estimation, misspecified models

References

[1] Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

[2] Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, **35**, 634–672.

2.208 Kernel Machine Score Test for Pathway Analysis in the Presence of Semi-Competing Risks

M. Neykov[1], B.P. Hejblum[2] and J.A. Sinnott[3,*]

[1]Dpt of Operations Research and Financial Engineering, Princeton Univ., Princeton, USA; mneykov@princeton.edu
[2]Department of Biostatistics, Harvard University, Boston, MA, USA; bhejblum@hsph.harvard.edu
[3,*]Department of Statistics, The Ohio State University, Columbus, OH, USA; jsinnott@stat.osu.edu;

Abstract: In cancer studies, patients often experience two different types of events: a non-terminal event such as recurrence or metastasis, and a terminal event such as cancer-specific death. Identifying pathways and networks of genes associated with one or both of these events is an important step in understanding disease development and targeting new biological processes for potential intervention. These correlated outcomes are commonly dealt with by modeling progression-free survival, where the event time is the minimum between the times of recurrence and death. However, identifying pathways only associated with progression-free survival may miss out on pathways that affect time to recurrence but not death, or vice versa. We propose a combined testing procedure for a pathway's association with both the cause-specific hazard of recurrence and the marginal hazard of death. The dependency between the two outcomes is accounted for through perturbation resampling to approximate the test's null distribution, without any further assumption on the nature of the dependency. Even complex nonlinear relationships between pathways and disease progression or death can be uncovered thanks to a flexible kernel machine framework. The superior statistical power of our approach is demonstrated in numerical studies and in a gene expression study.

Keywords: Kernel Machines; Pathway Analysis; Resampling; Score Test; Semi-Competing Risks

2.209 Semiparametric Bayesian Methods for Skewed Multivariate Response

Debajyoti Sinha

Florida State University

Abstract: Models and methods based on assumption of symmetric error density often lead to invalid and inefficient estimator of covariate effects on the highly skewed response data. We propose a new class of semiparametric models for heavily skewed clustered responses while taking into account the flexible structure for within cluster association. We explore the properties and advantages of our model compared to existing models for skewed clustered response. We illustrate the implementation and practical advantages of our methods forsemiparametric Bayesian analysis of clustered data from a real life biomedicalstudy. We examine the performance and robustness of our proposed methods for finite sample sizes via simulation studies. This is a joint work with Dr.S.Lipsitz of Harvard Medical School and A.Bhingare of FSU)

Keywords:

2.210 Nonlinear Statistical Inferences With Laplace Measurement Error

Weixing Song[1*]

[1] Kansas State University; weixing@ksu.edu

Abstract: When a p-dimensional parameter θ is defined through the moment condition $Em(X,\theta) = 0$, a simple estimation procedure of θ is proposed by Hong and Tamer when X, a k-dimensional random vector, are contaminated with Laplace measurement error U, that is, we can only observe $Z = X + U$. However, the estimation procedure was designed particularly for the cases where the components of the measurement error vector U are independent. In this paper, we first introduce a general multivariate Laplace distribution, then extend the Hong-Tamer moment estimation procedure to this general multivariate scenario. Moreover, the Hong-Tamer moment estimation procedure is based on the unconditional expectation $Em(X,\theta) = EH(X,\theta)$ for some function H. Example shows this techniques does not work in some cases. In this paper, we will propose an estimation procedure based on the condition expectation $E(m(X,\theta)|Z)$. Large sample properties of the proposed estimation procedure when X is one-dimensional are discussed.

Keywords: Nonlinear Statistical Inference; Measurement Error; Laplace Distribution; Bias Correction.

2.211 The tail empirical process of regularly varying functions of geometrically ergodic Markov chains

P. Soulier

Modal'X, univ. Paris 10

Abstract: We consider a stationary regularly varying time series which can be expressed as a function of a geometrically ergodic Markov chain. We obtain practical conditions for the weak convergence of weighted versions of the multivariate tail empirical process. These conditions include the so-called geometric drift or Foster-Lyapunov condition and can be easily checked for most usual time series models with a Markovian structure. We illustrate these conditions on several models and statistical applications.

2.212 On the Kernel Choice in RKHS-based Two-Sample Tests

B. Sriperumbudur[1,*], A. Gretton[2] and D. Sejdinovic[3]

[1] Pennsylvania State University; bks18@psu.edu
[2] University College London; arthur.gretton@gmail.com
[3] University of Oxford; dino.sejdinovic@gmail.com

Abstract: Embedding probability measures into a reproducing kernel Hilbert space as a mean element allows to define a metric on the space of probability measures [3]. Based on this metric, a non-parametric kernel two-sample test has been proposed by [1] where the test statistic, which is a U-statistic, measures the distance between the samples. While the test is asymptotically consistent, the choice of kernel is critical in ensuring that the test has high power. Usually, some heuristics are used to select the kernel as it is difficult to carry out the exact power analysis to understand the effect of kernel choice on the power. This difficulty raises from the fact that the U-statistic converges in distribution (under the null) to a countable sum of weighted chi-squared random variables. To address this issue, in this work, we propose an alternate test statistic whose asymptotic distribution under the null and the alternative are both Gaussian distributions, which therefore allows for exact power computations. Using this, we propose a kernel selection strategy that maximizes the test power for a given level and show the resultant test (using the selected kernel) to be consistent. Through numerical experiments, we demonstrate a significant improvement in performance of the proposed kernel test over various kernel selection heuristics. Since kernel based tests are equivalent [2] to energy distance based tests [4**?**], the ideas in our work can be used to address the question of the choice of distance function in the distance-based tests.

Keywords: Hilbert space embedding; Reproducing Kernel Hilbert Space; Two-sample tests

References

[1] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

[2] Sejdinovic, D., Sriperumbudur, B. K., Gretton, A., Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291.

[3] Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *J. of Machine Learning Research*, 11:1517–1561.

[4] Székely, G. and Rizzo, M. (2004). Testing for equal distributions in high dimension. *Interstat* 5.

[5] Székely, G. and Rizzo, M. (2005). A new test for multivariate normality. *J. Multivariate Anal.* 93:58–80.

2.213 Geometric Functional Data Analysis in Neuron Morphology

A. Duncan[1], E. Klassen[2], X. Descombes[3] and A. Srivastava[1,*]

[1] Dept of Statistics, Florida State University; a.duncan@stat.fsu.edu, anuj@stat.fsu.edu
[2] Dept of Mathematics, Florida State University; klassen@math.fsu.edu
[3] Morpheme Group, INRIA; xavier.descombes@inria.fr

Abstract: There is a great interest in statistical analysis of morphological structures of neurons, especially for evaluating cognitive abilities and detecting onset of cognitive diseases. Here one seeks tools for quantifying shape differences between neurons, modeling variabilities within and across neuron populations, and for testing/classifying neurons. The challenges include: (1) variability in size and shape of the main branch (axon), and (2) different numbers, sizes, and shapes of the side branches. In other words, the neurons differ in both geometry and topology, and that makes it difficult to model their shapes. The most difficult part in shape comparison is registration of points across neurons. In this paper we develop a framework for shape analysis of neuronal tree with following features: (1) a main branch viewed as a parameterized curve in $\mathbb{R}^3$, and (2) a finite number of secondary branches that emanate from the main branch at arbitrary points, each represented as a parametrized curve in $\mathbb{R}^3$. This framework is an extension of elastic functional data analysis, and shape analysis of Euclidean curves where one compares individual curves while being invariant to certain shape-preserving transformations. We define a shape space of these tree representations and impose an elastic metric on it compare different trees. The metric is based on a comparison of the shapes of main branch and the side branches, and locations and number of the side branches. The resulting geodesic paths between neurons show the main

branch of one tree deforming into the main branch of the other, while optimally deforming/sliding/creating/destroying the side branches of one into the side branches of other. Using this metric, we define sample mean and variances, and perform principal component analysis of shape data. Furthermore, we cluster and classify neurons into wild types and mutations using this approach. We present some preliminary results using axonal trees taken from the Neuromorpho database.

Keywords: neuron morphology; elastic shape analysis; functional data analysis; shape clustering.

2.214 Mixtures, Measurement Error, and Model Selection

L. A. Stefanski[1*]

[1*] Department of Statistics, North Carolina State University, stefansk@ncsu.edu.

Abstract: Model selection is more difficult in the presence of measurement error. Conversely, techniques and concepts from measurement error modelling can be used to facilitate variable selection in very general modeling frameworks. This talk surveys some recent and ongoing research involving both variable selection and measurement error modeling (MEM). Time permitting, methods for variable selection in parametric and nonparametric measurement error models will be described, in addition to the use of measurement error modelling techniques to facilitate variable selection in non-measurement error models.

Keywords: Lasso; Measurement error; Model selection; Penalty methods;Variable selection.

2.215 Tail chains for asymptotically independent Markov chains

I. Papastathopoulos[1], K. Strokorb[2,*], J.A. Tawn[3] and A. Butler[4]

[1] University of Edinburgh; i.papastathopoulos@ed.ac.uk
[2] University of Mannheim; strokorb@math.uni-mannheim.de
[3] Lancaster University; j.tawn@lancaster.ac.uk
[4] Biomathematics and Statistics Scotland; adam.butler@bioss.ac.uk

Abstract: The extremal behaviour of a Markov chain is typically characterized by its tail chain. For asymptotically independent chains recent results do not cover well-known asymptotically independent processes such as Markov processes with a Gaussian copula between consecutive values. We use more sophisticated limiting mechanisms that cover a broader class of asymptotically independent processes than current methods, including an extension of the canonical Heffernan-Tawn normalization scheme, and reveal features which existing methods reduce to a degenerate form associated with non-extreme states.

Keywords: Asymptotic independence; conditional extremes; Markov chains; hidden tail chain; tail chain

References

[1] R. Kulik and P. Soulier (2015). Heavy tailed time series with extremal independence. *Extremes*, **18(2)**, 273–299.

[2] I. Papastathopoulos and K. Strokorb and J.A. Tawn and A. Butler (2015). Extreme Events of Markov Chains. *arXiv preprint arXiv:1510.08920*.

2.216 On the weak convergence of the kernel density estimator in functional spaces

G. Stupfler[1]

[1] Aix Marseille Université, CNRS, EHESS, Centrale Marseille, GREQAM UMR 7316, 13002 Marseille, France

Abstract: Since its introduction, the pointwise asymptotic properties of the kernel estimator $\widehat{f}_n$ of a probability density function f on $\mathbb{R}^d$, as well as the asymptotic behaviour of its integrated or uniform errors, have been studied in great detail. Its weak convergence in functional spaces, however, is a more difficult problem. In this talk, we show that any Borel measurable weak limit of centered and rescaled versions of $\widehat{f}_n$ in a weighted L^p space on $\mathbb{R}^d$, with $1 \leq p < \infty$, must in fact be 0. An extension to the case $p = \infty$ of convergence in the supremum topology is discussed as well, and in both cases, we provide simple conditions for proving or disproving the existence of this Borel measurable weak limit.

Keywords: Kernel density estimator; Weak convergence; L^p space.

2.217 An Equation-by-Equation Estimator of a Multivariate Log-GARCH-X Model of Financial Returns

C. Francq[1] and G. Sucarrat[2,*]

[1] CREST and University of Lille; christian.francq@univ-lille3.fr
[2] BI Norwegian Business School; genaro.sucarrat@bi.no

Abstract: Estimation of large financial volatility models is plagued by the curse of dimensionality: As the dimension grows, joint estimation of the parameters becomes infeasible in practice. This problem is compounded if covariates or conditioning variables ("X") are added to each volatility equation. The problem is especially acute for non-exponential volatility models (e.g. GARCH models), since there the variables and parameters are restricted to be positive. Here, we propose an estimator for a multivariate log-GARCH-X model that avoids these problems. The model allows for feedback among the equations, admits several stationary regressors as conditioning variables in the X-part (including leverage terms), and allows for time-varying conditional covariances of unknown form. Strong consistency and asymptotic normality of an equation-by-equation least squares estimator is proved, and the results can be used to undertake inference both within and across equations. The flexibility and usefulness of the estimator is illustrated in two empirical applications.

Keywords: Exponential GARCH; Multivariate log-GARCH-X; VARMA-X; Equation-by-Equation Estimation (EBEE); Least Squares

2.218 Partially Informative Normal and Bayesian Partial Spline

Dongchu Sun[1*] and Sifan Liu[2]

[1] Univ. of Missouri and East China Normal Univesity; sund@missouri.edu, [2] Rutgers University; sifan.liu@rutgers.edu

Abstract: There is a well-known Bayesian interpretation of function estimation by spline smoothing using a limit of proper normal priors. This limiting prior has the same form with Partially Informative Normal (PIN), which was introduced in

Sun, Tsutakawa, and Speckman (1999). We derive that, under certain conditions, the linear transformation of PIN random variable and the linear combination of PIN random variables both follow PIN distributions. We apply these results to two extensions of univariate smoothing spline problem. One is large p, small n regression problem associated with the first case. We discuss about the conditions that the smooth component and response curve are estimable. The other is partial spline models associated with the second case. We provide necessary and sufficient conditions for the propriety of the posterior for both linear and smooth components, with non-informative priors on the variance of noise and the noise-signal variance ratio. We develop MC algorithms using constructive random posteriors, and perform simulation studies to show the advantages of partial splines. We also apply partial spline models to build multiple yield curves.

Keywords: Smoothing Spline; Partial Informative Normal; Bayesian Computation; Noise-signal Ratio; Constractive Random Poserior.

References

[1] Sun, D., Tsutakawa, R. K. and Speckman, P. L. (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika*, **86**, 341–350.

2.219 Measuring (Nonlinear) Granger Causality in Quantiles

X. Song[1] and A. Taamouti[2,*]

[1] Peking University; sxj@gsm.pku.edu.cn
[2] Durham University Business School; abderrahim.taamouti@durham.ac.uk

Abstract: In this paper we introduce first measures of Granger causality in quantiles between random variables. These measures are able to detect and *quantify* nonlinear causal effects. The new measures are based on nonparametric quantile regressions and defined as logarithmic functions of restricted and unrestricted expectations of check loss function. They

are easily and consistently estimated by replacing the unknown expectations of check loss function by their nonparametric kernel estimates. We derive the Bahadur representation of the nonparametric estimator of the measures, which we use to establish the consistency. We also provide the asymptote distribution of the estimator of the measure, which we use to build tests for their statistical significance and we study its properties under some local alternatives. We establish the validity of smoothed local bootstrap that one can use in finite sample settings to perform statistical tests. Monte Carlo simulations reveal that the proposed test has good finite sample size and power properties for a variety of data-generating processes and different sample sizes. Finally, the empirical importance of measuring nonlinear causality in quantile is also illustrated. We quantify the degree of nonlinear predictability of equity risk premium using variance risk premium, unemployment, inflation, and effective federal funds rate. Our empirical results show that the variance risk premium and effective federal funds rate are a good predictor of the quantiles of risk premium. The variance risk premium is able to predict both the centre and the lower and upper quantiles of the distribution of stock returns, whereas effective federal funds rate only helps to predict the lower and upper quantiles without any effect on the centre of distribution. Finally, the macro variables such as unemployment and inflation have no effect on the quantiles (distribution) of stock returns.

Keywords: Measures of Granger causality; Granger causality in quantiles; Conditional quantile function; Time series; Local linear estimator; Smoothed local bootstrap.

2.220 On making use of presumed values of a linear functional in its statistical estimation

Yu.G. Dmitriev[1], F.P.Tarasenko[1,*], P.F.Tarassenko[1]

[1] Tomsk State University; dmit@mail.tsu.ru, ftara@mfu.tsu.ru, ptara@mail.tsu.ru

Abstract: Methodology and techniques of usage of supplementary a priori information about the statistically estimated functional of unknown probability distribution in nonparametric setting were considered in [1], [2]. The additional information may be received from different sources and may have diverse relations to the underlying distribution (e.g. theoretical predictions, results of previous experiments and/or computer modeling of other manifestations of the same distribution, etc.). This paper considers a case when the supplementary information is an expert's imagination about the value to be estimated, made by guess based on his knowledge and experience. Let us call such information a priory guess. The problem was how to improve the quality of statistical inference by combining this information with a sample data during its processing. A solution to this problem for the case of a single guess was presented in [3], [4], [5] in a form of the 'combined' estimate of the functional. Now, in this paper, an approach to solving this problem in nonparametric case by combining sample (a posteriori) information with several (a priori) guesses is considered. The adaptive estimates of the functional are constructed, that are converging (in the minimal mean-square error sense) with growing size of a sample to the optimal ones. The conditions are defined under which the adaptive estimate demonstrates better MSE than traditional nonparametric estimate. Some numerical examples are given that illustrate interdependence between number of guesses, their deviations from real value of the functional, and probabilistic characteristics of the estimate.

Keywords: Linear functional; A priori information; Prior guess; Combined estimator; Nonparametric adaptive estimation.

References

[1] Dmitriev, Yu.G. and Tarasenko, F. P. (1978) On the use of a priori information in estimated liner functionals of distribution. *Problems of Control and Inform Theory*, **Vol.7, Issue 6**, 459–469.

[2] Dmitriev, Yu. G. and Tarassenko, P. F. (1992) The use of a priori information in the statistical processing of experimental data. *Russian Physics Journal*, **Vol. 35, Issue 9**, 888–893.

[3] Tarima, S. S. and Dmitriev, Yu. G. (2009) Statistical estimation with possibly incorrect model assumptions. *Tomsk State University Journal of Control and Computer Science*, **Vol. 8, issue 4**, 87–99.

[4] Dmitriev, Yu. G. and Tarassenko, P. F. and Ustinov, Yu. K. (2014) On Estimation of Linear Functional by Utilizing a Prior Guess. *A.Dudin et al. (Eds.): ITMM 2014, Communications in Computer and Information Science*, **487**, 82–90.

[5] Dmitriev, Yu. G. and Tarassenko, P. F. (2015) On Adaptive Estimation Using a Prior Guess. *Proceedings. The International Workshop, Applied Methods of Statistical Analysis. Nonparametric Approach. Novosibirsk, Russia. Novosibirsk: NSTU publisher*, 49–55.

2.221 Concentration results in extreme value theory

M. Thomas[1]

[1] Chalmers University of Technology, Gothenburg; maudt@chalmers.se

Abstract: Since univariate extreme value theory is essentially based on asymptotic theorems, the elaboration of adaptive estimation procedures or oracle inequalities is challenging. The purpose of this talk is to show how the development of concentration inequalities in extreme value theory allows us to derive oracle type results. The first part of this talk presents concentration inequalities for order statistics of a sample of random variables; combining Rényi representation for exponential order statistics and Bobkov and Ledoux Poincaré inequality for functions of independent exponentially distributed random variables. The second part deals with the development of an adaptive version of the Hill estimator based on non asymptotic bounds of this estimator.

Keywords: Concentration inequalities; Extreme value theory; Order statistics; Hill estimator, Lepski method.

2.222 Risk measures for regularly varying sequences of random length

Charles Tillier and Olivier Wintenberger

Univ. Paris Ouest-Nanterre, France
Univ. Of Copenhague, Denmark

Abstract: *Risks evaluation* is now a major issue in our society. In dietary risk, hydrology, nuclear security, finance or insurance, for which the risk analysis has become essential, *risk theory* plays a leading role and is now in the application field of the probability tools and the statistical methods. Increasingly present in the scientific literature and relevant for public authorities, it is relatively a new domain for mathematicians and has only been studied for the past couple years. In dietary risk for instance, once toxicologists have determined levels of contamination from which the exceedence can have some adverse effects for the human health, mathematicians can use these thresholds to build risk models and develop risk indicators. To face these risk problems, statisticians first called out to static models. However, due the dynamic nature of the studied phenomena (accumulation, elimination, delayed... phases), it seemed even more revelant to take into account the evolution through time. That is why they later resorted to *ruin models*, which are dynamic models that usually describe the evolution of a stock with entries and exits by means of stochastic processes, in a general manner at continuous time. Besides, as *risk theory* usually copes with rare events, it often deals with functions of heavy tailed random variables (like sums or products) and more specifically with regularly varying random variables. In this context, most of the stochastic processes that intervene in these applications fields can be written from a sequence of random length whose components are random variables. Therefore, the main aim of our paper is the following : to build regular variations for regularly varying sequences whose length is driven by a random variable in order to develop risk measures.

By way of applications, we suggest risk indicators for a class of processes covered by our framework : the Shot Noise Processes. The goal is to supplement the information given by the ruin probability and the tail process. Precisely, we first bear our interest on the *Expected Severity*, a widely-used risk indicator in insurance, because it is an alternative to Value at Risk that is more sensitive to the shape of the losses distribution in the tail of the distribution. Then, we introduce an indicator called *Integrated Expected Severity*, which in the previous context gives information on the total losses themselves. Lastly, we focus our interest on the *Expected Time Over a Threshold* which corresponds to the mean time spent by the process above a given threshold.

2.223 Root$-n$ consistent estimation of the marginal density of some stationary time series

L. Truquet[1]

[1] CREST Ensai & UMR 6625 CNRS IRMAR, University of Rennes 1; lionel.truquet@ensai.fr

Abstract: It is well known that, under some conditions, the density of a function of several independent random variables can be estimated at the usual parametric rate of convergence, using $U-$statistics arguments and kernel density estimation. In regression models and a few time series models, some similar results are also available for estimating the marginal density. In this talk, we will first make a review of the available results and next, we will extend these results to a general class of time series models, including GARCH processes.

Keywords: Kernel density estimation; Time series.

2.224 Sharp minimax and adaptive variable selection

Alexandre B.Tsybakov

CREST-ENSAE

Abstract: We derive non-asymptotic bounds for the minimax risk of variable selection under the expected Hamming loss in the problem of recovery of s-sparse vectors in $\mathbb{R}^d$ whose non-zero components are greater than $a > 0$. We get exact expressions for the non-asymptotic minimax risk as a function of (d, s, a) and find explicitly the minimax selectors. Analogous results are obtained for the probability of wrong recovery of the sparsity pattern. As corollaries, we derive necessary and sufficient conditions for such asymptotic properties as almost full recovery and exact recovery. Moreover, we propose data-driven selectors that provide almost full and exact recovery adaptive to the parameters (s, a) of the classes.
This is a joint work with Cristina Butucea and Natalia Stepanova.

2.225 Block maxima based tail index estimation

M. Vaičiulis

Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania; marijus.vaiciulis@mii.vu.lt

Abstract: We continue the investigation of the tail index estimator, which is based on maxima over blocks of data and which was proposed in [3]. We construct a new parameterized tail index estimator, based on local maxima, and we prove the asymptotic normality of this new estimator in the case of i.i.d. observations. The optimal values of the tuning parameters are discussed, using the same methodology as in [1] (see also [2]).

Keywords: Asymptotic normality–Tail index estimation

References

[1] Paulauskas V. and Vaičiulis, M. (2011). Several modifications of DPR estimator of the tail index. *Lithuanian Mathematical Journal*, **51(1)**, 36–50.

[2] Paulauskas V. and Vaičiulis, M. (2015). A class of new tail index estimators. *Accepted for publication in Annals of the Institute of Statistical Mathematics*, DOI: 10.1007/s10463-015-0548-3.

[3] Vaičiulis, M. (2014). Local-maximum-based tail index estimator. *Lithuanian Mathematical Journal*, **54(4)**, 503–526.

2.226 Solving inverse problems in econometrics using a mollification approach

P. Maréchal[1] and A. Vanhems[2,*]

[1] University of Toulouse; pr.marechal@gmail.com
[2] University of Toulouse, TBS and TSE; a.vanhems@tbs-education.fr

Abstract: The overall purpose of this research is to provide new tools for the analysis of structural econometric models. Our focus will be on nonparametric procedures that permit estimation of causal effects while avoiding the strong asumptions required by parametric procedures. More specifically, our research is motivated by the analysis of structural causal models with *endogeneity*. Endogeneity may be due to omitted variables, measurement errors, or simultaneity. Nonparametric regression with endogeneity gives rise to an ill-posed inverse problem. Our contribution is to solve this problem using a mollification approach. Mollification has been investigated for linear ill-posed equations, specifically for deconvolution in signal and image processing, by [2], and has been recently extended by [1] and [3] to more general situations. We investigate the use of this technique in the econometric context and apply it to consumer demand estimation and the analyse of Engel curves.

Keywords: Nonparametric Instrumental Regression; Inverse Problems; Mollification.

References

[1] Bonnefond, X. and Maréchal, P. (2009). A variational approach to the inversion of some compact operators. *Pacific Journal of Optimization*, **5**, 97–110.

[2] Lannes, A., Roques, S. and Casanove, M.-J. (1987). Stabilized reconstruction in signal and image processing; Part I: partial deconvolution and spectral extrapolation with limited field. *Journal of Modern Optics*, **34**, 161–226.

[3] Maréchal, P. and Mischra, S.K. (2016). Targeted solutions to linear ill-posed problems: a generalization of mollification. *submitted*.

2.227 Wilks' Phenomenon in Two-Step Semiparametric Empirical Likelihood Inference

F. Bravo[1], J.C. Escanciano[2] and I. Van Keilegom[3,*]

[1] Department of Economics, University of York; francesco.bravo@york.ac.uk
[2] Department of Economics, Indiana University; jescanci@indiana.edu
[3] Institute of Statistics, Université catholique de Louvain; ingrid.vankeilegom@uclouvain.be

Abstract: In both parametric and certain nonparametric statistical models, the empirical likelihood ratio satisfies a nonparametric version of Wilks' theorem. For many semiparametric models, however, the commonly used two-step (plug-in) empirical likelihood ratio is not asymptotically distribution-free, that is, Wilks' phenomenon breaks down. In this paper we suggest a general approach to restore Wilks' phenomenon in two-step semiparametric empirical likelihood inferences. The main insight consists in using as the moment function in the estimating equation the influence function of the plug-in sample moment. The proposed method is general, leads to distribution-free inference and it is less sensitive to the first-step estimator than alternative bootstrap methods. Several examples and a simulation study illustrate the generality of the procedure and its good finite sample performance.

Keywords: Empirical likelihood; Semiparametric inference; Stochastic equicontinuity; Wilks' phenomenon

2.228 Parameter estimation of distributions with heavy tailes

D. Politis[1], V. Vasiliev[2,*] and S. Vorobeychikov[2]

[1] University of California at San Diego; dpolitis@ucsd.edu
[2] Tomsk State University; vas@mail.tsu.ru, sev@mail.tsu.ru

Abstract: The estimation problem of the heavy tail index is considered. A new class of heavy tail distributions is introduced. It is shown that this class is very similar to the classical one. Examples of well-known distributions from both these classes are given. The new esimation approach based on the truncated estimation method is proposed. It makes possible to construct estimators with guaranteed accuracy based on a sample of fixed size. This method is applied to estimation of heavy tail indexes of the Pareto type distributions, as well as Hall's and log-gamma densities. It is shown that in some cases the proposed estimators achieve the optimal parametric rate of convergence. Given simulations confirm theoretical results.

Keywords: Heavy tails; Parameter estimation; Truncated estimation method; Finite sample size; Optimal convergence rate.

2.229 Nonparametric estimation of Markov switching model states under essential uncertainties

A.V. Dobrovidov[1] and V.O. Vasilyev[2,*]

[1] Laboratory at the V.A. Trapeznikov Institute of Control Sceinces, Russian Academy of Sciences, Moscow, Russia; dobrovidov@gmail.com
[2] Moscow Institute of Physics and Technology (State University) and junior researcher at the V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia; evil.vasy@gmail.com

Abstract: The autoregressive hidden Markov chain (AR-HMC) is widely used for various problems, including signal and image processing, econometrics, recognition, health sciences and etc. The AR-HMC enhances the HMC architecture by introducing a direct stochastic dependence between observations. In this paper, we develop a method of nonlinear filtering of a hidden (unobservable) Markov chain with two states (for simplicity) in AR-HMC model. This Markov chain controls coefficients of AR(p) model. The problem is to estimate the states of Markov chain by statistically dependent observations drawn from AR(p) model in the case of an unknown probability transition matrix and prior probabilities of the Markov chain, and completely unknown conditional distribution describing one of the two AR models. Another conditional distribution is proposed to be known. In this statement, the problem cannot be parameterized and consequently solved by the well-known EM-method [1]. Therefore, the proposed solution is based on the kernel nonparametric approach adapted to dependent observations [2]. We show, via simulations, that with an increase in sample size the performance of the proposed algorithm approaches the performance of the optimal (Bayes) nonlinear filtering developed for the first time by [3]. This method can be classified as an unsupervised estimation technique but, in contrast to EM-method, under nonparametric uncertainty.

Keywords: The autoregressive hidden Markov chain; Nonparametric uncertainty; Kernel estimation.

References

[1] Baum L.E., Petrie T., Soules G. and Weiss N. (1970). A maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, **41**, No. 1, 164–171.

[2] Alexander Dobrovidov, Gennady Koshkin and Vyacheslav Vasiliev (2012). *Non-parametric state space models*. Kendrick Press. USA.

[3] Stratonovich R.L. (1960). Conditional Markov Processes. *Theory of Probability and its Applications*, **5**, No. 2, 156–178.

2.230 On some high-dimensional tests for Principal Component Analysis.

Chr. Cutting[1], D.Paindaveine[1] and T.Verdebout[1,*]

[1] Université libre de Bruxelles; christine.cutting@ulb.ac.be, dpaindav@ulb.ac.be, tverdebo@ulb.ac.be

Abstract: Principal Component Analysis (PCA) is one of the most important tools in multivariate analysis. Testing procedures related to high-dimensional PCA have been widely studied recently. In particular, many papers tackled the problem of testing sphericity or unit covariance against spiked covariance matrices. In this work, we assume that the high-dimensional model has a spiked covariance matrix structure and we consider the problem of detecting the direction of the spike.

Keywords: Principal component analysis; High-dimensional statistics; Hypothesis testing.

2.231 Maximin tests for symmetry of circular data based on the characteristic function

Simos Meintanis[1] and Thomas Verdebout[2,*]

[1] University of Athens and North-West University; simosmei@econ.uoa.gr
[2] Université libre de Bruxelles (ULB); tverdebo@ulb.ac.be

Abstract: We consider inference for circular data based on the empirical characteristic function. More precisely, we provide tests for reflective symmetry on the circle based on the imaginary part of the empirical characteristic function. We show that the proposed tests enjoy many attractive features. In particular, we obtain that they are locally and asymptotically maximin in the Le Cam sense under sine-skewed alternatives in the specified mean direction case. For the unspecified mean direction case, we provide corrected versions of the original tests that keep very nice asymptotic power properties. Results are illustrated on a well-known dataset and checked via Monte-Carlo simulations.

Keywords: Circular data; Reflective symmetry; Characteristic function.

2.232 Detection and Feature Selection in Sparse Mixture Models

N. Verzelen[1*] and E. Arias-Castro[2]

[1] INRA, Montpellier; nicolas.verzelen@supagro.inra.fr
[2] University of California, San Diego; eariasca@ucsd.edu

Abstract: We consider Gaussian mixture models in high dimensions, focusing on the twin tasks of detection and feature selection. Under sparsity assumptions on the difference in means, we derive minimax rates for the problems of testing and of variable selection. We find these rates to depend crucially on the knowledge of the covariance matrices and on whether the mixture is symmetric or not. We establish the performance of various procedures, including the top sparse eigenvalue of the sample covariance matrix (popular in the context of Sparse PCA), as well as new tests inspired by the normality tests of [1]. This talk is based on [2].

Keywords: Gaussian mixture models; Detection of mixtures; Feature selection for mixtures; clustering; Projection tests

References

[1] Malkovich, J. F. and A. Afifi (1973). On tests for multivariate normality. *Journal of the American Statistical Association* **68** (341), 176–179.

[2] Verzelen, N. and E. Arias-Castro (2014). Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*.

2.233 The spatial sign covariance matrix

Daniel Vogel

Institute for Complex Systems and Mathematical Biology, University of Aberdeen, UK; daniel.vogel@abdn.ac.uk

Abstract: We explore the use of the spatial sign covariance matrix (SSCM) for various aspects of multivariate data analysis. A central research question concerning the SSCM is how its eigenvalues are related to those of the covariance at elliptical distributions. In dimension 2, this relation is known explicitly, which can be used to construct a robust pairwise correlation estimator. This estimator has a variety of nice properties. It is fast to compute, distribution-free within the elliptical model, as efficient as similarly robust estimators, and its asymptotic variance admits an explicit form, which facilitates inferential procedures. Its efficiency may be further improved by a prior componentwise standardization. We show that the loss due to having to estimate the marginal scales – as compared to known scales – is nil asymptotically. We derive a variance-stabilizing transformation in the same vein as Fisher's z-transformation, but which is valid for all elliptical distributions. We further study the SSCM in higher dimensions and propose a change-point test for multivariate dependence based on the SSCM. The asymptotic distribution under the null is derived for stationary, short-range dependent sequences without any moment assumption. The talk is based on joint work with Alexander

Dürre, Roland Fried and David Tyler.

Keywords: Change-point analysis, Correlation estimator; Elliptical distribution; Fisher's z-transformation; Near epoch dependence in probability.

2.234 A new two-sample test for skewed populations based on Edgeworth expansion and its application in high dimensional classification

Bo Tong[1], Haiyan Wang[2], Huaiyu Zhang[2]

[1] Research and Development Dept R43V, AbbVie, North Chicago, IL 60064;
[2] Department of Statistics, Kansas State University, Manhattan, Kansas, USA

Abstract: Various tests have been created to compare the means of two populations in many scenarios and applications. Tests for skewed data, however, are not well studied even though they are often needed in pharmaceutical study and agricultural economics. In this paper, we propose a new test to improve the accuracy and power under skewed populations

with moderate sample size. The proposed work starts with derivation of a first order Edgeworth expansion for the pooled two-sample t statistic. Then a new test rejection region was formed based on Cornish Fisher expansion of quantiles. This test corrects for first order population skewness that was ignored in two-sample t-test. We report the theoretical type I error and power properties of the newly proposed test and the large sample t tests. We also provide the detailed conditions under which the proposed test gives better power than the two-sample large sample test. Compared with commonly used two-sample parametric and nonparametric tests, our new test gives better power for skewed data. This new test combined with Naive Bayes method provide a powerful classification algorithm for classification and variable selection in high dimensional data.

Keywords: Edgeworth expansion; Hypothesis testing; Naive Bayes algorithm; High dimensional classification

2.235 Utilizing Latent Features in Clustered Functional Data

N. Wang

[1] Department of Statistics, University of Michigan, USA.; nwangaa@umich.edu

Abstract: In this presentation, we explore the use of latent features embedded in the clustered/grouped functional data to enhance model flexibility and prediction efficiency. Theoretical properties justifying our modeling strategies will be presented. We will illustrate the shared and contrast information reflects by these latent features embedded in different sub-groups of subjects. Criterion, such as out-of-sample prediction, were employed to gauge the use of different types of features and the bases on which they were evaluated. Effectiveness of the new methods is demonstrated using both synthetic data and data collected through medical studies.

Keywords: Clustering; Latent features; Orthonormal basis; Spline.

2.236 Sufficient Dimension Reduction under Dimension Reduction Based Imputation with Predictors Missing at Random

Xiaojie Yang and Qihua Wang

Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100190, China

Abstract: In some practical problems, a subset of predictors is subject to missingness when the dimension of the predictor vector is high. In this situation, the standard sufficient dimension reduction(SDR) methods cannot be applied directly to avoid the " curse of dimension ". In this paper, a dimension reduction based kernel imputation method assisted by the complete-case analysis is developed to handle the dimension-reduction problems with predictors missing at random. The sliced inverse regression(SIR) is used to illustrate this procedure. It is shown that the proposed kernel imputation estimator of SIR is asymptotically normal under some mild conditions. The finite-sample performances of the proposed method are evaluated through comprehensive simulations and a real data set is analyzed to illustrate the proposed method. Furthermore, the general idea of this article is extended to other two popular SDR methods, namely sliced average variance(SAVE) and principal Hessian direction(PHD).

Keywords:

2.237 A new nonparametric test for checking the equality of the correlation structures of two time series

Lei Jin[1] and Suojin Wang[2,*]

[1]Department of Mathematics and Statistics, Texas A&M University, Corpus Christi, TX 78412, USA; lei.jin@tamucc.edu
[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA; sjwang@stat.tamu.edu

Abstract: In this talk, we consider an order selection test to check the equality of two independent stationary time series in their correlation structures. The asymptotic distribution of the order selection test statistic under the null hypothesis is obtained. For many existing tests, consistency against general alternative hypotheses has not been established. On the other hand, we show that the proposed test is consistent not only under any fixed alternative hypothesis but also under a sequence of local alternative hypotheses. A simulation study is conducted to examine the finite sample performance of the test in comparison to some existing methods. We also apply the proposed test to an analysis of a biomedical data set.

Keywords: Autocorrelation; Local alternative; Order selection test; Spectral density.

2.238 Statistical and computational tradeoffs in estimation of sparse principal components

T. Wang[1], Q. Berthet[1] and R. J. Samworth[1,*]

[1] Cambridge University; t.wang@statslab.cam.ac.uk, q.berthet@statslab.cam.ac.uk, r.j.samworth@statslab.cam.ac.uk

Abstract: In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. In this paper we show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a Restricted Covariance Concentration condition, we show that there is an effective sample size regime in which no randomised polynomial time algorithm can achieve the minimax optimal rate. We also study the theoretical performance of a (polynomial time) variant of the well-known semidefinite relaxation estimator, revealing a subtle interplay between statistical and computational efficiency.

Keywords: Computational lower bounds; Planted Clique problem; Polynomial time algorithm; Sparse principal component analysis

2.239 Estimators for Markov chains with missing observations

W. Wefelmeyer

Department of Mathematics, University of Cologne, Weyertal 86–90, 50931 Cologne, Germany; wefelm@math.uni-koeln.de

Abstract: Discrete-time real-valued Markov chains are sometimes observed at certain time points only. The unobserved time points may be deterministic and periodic, or the gap length may be random and may also depend on the last observed state. This is no problem for estimating the one-dimensional marginal distribution. However, if we want to estimate joint or conditional distributions of the chain, in particular the transition distribution, can we exploit the information in pairs of observations separated by an unobserved gap? We discuss when and how this is possible in nonparametric and in autoregressive models. We point out similarities to mixture models and to regression models with observations missing at random. This is joint work with Anton Schick and Ursula U. Müller.

Keywords: Nonparametric estimator; Markov chain; Missing Observations.

2.240 Robustness of quadratic inference function estimators

Samuel Müller[1], Suojin Wang[2] and A.H.Welsh[3,*]

[1] University of Sydney; samuel.mueller@sydney.edu.au
[2] Texas A&M University; sjwang@stat.tamu.edu
[3] The Australian National University; Alan.Welsh@anu.edu.au

Abstract: Quadratic inference function estimators for the regression parameter in regression models for longitudinal data were introduced by [1] to improve on the efficiency of generalized estimating equations estimators. [2] argued that quadratic inference function estimators are also robust against outliers, making them preferable to generalized estimating equations estimators. In this talk, we discuss the robustness properties of quadratic inference function estimators, revisiting particular cases to understand more deeply the generality of the conclusions of [2]. We show that robustness issues in generalised estimating equations estimation are more subtle than is generally believed and we had anticipated.

Keywords: Generalized estimating equations; Longitudinal data.

References

[1] Qu, A. and Lindsay, B.G. and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions *Biometrika*, **87**, 823–836.

[2] Qu, A. and Song, P.X.-K. (2004). Assessing robustness of generalized estimating equations and quadratic inference functions. *Biometrika*, **91**, 447–459.

2.241 Nonparametric Instrumental Variable Estimation Under Monotonicity

D. Chetverikov[1] and D. Wilhelm[2,*]

[1] Department of Economics, University of California at Los Angeles, USA; chetverikov@econ.ucla.edu
[2] Department of Economics, University College London, UK; d.wilhelm@ucl.ac.uk

Abstract: The ill-posedness of the inverse problem of recovering a regression function in a nonparametric instrumental variable model leads to estimators that may suffer from a very slow, logarithmic rate of convergence. In this paper, we show that restricting the problem to models with monotone regression functions and monotone instruments significantly weakens the ill-posedness of the problem. In stark contrast to the existing literature, the presence of a monotone instrument implies boundedness of our measure of ill-posedness when restricted to the space of monotone functions. Based on this result we derive a novel non-asymptotic error bound for the constrained estimator that imposes monotonicity of the regression function. For a given sample size, the bound is independent of the degree of ill-posedness as long as the regression function is not too steep. As an implication, the bound allows us to show that the constrained estimator converges at a fast, polynomial rate, independently of the degree of ill-posedness, in a large, but slowly shrinking neighborhood of constant functions. Our simulation study demonstrates significant finite-sample performance gains from imposing monotonicity even when the regression function is rather far from being a constant. We apply the constrained estimator to the problem of estimating gasoline demand functions from U.S. data.

Keywords: Nonparametric Instrumental Variable Model; Monotonicity.

2.242 On statistical inference based on numerical inversion of the empirical characteristic function

Viktor Witkovský

Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia; witkovsky@savba.sk

Abstract: The methods for making the exact statistical inference frequently require evaluation of the probability density function (PDF), the cumulative distributions function (CDF), and/or the quantile function (QF) of a random variable from its characteristic function (CF), which is defined as a Fourier transform of its probability distribution function. Working with CFs provides an alternative (frequently more simple) route, than working directly with PDFs and/or CDFs. However, the analytical derivation of the PDF and/or CDF by using the inverse Fourier transform is available only in special cases. Thus, in most practical situations, a numerical derivation of the PDF/CDF from the CF is an indispensable tool, see e.g. [1**? ? ? ?**]. The methods based on numerical inversion of the CFs can be used also in nonparametric settings. For example, in some (simple) situations the bootstrap distribution can be quickly and efficiently evaluated by numerical inversion of the CF, which is based on the empirical characteristic function. Here we shall present brief overview of some simple approaches for numerical inversion of the CFs, as e.g. the method based on approximation by discrete Fourier transform (DFT) and by application of the FFT (fast Fourier transform) algorithm for computing PDF/CDF of (univariate) continuous random variables. We shall also present several examples based on the empirical characteristic functions, to illustrate the applicability of this approach for making the exact statistical inference.

Keywords: Empirical characteristic function, Inverse Fourier transform; Bootstrap distribution; Gil-Pelaez inversion formula, Fast Fourier transform (FFT) algorithm.

References

[1] Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika* **38**, 481–482.

[2] Abate, J., Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10**, 5–88.

[3] Witkovský, V. (2001). On the exact computation of the density and of the quantiles of linear combinations of t and F random variables, *Journal of Statistical Planning and Inference* **94**, 1–13.

[4] Witkovský, V. (2001). Computing the distribution of a linear combination of inverted gamma variables, *Kybernetika* **37**, 79–90.

[5] Witkovský, V., Wimmer, G., Duby, T. (2015). Logarithmic Lambert $W \times \mathcal{F}$ random variables for the family of chi-squared distributions and their applications, *Statistics & Probability Letters* **96**, 223–231.

2.243 Big network data

P. J. Wolfe[1]

[1] University College London; p.wolfe@ucl.ac.uk

Abstract: How do we draw sound and defensible data-analytic conclusions from networks? This question has recently risen to the forefront of mathematical statistics, and it represents a fundamental challenge for data science. In this talk I will describe new large-sample theory that helps us to view and interpret networks as statistical data objects, along with the transformation of this theory into new statistical methods to model and draw inferences from network data in the real world. The insights that result from connecting theory to practice also feed back into pure mathematics and theoretical computer science, prompting new questions at the interface of combinatorics, analysis, probability, and algorithms.

Keywords: Big data; Network analysis; Nonparametric statistics.

2.244 Kernel Estimation Of Hazard Functions When Observations Have Dependent and Common Covariates

J. Wolter[1]

[1] University of Oxford, Oxford-Man Institute; james.wolter@economics.ox.ac.uk

Abstract: We propose a hazard model where dependence between events is achieved by assuming dependence between covariates. This model allows for correlated variables specific to observations as well as macro variables which all observations share. This setup better fits many economic and financial applications where events are not independent. Nonparametric estimation of the hazard function is then studied. Kernel estimators are shown to have similar asymptotic properties compared with the i.i.d. case. Mixing conditions ensure the asymptotic results follow. These results depend on adjustments to bandwidth conditions. Simulations are conducted which verify the impact of dependence on estimators. Bandwidth selection motivated by theoretical results is shown to improve performance.

Keywords: Hazard Estimation; Correlated Events; Dependent Covariates, Common Covariates, Kernel Estimation.

2.245 Variable Selection in Kernel Regression Using Measurement Error Selection Likelihoods

Kyle R. White, Leonard A. Stefanski and Yichao Wu

Dpt Statistics, North Carolina State University, USA; kwhite3@ncsu.edu, stefansk@ncsu.edu, wu@stat.ncsu.edu

Abstract: This paper develops a nonparametric shrinkage and selection estimator via the measurement error selection likelihood approach recently proposed by [**?**]sww2014. The Measurement Error Kernel Regression Operator (MEKRO) has the same form as the Nadaraya–Watson kernel estimator, but optimizes a measurement error model selection likelihood to estimate the kernel bandwidths. Much like LASSO or COSSO solution paths, MEKRO results in solution paths depending on a tuning parameter that controls shrinkage and selection via a bound on the harmonic mean of the pseudo-measurement error standard deviations. We use the small-sample- corrected AIC to select the tuning parameter. Large-sample properties of MEKRO are studied and small-sample properties are explored via Monte Carlo experiments and applications to data.

Keywords: Bandwidth selection; Feature selection; LASSO; Nadaraya-Watson; Nonparametric regression.

References

[1] Lenard Stefanski, Yichao Wu and Kyle White (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association*, **109**, 574–589.

2.246 Lead Lag Relationship among High Dimensional Time Series

H. Xiao[1,*] and D. Sun[1]

[1] Rutgers University, USA; hxiao@stat.rutgers.edu, diesun@eden.rutgers.edu

Abstract: Multiple time series often exhibits lead-lag relationship among its component series. It is very challenging to identify this type of relationship when the number of series is large. We study the lead-lag relationship in the high dimensional context, using the maximum cross correlations and some other variants. Our result can also be used to test whether there is a correlation between two time series.

Keywords: Lead lag; High Dimensional Time Series; Extreme value theory

2.247 "Nonparametric" Meta Analysis with Unknown Study-specific Parameters

Min-ge Xie[1*]

[1] Rutgers University, mxie@stat.rutgers.edu,

Abstract: Meta-analysis is a valuable tool for combining information from independent studies in health care studies and other fields. However, most common meta-analysis techniques rely on distributional assumptions that are difficult, if not impossible, to verify. For instance, in the commonly used fixed-effects and random-effects models, we take for granted that the underlying study-level parameters are either exactly the same across individual studies or that they are realizations of a random sample from a population, often under a parametric distributional assumption. In this talk, we present a new framework for summarizing information obtained from multiple studies and make inference that is not dependent on any distributional assumption for the study-level parameters. Specifically, we assume the study-level parameters are unknown, fixed parameters and draw inferences about, for example, the quantiles of this set of parameters using study-specific summary statistics. This type of problem is known to be quite challenging in statistical inference (c.f., Hall & Miller, 2010). We utilize a novel resampling method via the confidence distributions of the study-level parameters to construct confidence intervals for the above quantiles. We justify the validity of the inference procedure asymptotically and compare the new procedure with the standard bootstrapping method. We also illustrate our proposal with simulations and real data related to health care policy studies. (Joint work with Brian Claggett and Lu Tian)

Keywords: Confidence distribution, Fusion Learning, Heterogeneous studies.

2.248 Statistical challenges in analyzing observational data on pregnancy

Ronghui Xu*

University of California, San Diego; rxu@ucsd.edu

Abstract: We consider exposures during pregnancy including vaccines and medication. There are multiple outcomes, including for example spontaneous abortion, preterm delivery, birth defects, etc. Some of these outcomes are intrinsically linked. Some, like spontaneous abortion, has two aspects: yes or no, and if yes, the timing of the event. This latter case has close connection with the cure rate models in the literature, although our data are somewhat different. In addition, the data are obtained in the context of observational studies, so there are features including left truncation, partial interval censoring, etc. We describe some of the methods of inference we have developed so far, including nonparametric maximum likelihood with an EM algorithm, and smoothing, as well as discuss future works to be done.

Keywords: Cure rate; I-spline; Interval censor; Left truncation; Nonparametric maximum likelihood.

2.249 Self-weighted LAD-based Inference for Heavy-tailed Threshold Autoregressive Models

Yaxing Yang[1] and Shiqing Ling[2]

[1]Department of Mathematics, Hong Kong University of Science and Technology; yyangaj@ust.hk
[2] Department of Mathematics, Hong Kong University of Science and Technology; maling@ust.hk

Abstract: The least squares estimator of the threshold autoregressive (TAR) model may not be consistent when its tail is less than or equal 2. Neither theory nor methodology can be applied to model fitting in this case. This paper is to develop a systematic procedure of statistical inference for the heavy-tailed TAR model. We first investigate the self-weighted least absolute deviation estimation for the model. It is shown that the estimated slope parameters are $\sqrt{n}$-consistent and asymptotically normal, and the estimated thresholds are n-consistent, each of which converges weakly to the smallest minimizer of a compound Poisson process. Based on this theory, the Wald test statistic is considered for testing the linear restriction of slope parameters and a procedure is given for inference of threshold parameters. We finally construct a sign-based portmanteau test for model checking. Simulations are carried out to assess the performance of our procedure and a real example is given.

Keywords: Asymptotic distribution, compound Poisson process, consistency, threshold autoregressive models, self-weighted least absolute deviation estimation, wald test, sign-based portmanteau test.

2.250 A new Framework for Statistical Analysis and Simulations in Medical Studies

C. Samir[1], A.-F. Yao [1]

[1] University of Clermont Auvergne, Clermont-Ferrand, France

Abstract: Statistical shape analysis plays an important role in various medical imaging applications. In particular, such methods provide tools for registering, deforming, comparing, averaging, and modeling anatomical shapes. Recent advances in medical applications offer increasingly detailed information on typical anatomical structures. However, there is a lack of validation techniques for automatic strategies, especially for multimodal data, i.e. coming from different types or levels of measurement. For example, standard methods to assess an accurate diagnosis use multiple modalities. However, some limitations due to non-localized pertinent information cannot be directly avoided. An interesting solution would be to statistically analyze shapes of real clinical data and provide enough random or simulated samples to validate registration step; registration of different modalities is key for fusing complementary information for diagnostic purposes. In this talk, we focus on recent methods for statistical shape analysis of elastic parametrized data (surfaces, images, signals) to disease classification and simulation of realistic samples. In particular, we present new statistical frameworks to generate realistic simulated data that can be used as ground truth when dealing with deformability of observations (e.g. cells).

Keywords: Shape Analysis; Fréchet Mean of Surfaces; Simulation; Warping; Registration.

2.251 Bootstrap Inference for Multiple Change-points in Time Series

Chun Yip Yau

Chinese University of Hong Kong; cyyau@sta.cuhk.edu.hk

Abstract: In this paper we propose two bootstrap procedures, namely parametric and block bootstrap, to approximate the asymptotic distribution of change-point estimator for piecewise stationary time series. The bootstrap procedures are then used to develop a generalized likelihood ratio scan method (GLRSM) for multiple change-points inference in piecewise stationary time series, which estimates the number and positions of change-points and provides confidence interval for each change-point. The computational complexity of using GLRSM for multiple change-points detection is as low as $O(n(\log n)^3)$ for a series of length n. Extensive simulation studies are provided to demonstrate the effectiveness of the proposed methodology under different scenarios. Applications to financial time series are also illustrated. Research supported in part by HKSAR-RGC Grants.

Keywords: Confidence interval; Likelihood ratio; Piecewise stationary time series models; Scan statistics; Structural break

2.252 Change Point Estimation of Brain Shape Data in Relation with Alzheimer's Disease.

L. Younes[1]

[1] Dpt of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, USA; laurent.younes@jhu.edu

Abstract: The manifestation of an event, such as the onset of a disease, is not always immediate and often requires some time for its repercussions to become observable. Slowly progressing diseases, and in particular neuro-degenerative disorders such as Alzheimer's disease (AD), fall into this category. The manifestation of such diseases is related to the onset of cognitive or functional impairment and, at the time when this occurs, the disease may have already had been affecting the brain anatomically and functionally for a considerable time. We consider a statistical two-phase regression model in which the change point of a disease biomarker is measured relative to another point in time, such as the manifestation of the disease, which is subject to right-censoring (i.e., possibly unobserved over the entire course of the study). We develop point estimation methods for this model, based on maximum likelihood, and bootstrap validation methods. The effectiveness of our approach is illustrated by numerical simulations, and by the estimation of a change point for atrophy in the context of Alzheimer's disease, wherein it is related to the cognitive manifestation of the disease. This work is a collaboration with Xiaoying Tang and Michael Miller, and was partially supported by the NIH.

Keywords: Change point estimation; Two-phase regression; Right censoring; Medical imaging

References

[1] X. Tang, M.I. Miller and Younes, L. (2016). Biomarker Change Point Estimation with Right Censoring in Longitudinal Studies (Submitted).

[2] L. Younes, M. Albert, M. I. Miller (2014). Inferring changepoint times of medial temporal lobe morphometric change in preclinical Alzheimer's disease, NeuroImage: Clinical, Volume 5, Pages 178-187

2.253 Multivariate Hyperrectangular Tolerance Regions Based on Data Depth

D. S. Young[1,*] and T. Mathew[2]

[1] Department of Statistics, University of Kentucky, USA; derek.young@uky.edu
[2] Department of Mathematics and Statistics, University of Maryland, USA; mathew@umbc.edu

Abstract: Multivariate tolerance region procedures have been developed primarily for the multivariate normal. In practice, multivariate normality will not always be an appropriate assumption, so alternatively, a nonparametric multivariate tolerance region could be used. Certain scientific applications require the construction of hyperrectangular regions, which could be constructed using a tolerance region approach. Such practical applications include the calculation of statistically-based design limits for the performance of various materials used in constructing nuclear cores [2], identifying editing thresholds for macro-level outliers in survey data [3], and setting reference regions in laboratory medicine and clinical chemistry to describe the variations of multivariate measurements or values in healthy individuals [1]. In this talk, we present two approaches for developing multivariate hyperrectangular tolerance regions – a fully nonparametric approach and a semiparametric approach. Both approaches are similar in that a multivariate ordering of the data using data depth is obtained and then a subjective ordering of the dimensions of the data is used to identify which extreme points are trimmed in order to construct the hyperrectangular tolerance region. The result is a highly-flexible multivariate method that provides bounds for diverse applications in areas like engineering, survey data analysis, and laboratory medicine.

Keywords: Calibration; Data depth; Multivariate density estimation; Order statistics; Trimming.

References

[1] Dong, X. and Mathew, T. (2015). Central Tolerance Regions and Reference Regions for Multivariate Normal Populations. *Journal of Multivariate Analysis*, **134**, 50–60.

[2] International Atomic Energy Agency (2008). *Best Estimate Safety Analysis for Nuclear Plants: Uncertainty Evaluation. Safety Report Series No. 52.* Vienna, Austria. IAEA Publishing Section.

[3] Young, D. S. and Mathew, T. (2015). Ratio Edits Based on Statistical Tolerance Intervals. *Journal of Official Statistics*, **31**, 77–100.

2.254 Community Detection of Sparse Networks: A Review and Two New Methods

B.Y. Jing[1], T. Li[1], N.C. Ying[1] and X.S. Yu[1]

[1] the Hong Kong University of Science and Technology;
majing@ust.hk, tlial@ust.hk, nying@ust.hk, xyuai@ust.hk

Abstract: Network is ubiquitous in the study of both natural science and sociology. It provides an effective way to represent the key information in a wide variety of systems. Examples include social networks, protein interaction networks and scientific publication networks. In this talk, I will start with a review of the motivation of exploring networks. Classical methods will be introduced, emphasizing the research on community detection, where groups of tightly connected nodes are detected. However, this task remains a problem when the connections in a network are very sparse. That is, most of the nodes are connected to only a very small set of nodes. We call this scenario a *sparse network*. In this case, if the only information we know is whether pairs of nodes are connected or not, we might not have enough information to do community detection, because the number of connections are also small. A more delicate description of the relationship between nodes might be in need. We propose two new methods to deal with this issue. In both methods, we design ways to develop a distance/similarity matrix from the adjacency matrix of a network. More information is extracted by either manipulating the distances between nodes or analyzing the random walks on the networks. Numerical results show the superiority of our methods. And I will explain the reason why they work.

Keywords: Stochatic Block Model; Community detection; Sparse Network; Spectral clustering.

2.255 Capture the neglected recovery of the dependence structure in directed and dynamic networks

T. Sit[1], Z. Ying[2] and Y. Yu[3,*]

[1] Department of Statistics, Chinese University of Hong Kong; tonysit@sta.cuhk.edu.hk
[2] Department of Statistics, Columbia University; zying@stat.columbia.edu
[3] Statistical Laboratory, University of Cambridge; y.yu@statslab.cam.ac.uk

Abstract: Directed and/or dynamic networks are of increasing interest but also under-studied. In this paper, we aim to recover the dependence structures across both time course and network structure. Multivariate counting processes with recurrent survival analysis techniques are used to model the individual behaviour with time dependent properties; whilst the network structure is modelled as the graphical model. The contribution of this paper is two-fold: a) a sandwich covariance estimator is available for more robust statistical inference and b) a conditional independent technique in *undirected* network is integrated into the *directed* network making the dependence structure characterisable and computationally feasible.

Keywords: Directed networks; Dynamic networks; Dependence structure; Survival analysis.

2.256 Joint inference on market and estimation risks in dynamic portfolios

C. Francq[1], and J-M. Zakoïan[2,*]

[1] CREST and Lille University; christian.francq@univ-lille3.fr
[2] CREST and Lille University; zakoian@ensae.fr

Abstract: We study the estimation risk induced by univariate and multivariate methods for evaluating the conditional Value-at-Risk (VaR) of a portfolio of assets. The composition of the portfolio can be time-varying and the individual returns are assumed to follow a general multivariate dynamic model. Under ellipticity of the conditional distribution, we introduce in the multivariate framework a concept of VaR parameter, and we establish the asymptotic distribution of its estimator. A multivariate Filtered Historical Simulation method, which does not rely on ellipticity, is studied. We also consider two univariate approaches based on past real or reconstituted returns. We derive asymptotic confidence intervals for the conditional VaR, which allow to quantify simultaneously the market and estimation risks. Potential usefulness, feasibility and drawbacks of the different univariate and multivariate approaches are illustrated via Monte-Carlo experiments and an empirical study based on stock returns.

Keywords: Confidence Intervals for VaR; Multivariate GARCH; Estimation risk; Filtered Historical Simulation; Elliptical Distribution

2.257 Scaling by subsampling for big data

M. Zetlaoui[1]

[1] Paris West University Nanterre La Défense; melanie.zetlaoui@u-paris10.fr

Abstract: The increasing capacity to collect data has improved much faster than our ability to process and analyze Big Datasets. The availability of massive information in the Big Data era suggests to use subsampling techniques as a remedy to the apparent intractability of learning from datasets of explosive size in order to break the current computational barriers. The aim is to recall some basic methods which were developed earlier (and not directed to big data), to actually show that it is possible to drop the hypothesis concerning the explicit knowledge of standardization in order to use subsampling for the construction of confidence regions or predictors at a larger scale. Several methods based on pattern recognition are tested on real data for classifying digits.

Keywords: Resampling methods; Big data; Rate of convergence estimation.

2.258 Probability-enhanced Sufficient Dimension Reduction for Binary Classification

Hao Zhang

University of Arizona (United States)

Abstract: Many sufficient dimension reduction (SDR) methods have been developed since the introduction of sliced inverse regression (SIR; Li, 1991). For binary classification problems, SIR suffers the limitation of estimating at most one direction since only two slices are available. We propose a new and flexible probability-enhanced SDR method for binary classification problems using the weighted support vector machine (WSVM). The key idea is to slice the data based on conditional class probabilities of observations rather than their binary responses. We show that the central subspace based on the conditional class probability is the same as that based on the raw binary response, which justifies the proposed slicing scheme and assures no information loss. Furthermore, in order to implement the new slicing scheme, one does not need exact probability values since the only required information is the relative ordering of probability values. The new SDR bypasses the probability estimation and employs the WSVM to directly estimate the order of probability values, based on which the slicing is performed. The performance of the proposed probability-enhanced SDR scheme is evaluated by both simulated and real data examples.

2.259 Variable Selection with Prior Information for Generalized Linear Models via the Prior LASSO Method

Y. Jiang[1], Y. He[2] and H. Zhang[3,*]

[1] Department of Statistics, Oregon State University, Corvallis, Oregon 97331-4606; yuan.jiang@stat.oregonstate.edu
[2] Nielsen Company, 770 Broadway, New York, New York 10003-9595; yunxiaohe@gmail.com
[3]Department of Biostatistics, Yale University School of Public Health, USA; heping.zhang@yale.edu

Abstract: LASSO is a popular statistical tool often used in conjunction with generalized linear models that can simultaneously select variables and estimate parameters. When there are many variables of interest, as in current biological and biomedical studies, the power of LASSO can be limited. Fortunately, so much biological and biomedical data have been collected and they may contain useful information about the importance of certain variables. This paper proposes an extension of LASSO, namely, prior LASSO (pLASSO), to incorporate that prior information into penalized generalized linear models. The goal is achieved by adding in the LASSO criterion function an additional measure of the discrepancy between the prior information and the model. For linear regression, the whole solution path of the pLASSO estimator can be found with a procedure similar to the Least Angle Regression (LARS). Asymptotic theories and simulation results show that pLASSO provides significant improvement over LASSO when the prior information is relatively accurate. When the prior information is less reliable, pLASSO shows great robustness to the misspecification. We illustrate the application of pLASSO using a real data set from a genome-wide association study.

Keywords: Asymptotic efficiency; Oracle inequalities; Solution path; Weak oracle property

2.260 Radial-angular decomposition of regularly varying time series in star-shaped metric spaces

Johan Segers[1], Yuwei Zhao[1,*] and Thomas Meinguet[2]

[1]Université catholique de Louvain, Belgium; johan.segers@uclouvain.be, yuwei.zhao@uclouvain.be
[2] ING Bank N.V.; thomas.meinguet@skynet.be

Abstract: There exist two ways of defining regular variation of a time series in a star-shaped metric space: either by the distributions of finite stretches of the series or by viewing the whole series as a single random element in a sequence space. The two definitions are shown to be equivalent. The introduction of a norm-like function, called *radius*, yields a radial-angular decomposition similar to the polar decomposition in Euclidean spaces. The angular component of the time series, called *angular* or *spectral tail process*, captures all aspects of extremal dependence. The stationarity of the underlying series induces a transformation formula of the spectral tail process under time shifts.

Keywords: Regular variation; Spectral tail process; Tail dependence; Time series

References

[1] Hult, H. and Lindskog, F. (2006) Regular variation for measures on metric spaces. *Publ. Inst. Math.* **80**, 121–140.

[2] Meinguet, T. and Segers, J. (2010) Regularly varying time series in Banach spaces. Technical Report, available at *http://arxiv.org/abs/1001.3262*.

[3] Segers, J., Zhao, Y. and Meinguet, T. (2013) Radial-angular decomposition of regularly varying time series in star-shaped metric spaces. *Submitted,* available at *http://arxiv.org/abs/1604.00241*.

2.261 Semiparametric Inference in a Covariate-Adjusted Response-Adaptive RCT with Data-Adaptive Estimation

W. Zheng [1,*], A. Chambaz [2] and M. van der Laan[1]

[1] School of Public Health, University of California, Berkeley, USA; wenjing.zheng@berkeley.edu, laan@berkeley.edu
[2] Université Paris Ouest Nanterre, Paris, France; achambaz@u-paris10.fr

Abstract: In this talk, we present a novel group-sequential Covariate-Adjusted Response-Adaptive (CARA) randomized controlled trial (RCT) design and inferential procedure. The proposed framework adopts a loss-based approach to construct more flexible CARA randomization schemes while exploiting machine learning/data-adaptive estimators for the response model and the design. In general, this approach allows better adaptation towards the user-supplied optimal randomization scheme through better variable adjustments and the targeted construction of an instrumental loss function. Under the proposed framework, the parameter of interest is non-parametrically defined and is estimated using the paradigm of Targeted Maximum Likelihood Estimation (TMLE) based on such an adaptive sampling scheme. We establish that under appropriate empirical process conditions, the sequence of randomization schemes converges to a fixed scheme, and the proposed parameter estimate is robust to model misspecification and is asymptotically normal. We illustrate the proposed framework with the use of LASSO regressions to estimate the conditional response given treatment and baseline covariates.

Keywords: Adaptive Trial Designs; Semiparametric Inference; Targeted Maximum Likelihood Estimation; Response-Adaptive Design; Martingales; Robust Estimation

2.262 A General Framework for Bayes Structured Linear Models

Chao Gao, Aad W. van der Vaart and Harrison H. Zhou*

Yale University, Leiden University and Yale University

Abstract: High dimensional statistics deals with the challenge of extracting structured information from complex model settings. Compared with a large number of frequentist methodologies, there are rather few theoretically optimal Bayes methods for high dimensional models. This paper provides a unified approach to both Bayes high dimensional statistics and Bayes nonparametrics in a general framework of structured linear models. With a proposed two-step prior,

we prove a general oracle inequality for posterior contraction under an abstract setting that allows model misspecification. The general result can be used to derive new results on optimal posterior contraction under many complex model settings including recent works for stochastic block model, graphon estimation and dictionary learning. It can also be used to improve upon posterior contraction results in literature including sparse linear regression and nonparametric aggregation. The key of the success lies in the novel two-step prior distribution: one for model structure, i.e., model selection, and the other one for model parameters. The prior on the parameters of a model is an elliptical Laplace distribution that is capable of modeling signals with large magnitude, and the prior on the model structure involves a factor that compensates the effect of the normalizing constant of the elliptical Laplace distribution, which is important to attain rate-optimal posterior contraction.

Keywords: Oracle inequality, Stochastic block model, Graphon, Sparse linear regression, Aggregation, Dictionary learning, Posterior contraction

2.263 Comparing Large Covariance Matrices under Weak Conditions on the Dependence Structure and its Application to Gene Clustering

J. Chang[1], L. Wang[2], W. Zhou[3,*], and W.-X. Zhou[4]

[1] School of Mathematics and Statistics, The University of Melbourne, Australia
[2] School of Statistics, University of Minnesota, Minneapolis, U.S.A.
[3] Department of Statistics, Colorado State University,U.S.A.
[4] Department of Operations Research and Financial Engineering, Princeton University, U.S.A.

Abstract: Comparing large covariance matrices has important applications in modern genomics, where scientists are often interested in understanding whether relationships (e.g., dependencies or co-regulations) among a large number of genes vary between different biological states. We propose a computationally fast procedure for testing the equality of two large covariance matrices when the dimensions of the covariance matrices are much larger than the sample sizes. A distinguishing feature of the new procedure is that it imposes no structural assumptions on the unknown covariance matrices. Hence the test is robust with respect to various complex dependence structures that frequently arise in genomics. We prove that the proposed procedure is asymptotically valid under weak moment conditions. As an interesting application, we derive a new gene clustering algorithm which shares the same nice property of avoiding restrictive structural assumptions for high-dimensional genomics data. Using an asthma gene expression dataset, we illustrate how the new test helps compare the covariance matrices of the genes across different gene sets/pathways between the disease group and the control group, and how the gene clustering algorithm provides new insights on the way gene clustering patterns differ between the two groups.

Keywords: Bootstrap; Differential expression analysis; Gene clustering; High dimension; Hypothesis testing; Sparsity.

References

[1] Chang, J., Zhou, W., and Zhou, W.-X. (2015). Bootstrap Tests on High Dimensional Covariance Matrices with Applications to Understanding Gene Clustering. ArXiv:1505.04493

2.264 Gradient-based structural change detection for non-stationary time series M-estimation

Weichi Wu[1] and Zhou Zhou[2]

[1] University College London; w.wu@ucl.ac.uk
[2] University of Toronto; zhou@utstat.toronto.edu

Abstract: We consider structural change testing for a wide class of time series M-estimation with non-stationary regressors and errors. New uniform Bahadur representations are established with nearly optimal approximation rates. A CUSUM-type test statistic based on the gradient vectors of the regression is considered. Two of the most frequently used change point testing procedures, pivotalization and independent wild bootstrap, are shown to be inconsistent for non-stationary time series M-estimation. A simple bootstrap method is proposed and is proved to be consistent for M-estimation structural change detection under both abrupt and smooth non-stationarity and temporal dependence. Our bootstrap procedure is shown to have certain asymptotically optimal properties in terms of accuracy and power.

Keywords: M-estimation, quantile regression, piece-wise local stationarity, bootstrap, structural change.

2.265 A scalable nonparametric specification testing in massive data

Y. Zhao[1], C. Zou[1,*] and Z. Wang[1]

[1] Institute of Statistics and LPMC; zyysdjn1988824@163.com, nk.chlzou@gmail.com, zjwang@nankai.edu.cn

Abstract: Lack-of-fit checking for parametric models is essential in reducing misspecification. However, for massive datasets which are increasingly prevalent, classical tests become prohibitively costly in computation and its feasibility is questionable even with modern parallel computing platforms. Building on the divide and conquer strategy, we propose a new nonparametric testing method, that is fast to compute and easy to implement with only one tuning parameter determined by a given time budget. Under mild conditions, we show that the proposed test statistic is asymptotically equivalent to that based on the whole data. Benefiting from using the sample-splitting idea for choosing the smoothing parameter, the proposed test is able to retain the type-I error rate pretty well with asymptotic distributions and achieves adaptive rate-optimal detection properties. Its advantage relative to existing methods is also demonstrated in numerical simulations and a data illustration.

Keywords: Adaptive test; Asymptotic normality; Lack-of-fit test; Rate-optimal; Sample-splitting method

3 Contributed sessions

3.1 Nonparameteric EWMA Sign Control Chart with Estimated Parameters

Abbasi, S. A.

Dept. of Mathematics and Statistics, King Fahd Univ. of Petroleum and Minerals, KSA; saddamaa@kfupm.edu.sa

Abstract: Control charts are widely used for the monitoring of industrial and analytical process parameters. Most control chart structures are based on the assumption that process observations are normally distributed. In real life, many quality characteristics such as capacitance, roundness, insulation resistance etc. follow non-normal distributions. In such situations the control charts based on normality assumption can be misleading for making decisions regarding the state of the process. In recent years, a lot of work has been done for the development of efficient nonparametric control charts for the monitoring of process location and dispersion (cf. [1], [2]). But most of the proposed nonparametric methods are based on known parameter values, which is usually not the case to be. Parameters (location or dispersion) are mostly estimated from Phase I reference sample. Recently [2] emphasized the need of an effective Phase I analysis for nonparametric control charts. Following his recommendation, this study investigates the effect of parameter estimation on the performance of nonparametric EWMA (NPEWMA) sign chart proposed by [3]. The performance of the NPEWMA sign chart is evaluated using average run length (ARL) as a performance measure. The ARL results indicated a significant impact of estimated parameters on the detection ability of nonparametric EWMA charts. The Phase I sample size required, to ensure a consistent NPEWMA sign chart performance (compared to the known parameter case) is also provided as part of this study.

Keywords: Nonparametric; EWMA; Process Monitoring; Average run length; Location.

References

[1] Abbasi, S. A., Miller, A. and Riaz, M. (2013). Nonparametric Progressive Mean Control Chart for Monitoring Process Target. *Quality and Reliability Engineering International*, **29**, 1069–1080.

[2] Capizzi, G. (2015). Recent Advances in Process Monitoring: Nonparametric and Variable-Selection Methods for Phase I and Phase II *Quality Engineering*, **27 (1)**, 44–67.

[3] Graham, M. A. , Chakraborti, S. and Human, S. W. (2011). A Nonparametric EWMA Sign Chart for Location Based on Individual Measurements *Quality Engineering*, **23 (3)**, 227–241.

3.2 Multivariate intensity estimation via wavelet model selection

N. Akakpo[1]

[1] UPMC, LPMA, UMR CNRS 7599; nathalie.akakpo@upmc.fr

Abstract: We propose a new wavelet procedure to estimate a multivariate intensity function. Its originality lies in the use of the hyperbolic wavelet basis of $\mathbb{L}_2([0,1]^d)$, that seldomly appears in statistics. Our procedure selects from the data the adequate set of wavelet coefficients to estimate, by using a penalized least-squares criterion. We impose a structural constraint on the set of coefficients to be selected, which yields both good theoretical performances and a reduced computational complexity. In some sense, we are able to adapt to anisotropy and to overcome the curse of dimensionality. Besides, this approach is valid in a wide framework, encompassing for instance density estimation, copula density estimation, estimating the intensity of a spatial Poisson process or the jump intensity of a Lévy process.

Keywords: Anisotropy, Model selection, Spatial adaptation, Wavelets.

References

[1] Akakpo, N. (2016). New wavelet selection strategies for multivariate intensity estimation. Working paper.

3.3 Separation rates for non-parametric independence tests based on wavelet decomposition and permutation

M. Albert[1]

[1] Gipsa-lab, University Grenoble Alpes, France; melisande.albert@gipsa-lab.grenoble-inp.fr

Abstract: Testing independence is one of the central goals in data analysis, as well as in many application fields, such as synchrony detection in neuroscience for instance. This work presents a non-parametric independence testing procedure in the density framework. First, based on a wavelet thresholding method, single coefficient tests are constructed. Their corresponding critical values are obtained from a permutation approach [2]. In particular, each of these single tests is known to be non-asymptotically of prescribed level. Then, a multiple testing procedure based on aggregation is introduced, avoiding the delicate question of the coefficient choice. A non-asymptotic study of the resulting independence test is performed, providing conditions on the alternative ensuring a control of the second kind error rate by a prescribed value. This study is based on a new Bernstein-type concentration inequality for randomly permuted sums, derived from the fundamental inequalities of [4]. This leads to an upper-bound for the uniform separation rates of the aggregated test with respect to L_2-metric over weak Besov bodies, which should be, in view of the literature [3], optimal in the minimax sense as introduced by [1]. Moreover the whole procedure is adaptive as it is entirely data-driven, and does not require any knowledge on the smoothness of the alternative.

Keywords: Independence tests; Permutation method; Uniform separation rate; Concentration inequalities; Adaptive tests in the minimax sense.

References

[1] Baraud, Y. (2002) Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, **8(5)**, 577–606.

[2] Fisher, R. A. (1935). *The design of experiments*. Edinburgh & London: Oliver & Boyd.

[3] Ingster Yu. I. (1989). Asymptotic minimax testing of independence hypothesis. *Journal of Soviet Mathematics*, **44(4)**, 466–476.

[4] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques.*, **81(1)**, 73–205.

3.4 A comparison of unsupervised curve classification methods for sport training data

G. Lefort[1,2], M. Avalos[1,3,4,*], P. Soret[1,3,4], D. Pyne[5], J.-F. Toussaint[6,7] and P. Hellard[6,8]

[1]INRIA SISTM Bordeaux, France; [2]ENSAI Rennes, France; [3]INSERM U1219, Bordeaux, France; [4]University of Bordeaux, France; [5]Department of Physiology, Australian Institute of Sport, Canberra, Australia; [6]IRMES, INSEP, Paris, France; [7]CIMS, Hôtel-Dieu, AP-HP, Paris, France; [8]Research department, French Swimming Federation, Pantin, France; *Corresponding author: marta.avalos@isped.u-bordeaux2.fr

Abstract: Achieving peak performance at a specified time is the primary goal of athletes' training programs. To optimize performance and reduce the risk of injury,

a comprehensive list of training program parameters (e.g. intensity, volume, frequency, distribution, duration and type) requires careful management. This work focuses on clustering of time evolution curves of training measurements.

Training data are recorded densely over time.

However, duration of follow-up and duration of the seasons vary among subjects. Also, subject-specific variation

can induce substantial error. Functional data analysis (FDA) and longitudinal data analysis (LDA) are the main approaches to analyze repeated measures data (in which multiple measurements are made on the same subject across time). Typically, FDA is applied when the data are dense, assumed to be observed in the continuum, and a function of time. LDA is usually applied when data are sparse, possibly with different number of measurements across individuals, and subject to error. We compared several FDA and LDA methods implemented through publicly available R code: k-means based on the standard Euclidian distance, a discrete Fréchet distance [2], and a functional distance [1]; Gaussian mixture model–based clustering for standard [4], longitudinal [5] and functional [3] data; and latent class mixed models [6]. We discuss advantages and limitations including computational and practical aspects.

Keywords: Functional data analysis; Longitudinal data analysis; Sport science data.

References

[1] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda.usc. *Journal of Statistical Software*, **51**, 1–28.

[2] Genolini, C. and Falissard, B. (2011). Kml : A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*.

[3] Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. Neurocomputing, **112**, 164–171.

[4] Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2014). Rmixmod: The R package of the model–based unsupervised, supervised and semi–supervised classification mixmod library. *Journal of Statistical Software*.

[5] McNicholas, P. D. and Murphy, T. B. (2010). Model–based clustering of longitudinal data. Canadian *Journal of Statistics*, **38**, 153–168.

[6] Proust-Lima, C., Philipps, V., and Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *Technical report, University of Bordeaux*. arXiv:1503.00890v2.

3.5 Identification and estimation in the functional linear instrumental regression

Andrii Babii[1]

[1] Toulouse School of Economics; babii.andrii@gmail.com.

Abstract: This paper studies a particular type of a linear IV regression model with high-dimensional endogenous component, called the functional linear instrumental regression (FLIR). It is shown that identification in this model can be achieved with a single real-valued instrumental variable under the weak completeness condition. Two estimators based on the Tikhonov and Galerkin regularizations are studied. We obtain the non-asymptotic upper bounds on the mean-integrated squared errors and corresponding convergence rates for both estimators. Estimators are simple to implement and demonstrate good small-sample performance in Monte Carlo experiments.

Keywords: Instrumental varibles; Ill-posed inverse problem; Functional data.

References

[1] Jean-Pierre Florens and Sébastien Van Bellegem. (2015). Instrumental variable estimation in functional linear models. *Journal of Econometrics*, **186(2)**, 465–476.

[2] Hervé Cardot and Jan Johannes. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, **101(2)**, 395–408.

[3] Marc Hoffmann and Markus Reiss. (2010). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, **36(1)**, 310–336.

3.6 Inference for Monotone Functions Under Short and Long Range Dependence: New Universal Limits

Pramita Bagchi

University of Michigan

Abstract: Isotonic regression problem is an important problem arising in many applications such as climate studies, economics, current status data in biostatistics, among many others. When the data are independent and Gaussian, the well–known Pooled Adjacent Violators Algorithm provides the maximum likelihood estimate (MLE) of the monotone function we want to estimate. Its asymptotic properties are well understood. In many applications, however, the data are dependent. Motivated by the work of Banerjee & Wellner (2001), our goal is to study the pseudo LRS and L_2 distance between constrained and unconstrained MLE. The limit distribution of these statistics for independent identically distributed data do not involve the slope of the regression function, which is difficult to estimate and arises in the asymptotic distribution of MLE. Here we present some results,which show that these statistics with suitably modified scaling will have the same asymptotic distribution in the case when the data are weakly (short–range) dependent. We have also presented some calculations to obtain the limit distributions of these statistics under strong (long–range) dependence of the data. As these two statistics converge jointly the ratio of these two statistics has a limit distribution free of the the slope of the regression function and other scaling factors under both kinds of dependence. We expect to use this ratio statistic to construct confidence interval for the regression function at a particular point for many practical applications as it does not involve independence assumption on the data and estimating the scaling factors.

3.7 Maximum likelihood estimation of a unimodal probability mass function

F. Balabdaoui[1,2,*] and H. J. Jankowski[3]

[1] CEREMADE, Université Paris-Dauphine; fadoua@ceremade.dauphine.fr
[2] Seminar für Statistik, E.T.H Zürich
[3] York University, hkj@mathstat.yorku.ca

Abstract: We develop an estimation procedure for a discrete probability mass function (pmf) with unknown support. We derive its maximum likelihood estimator under the mild and natural shape-constraint of unimodality. Shape-constrained estimation is a powerful and robust technique that additionally provides smoothing of the empirical distribution yielding gains in efficiency. We show that our unimodal estimator is consistent when the model is specified, and that it converges to the best projection of the true pmf on the unimodal class under model misspecification. We derive the limiting distribution of the the estimator, and use this to build asymptotic confidence bands for the unknown pmf when the latter is unimodal. We illustrate our approach using time-to-onset data of the Ebola virus during the 1976 outbreak in the former republic of Zaire.

Keywords: Maximum likelihood estimation; Probability mass function estimation; Shape constrained estimation; Unimodal.

3.8 Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data

I. Barbeito[1,*] and R. Cao[1]

[1] Research group MODES, Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain; ines.barbeito@udc.es, rcao@udc.es

Abstract: In this work a smoothed version of the stationary bootstrap introduced by [1] is established for the purpose of bandwidth selection in density estimation for dependent data. An exact expression for the bootstrap version of the MISE under dependence is obtained in this context. This is very useful since implementation of the bootstrap selector does not require Monte Carlo approximation. A simulation study is carried out to show the good practical performance of the new bootstrap bandwidth selector with respect to other existing competitors. The method is illustrated by applying it to two real data sets.

Keywords: Kernel method; Mean integrated squared error; Smoothing parameter; Stationary processes.
2010 MSC: 62F40, 62G07, 60G10

References

[1] Politis, D.N. and Romano, J.R. (1994). The stationary bootstrap. J. Amer. Statist. Assoc., 89, 1303-1313.

3.9 On testing for Hermite rank in Gaussian subordination series

J. Beran[1,*]

[1,*] University of Konstanz, Germany; jan.beran@uni-konstanz.de

Abstract: Part of this talk is joint work with Sven Moehrle and Sucharita Ghosh. Statistical methods for non-Gaussian long-memory processes are often based on the assumption of Gaussian subordination with Hermite rank one. The main reasons for this assumption are mathematical convenience and the absence of methods for checking the assumption empirically. In this talk, an asymptotically consistent computational method for testing the null hypothesis of Hermite rank one is introduced. Simulations and data examples illustrate the finite sample performance. General implications for statistical inference are discussed.

Keywords: Long-range dependence; Long memory; Gaussian subordination; Hermite rank; Bootstrap

3.10 New proposals for multidirectional supervised classification

J.R. Berrendero[1,*], J. Cárcamo[1] and J.G. Ponce[2]

[1] Department of Mathematics, Universidad Autónoma de Madrid; joser.berrendero@uam.es, javier.carcamo@uam.es
[2] Universidad Nacional de Ingeniería, Lima, Perú; jahazielponce@gmail.com

Abstract: Multidirectional discriminant analysis is a generalization of linear discriminant analysis in which two or more discriminant hyperplanes are considered instead of just one. The goal of multidirectional classification rules is to combine the interpretability of linear rules with the flexibility of more involved nonlinear classifiers. See [1] for a discussion of the advantages of these methods and a proposal based on a generalization of support vector machines. Here, we introduce several new proposals to obtain multidirectional rules and address their potential use in variable selection and dimension reduction. The performance and usefulness of the new classification rules are illustrated through simulations and real data examples.

Keywords: Classification; Discriminant analysis; Variable selection; Dimension reduction.

References

[1] Huang, H., Liu, Y. and J.S. Marron (2012). Bidirectional discrimination with application to data visualization. *Biometrika*, **99**, 851–864.

3.11 Analysing Rankings with the Sign Test, Using p-values Conditional on the Rank Order of the Sample

L. Bohlin[1]∗

[1] Affiliation; lars.bohlin@mdh.se

Abstract: This paper deals with the problem of making inference from a survey question where the respondents are asked to rank a couple of objects from the best one to the worst. Pair wise sign tests could be used to investigate what object/objects that differs from the other but there is a problem to choose the combinations to be tested. The standard recommendation in statistics is to make these choices before the data is collected to avoid that the choice of method will depend upon the data in a specific sample. We show that the rejection frequencies under a true H_0 differ quite a lot from the level of significance if the choice of objects is dependent on the rank order of the sample and conventional p-values are used.
A method to calculate the p-values of the sign test conditional on the mean ranks in the specific sample is developed. The advantage with this methodology is, that the choice of objects to be compared may be done after the descriptive statistics are calculated and still yield consistent p-values. Since it is fairly difficult for the referees to control how authors choose what tests to make, results based upon conditional p-values would be much more trustworthy than results based on conventional p-values. The method of conditional p-values is evaluated in comparison with a strategy where all pairs of objects are tested and the p-values are adjusted with the Holm-Bonferoni method to avoid a too high family wise error rate.

Keywords: Conditional p-values; Non-Parametric Methods; Sign Test; Data Mining.

3.12 Bootstrapping kernel density estimators for length-biased data

Borrajo, M.I.[1,*], González-Manteiga, W.[1] and Martínez-Miranda M.D.[2]

[1] Faculty of Mathematics, University of Santiago de Compostela, E15782 Santiago de Compostela, Spain;
mariaisabel.borrajo@usc.es, wenceslao.gonzalez@usc.es
[2] Faculty of Sciences, University of Granada, E18071 Granada, Spain; mmiranda@ugr.es

Abstract: Length-biased data arise in sampling proccesses in many different fields such as ecology, epidemiology or industry. The main problem is that the observed data are not representative of the original variable of interest, indeed the probability of data being sampled is proportional to its length. Even though there is many literature on the topic of density estimation in general, it was not so prolific in the present context. [1] proposed a kernel density estimator for length-biased data, and later on [2] define an adapted cross-validation method for bandwidth selection.
In this work we provide asymptotic derivations about the mean and the variance of Jones' density estimator. We define a consistent bootstrap method for length-biased data, we study its asymptotic properties and we apply it to build new bandwidth selectors. Finally we study the finite sample performance of the different proposed methods through an extensive simulation study.

Keywords: Length-biased data; Density; Bandwidth; Bootstrap

References

[1] Jones, M.C. (1991). Kernel density estimation for length-biased data. *Biometrika*, **3**(78), 511–519.

[2] Guillamón, A., Navarro J. and Ruiz, J.M. (2002). Kernel density estimation using weighted data. *Communications in Statistics-Theory and Methods*, **9**(27), 2123–2135.

3.13 Strong approximations for a class of integrated empirical processes with applications to statistical tests

S. Alvarez-Andrade[1], S. Bouzebda[1,*] and Aimé Lachal[2]

[1] Sorbonne Universités, Université de Technologie de Compiègne
Laboratoire de Mathématiques Appliquées de Compiègne salim.bouzebda@utc.fr
[2] Université de Lyon, Institut National des Sciences Appliquées de Lyon

Abstract: The main purpose of this paper is to investigate the strong approximation of a class of integrated empirical processes. More precisely, we obtain the exact rate of the approximations by a sequence of weighted Brow- nian bridges and a weighted Kiefer process. Our arguments are based in part on the Komlós et al. (1975)'s results. Applications include the two-sample testing procedures together with the change-point problems.

Keywords: Integrated empirical process; Brownian bridge; Kiefer process; Rates of convergence.

References

[1] Alvarez-Andrade, S., Bouzebda, S. and Lachal, A. (2016a). Some asymptotic results for the integrated empirical process with applications to statistical tests. *Comm. Statist. Theory Methods*, to appear.

[2] Alvarez-Andrade, S., Bouzebda, S. and Lachal, A. (2016b). Strong approximations for a class of integrated empirical processes with applications to statistical tests. Preprint

3.14 A note on the CLT for the discrete Fourier transforms of functional time series

C. Cerovecki[1,*] and S. Hörmann[1]

[1] Department of Mathematics, Université Libre de Bruxelles, Belgium. clement.cerovecki@ulb.ac.be

Abstract: Functional data often arise by segmenting a continuous time process into natural units, such as days. Then a certain degree of dependence between the observed curves $X_1, \ldots, X_T$ is well expected and, consequently, a thorough statistical investigation requires time series methodology. During recent years functional time series (FTS) analysis has seen an upsurge in the scientific community and diverse related practical and theoretical problems have been addressed. Some of the latest publications are devoted to frequency domain topics for FTS such as [2] or [1]. Motivated by this fact, we study the key ingredient for the frequency domain approach: the discrete Fourier transform of the function valued random process. We derive its weak convergence to a complex Gaussian (functional) random element under very mild assumptions.

Keywords: Central limit theorem; Functional time series; Fourier transform; Periodogram; Stationarity.

References

[1] Hörmann, S., Kidziński, L. and Hallin, M. (2014). *Dynamic Functional Principal Component.* Journal of the Royal Statistical Society: Series B, 77, 319–348.

[2] Panaretos, V.M. and Tavakoli, S. (2013). *Fourier analysis of stationonary time series in function spaces.* The Annals of Statistics, 41, No. 2, 568–603.

3.15 On Robbins-Monro type conditional variance estimation with functional ergodic data

Mohamed Chaouch

Department of Statistics, United Arab Emirates University; m.chaouch@uaeu.ac.ae

Abstract: In this paper, we are interested in nonparametric estimation of the conditional variance when the predictor takes values in an infinite dimensional space and the response variable is scalar. In the statistical literature, a lot of attention has been given to the nonparametric estimation of the regression function when the covariate is of functional nature. Many authors studied its asymptotic properties such as the almost complete consistency with rate and the asymptotic distribution when the underlying process satisfies an α-mixing condition (see [1] and the references therein). Recently [3] and [2] generalized those results to ergodic processes. However, almost nothing has been done for the conditional variance estimation with functional stationary ergodic processes. Firstly, a kernel-type estimator of the conditional variance function is defined, then a uniform almost sure consistency rate as well as the asymptotic distribution are established. Nowadays, with the progress of measurement apparatus and the development of automatic sensors, we can get access to large samples of observations taking values in high dimensional spaces. Therefore, within this new framework of "Massive Data", the computation of the nonparametric estimator of the conditional variance presented

in first part will be a challenge. To deal with this constraint, recursive algorithm will be introduced to perform the conditional variance estimation without any full data storage requirement. More precisely, when the data arrive sequently the value of each successive estimator is obtained from its value at the previous step by a simple adjustment that takes into account the recently received data. A mean square consistency of the Robbins-Monro type conditional variance estimator is then established. Finally, a simulation study will be given to show how the recursive estimator performs better than the static one in term of computation time without affecting significantly the accuracy.

Keywords: Conditional variance; functional ergodic data; massive data; nonparametric estimation; Robbins-Monro approximation

References

[1] Ferraty, F. and Vieu, P. (2006). *Nonparametric Modelling for Functional Data. Methods, Theory, Applications and Implementations.* Springer-Verlag, London.

[2] Laib, N. and Louani, D. (2011). Rates of strong consistencies of the regression function estimator for functional stationary ergodic data. *J. Statist. Plann. Inference*, **141**(1), 359–372.

[3] Laib, N. and Louani, D. (2010). Nonparametric kernel regression estimation for functional stationary ergodic data: asymptotic properties. *J. Multivariate Anal.*, **101**(10), 2266–2281.

3.16 Asymptotic AIC post-selection confidence intervals

Ali Charkhi* and Gerda Claeskens

ORStat and Leuven Statistics Research Center, KU Leuven, Belgium.

Abstract: Given a set of models, one selects a model to work with. This selection procedure usually is a data driven method, hence it affects the distribution of the parameter estimators in the selected model and naive inference fails to cover these effects. In this paper we consider the Akaike information criterion for model selection and study the asymptotic distribution of parameter estimators after selection. The method is applicable to any likelihood model, including though not restricted to generalized linear models. It turns out that the parameter estimator's distribution depends on the competitive models and the smallest true model. Regarding the fact that the smallest true model is unknown, we can calculate conservative confidence intervals by assuming that the smallest model in the model set is true. Moreover, if the models are misspecified, we can still calculate the confidence intervals for the pseudo-true parameters. Also, the distribution of a linear combination of the parameter estimators in the selected model has been calculated. In order to be able to calculate the confidence intervals for the parameters, we propose a simulation method from the asymptotic distribution. This method is based on considering the constraints that are imposed by the model selection criterion. Simulations for both linear models and generalized linear models show the validity of the method in finite samples.

Keywords: Model selection; Post-selection inference; Akaike information criterion; Likelihood model; Confidence region.

3.17 Risk Bounds for the Least Squares Estimator in Unimodal Regression

S. Chatterjee[1*] and J. Lafferty[2]

[1*] University of Chicago; sabyasachi.chatterjee@uchicago.edu
[2] University of Chicago; lafferty@galton.uchicago.edu

Abstract: There has been a recent surge in interest in estimation of functions with shape constraints. The main phenomenon of interest, which all of these papers concern with, is an automatic adaptive property of the constrained Least Squares Estimator(LSE). For example in isotonic regression, [1] showed that the LSE, under the mean squared error loss, achieves a near parametric rate of convergence $s\frac{\log n}{n}$ whenever the true isotonic function is piecewise constant with s pieces. This rate of convergence is in contrast with the well known and usual $n^{-2/3}$ rate of convergence of the LSE. We stress that this adaptation to piecewise constant isotonic functions is completely automatic. Similar phenomenon has been shown to hold for the LSE in problems such as convex regression(see [3]) and isotonic matrices (see [4]. In fact, it was shown in [2] that the LSE has this automatic adaptive property whenever the shape constraints mean that

the parameter space is a convex cone of an appropriate type. An important shape constraint that does not imply that the parameter space is a convex cone is unimodality. Infact, after an appropriate formulation of the unimodal regression problem, the parameter space is not even convex. Although closely related to isotonic regression, unimodal regression has not been as extensively studied and nothing is known about the behaviour of the LSE as an estimator of the true unimodal function. In this paper we study the statistical properties of the least squares estimator in unimodal function estimation and examine whether the automatic adaptivity property of the LSE still holds in this setting. Our results show that the risk of the LSE, in general, in mean squared error, is infact still $O(n^{-2/3})$. We are also able to show near parametric rates of convergence $O(s^{3/2}(\log n)^{3/2}/n)$ for unimodal functions which are piecewise constant with s pieces. Our technique of analyzing the mean squared error of the LSE includes a general variational representation of the risk that holds whenever the parameter space can be expressed as a finite union of convex sets, and may be of interest in other settings. The reader may note that the exponent of s and the log term is $3/2$ intead of 1 which one obtains in isotonic regression. This raises a natural question whether the true exponent is still 1. We do not know the answer to this question. Another open question is whether one can obtain oracle type risk bounds as has been obtained in isotonic regression by [1] and [2]. These oracle risk bounds imply near parametric rates of convergence of the LSE even if the true function is very close to a piecewise isotonic function with a few pieces. Existing proofs of oracle risk bounds seem to rely heavily on convexity of the parameter space and hence cannot be directly applied to our setting.

Keywords: Unimodal; Least Squares Estimate; Automatic Adaptation.

References

[1] Chatterjee, S., Guntuboyina, A., and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. The Annals of Statistics, 43(4), 1774-1800.

[2] Bellec, P. C. (2015). Sharp oracle inequalities for Least Squares estimators in shape restricted regression. arXiv preprint arXiv:1510.08029.

[3] Guntuboyina, A., Sen, B. (2013). Global risk bounds and adaptation in univariate convex regression. Probability Theory and Related Fields, 1-33.

[4] Chatterjee, S., Guntuboyina, A., Sen, B. (2015). On matrix estimation under monotonicity constraints. arXiv preprint arXiv:1506.03430.

3.18 Estimating the Conditional Error Distribution in Non-parametric Regression for Functional Data and Applications

M. Cherfi

Faculté des Sciences, Université Hassiba Benbouali de Chlef, Algérie; mohamed.cherfi@mail.com

Abstract: In this paper, a new nonparametric estimation procedure of the conditional error distribution in regression of a scalar response variable given a random variable taking values in a semi-metric space. The unknown error distribution is estimated by a kernel estimator of residuals, where the functional Nadaraya–Watson estimator is used to estimate the regression function.

Under some general conditions, we establish both the pointwise and the uniform almost-complete consistencies with convergence rates of the conditional distribution estimator. Procedures are constructed for testing independence between the regressor and the error. The tests are based on The test statistic is based on a kernel estimator for the L_2-distance between the conditional distribution and the unconditional distribution of the error. Simulation results and a real data example are presented.

Keywords: Conditional density; Nonparametric regression; Functional data; Test for independence.

References

[1] Ferraty F, Vieu P (2006). *Nonparametric functional data analysis: theory and practice*. Springer, New York.

[2] Kiwitt, S. and Neumeyer, N. (2012). Estimating the conditional error distribution in non-parametric regression. *Scand. J. Stat.*, **39**(2), 259–281.

[3] Laksaci, A., Madani, F., and Rachdi, M. (2013). Kernel conditional density estimation when the regressor is valued in a semi-metric space. *Comm. Statist. Theory Methods*, **42**(19), 3544–3570.

[4] Neumeyer, N. (2009). Testing independence in nonparametric regression. *J. Multivariate Anal.*, **100**(7), 1551–1566.

[5] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.*, **101**(5), 1067–1078.

[6] Shang, H. (2013). Bayesian bandwidth estimation for a nonparametric functional regression model with unknown error density. *Computational Statistics and Data Analysis*, **67**, 185–198.

3.19 Fast Two-Sample Testing with Analytic Representations of Probability Measures

Kacper Chwialkowski[1], Aaditya Ramdas[2], Dino Sejdinovic[3], Arthur Gretton[1]

[1] Gatsby Computational Neuroscience Unit, UCL; kacper.chwialkowski@gmail.com, arthur.gretton@gmail.com
[2] Department of EECS and Statistics, UC Berkeley; aramdas@cs.berkeley.edu
[3] Department of Statistics, University of Oxford,dino.sejdinovic@gmail.com

Abstract: We propose a class of nonparametric two-sample tests with a cost linear in the sample size. Two tests are given, both based on an ensemble of distances between analytic functions representing each of the distributions. The first test uses smoothed empirical characteristic functions to represent the distributions, the second uses distribution embeddings in a reproducing kernel Hilbert space. Analyticity implies that differences in the distributions may be detected almost surely at a finite number of randomly chosen locations/frequencies. The new tests are consistent against a larger class of alternatives than the previous linear-time tests based on the (non-smoothed) empirical characteristic functions [**? ? ?**], while being much faster than the current state-of-the-art quadratic-time kernel-based [**?**] or energy distance-based tests [**? ? ?**]. Experiments on artificial benchmarks and on challenging real-world testing problems demonstrate that our tests give a better power/time tradeoff than competing approaches, and in some cases, better outright power than even the most expensive quadratic-time tests. This performance advantage is retained even in high dimensions, and in cases where the difference in distributions is not observable with low order statistics.

Keywords: Kernel Methods, Statistical Hypothesis Testing

3.20 Goodness-of-fit tests in semiparametric transformation models using the integrated regression function

B. Colling[1,*], I. Van Keilegom[1]

[1] Institut de statistique, Université catholique de Louvain, Belgium; benjamin.colling@uclouvain.be, ingrid.vankeilegom@uclouvain.be

Abstract: Consider the following semiparametric transformation model $\Lambda_\theta(Y) = m(X)+\varepsilon$, where X is a d-dimensional covariate, Y is a univariate dependent variable and ε is an error term with zero mean and which is independent of X. We assume that m is an unknown regression function and that $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions. We use a profile likelihood estimator for the parameter θ and a semiparametric local polynomial estimator for m. Our goal is to develop a new test for the parametric form of the regression function m and to compare its performance to that proposed by Colling and Van Keilegom (2015). The basic idea of the test developed by Colling and Van Keilegom (2015) was to compare the distribution function of ε estimated in a semiparametric way to the distribution function of ε estimated under the null hypothesis whereas here we compare the integrated regression function estimated in a semiparametric way to the integrated regression function estimated under the null hypothesis. We consider two different test statistics, a Kolmogorov-Smirnov and a Cramér-von Mises type statistic. We establish the limiting distributions of these two test statistics under the null hypothesis and under a local alternative. Finally, a simulation study is carried out to illustrate the performance of our testing procedure, to compare this new test to the previous one and to see under which model conditions which test behaves the best.

Keywords: Goodness-of-fit; Integrated regression function; Profile likelihood; Semiparametric regression; Transformation model.

References

[1] Bierens, H.J. (1982). Consistent model specification tests. *Journal of Econometrics*, **20**, 105-134.

[2] Colling, B. and Van Keilegom, I. (2015). Goodness-of-fit tests in semiparametric transformation models. *TEST* (in press)

[3] Escanciano, J.C. (2006). A consistent test for regression models using projections. *Econometric Theory*, **22**, 1030-1051.

[4] Linton, O., Sperlich, S., Van Keilegom, I. (2008). Estimation of a Semiparametric Transformation Model. *Annals of Statistics* **36**, 686–718.

[5] Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* **25**, 613–641.

3.21 Second Order Correctness of Perturbation Bootstrap M-Estimator of Multiple Linear Regression Parameter

Debraj Das[1], Soumendra Nath Lahiri[2]

[1] North Carolina State University; ddas3@ncsu.edu
[2] North Carolina State University; snlahiri@ncsu.edu

Abstract: Consider the multiple linear regression model where errors are independent and identically distributed and design vectors are non-random completely known. The idea of random perturbation of the objective function in a semiparametric setting was introduced by Jin, Ying and Wei (2001). In this work, this idea of random perturbation is adapted for perturbing the score function of regression M-estimator to define a new bootstrap method, named "Perturbation Bootstrap" and also second order properties have been shown. It has been found that although perturbation bootstrap method is second order correct in the standardized setup, the standard way of studentization of the bootstrapped estimator fails to be second order correct; unlike the classical residual bootstrap. An innovative way of studentization in perturbation bootstrap setting has been introduced which corrects the distribution of the studentized M-estimator upto second order.

Keywords: M-Estimation; Perturbation Bootstrap; Residual Bootstrap; Studentization; Edgeworth Expansion.

3.22 Bayesian Inferences of Welfare Reform with Improved Credible Interval Estimation via Semiparametric Out of Sample Fusion

K. Dayaratna[1]

[1] Senior Statistician and Research Programmer, The Heritage Foundation; kevin.dayaratna@heritage.org

Abstract: First popularized by former U.S. President Ronald Reagan in his classic 1964 "A Time for Choosing" speech, the concept of welfare reform has been a hot topic of public policy research for decades. The Personal Responsibility and Work Authorization Act of 1996 was one of the United States' most comprehensive efforts at welfare reform. The law's aim was to transform one of America's major welfare programs away from a system fostering dependency and into a program providing temporary assistance to enable people to become contributing members of society. In this study, we utilize non-parametric Bayesian methods to quantify the impact of this law. In the process, we improve upon existing Bayesian interval estimation methods by calling upon semi-parametric estimation techniques thus far used only in frequentist statistical modeling. We find that the welfare reform of the 1990s was quite successful in getting people back to work and can be improved upon even further. We conclude by discussing the resulting policy implications and conclude with avenues of future research.

Keywords: Approximate Bayesian Computation, Non-Parametric Modeling, Density Ratio Estimation

References

[1] Dayaratna, K.D (2014). Contributions to Bayesian Statistical Modeling in Public Policy Research. *Ph.D. Dissertation.*

3.23 Copula Quantile Regression with Censored Data

M. De Backer[1,*], A. El Ghouch[1] and I. Van Keilegom[1]

[1] Université Catholique de Louvain; mickael.debacker@uclouvain.ac.be, anouar.elghouch@uclouvain.be, ingrid.vankeilegom@uclouvain.be

Abstract: When facing multivariate covariates, general semiparametric regression techniques come at hand to propose flexible models that are unexposed to the curse of dimensionality. In this work, in the context of (possibly) right-censored response observations, a copula-based estimator for conditional quantiles is investigated that would allow practitioners to analyse in a flexible way multivariate data. In spirit, our methodology is an extension of the recent work of [1] and [2] to allow for the presence of censoring, as the main idea consists of expressing the characterization of the quantile regression in terms of a multivariate copula and marginal distributions, hereby making use of the advantages of copulas in dependence modelling. However, in order to bypass the effects of a possible misspecification of the underlying copula, we further propose in this work an alternative semiparametric estimation scheme for the multivariate copula density, driven by the regression context. The resulting estimator has the valuable property of being automatically monotonic across quantile levels, and asymptotic normality is obtained under classical regularity conditions. Finally, numerical examples as well as a real data application are used to illustrate the validity and finite sample performance of the proposed procedure in comparison with competing methodologies.

Keywords: Censored quantile regression; Multidimensional copula modeling; Semiparametric regression; Multivariate survival analysis.

References

[1] Noh, H., El Ghouch, A. and Bouezmarni, T. (2013). Copula-Based Regression Estimation and Inference. *Journal of the American Statistical Association*, **108**, 676–688.

[2] Noh, H., El Ghouch, A. and Van Keilegom, I. (2015). Semiparametric Conditional Quantile Estimation through Copula-Based Multivariate Models. *Journal of Business and Economic Statistics*, **33(2)**, 167–178.

3.24 Probabilistic Index Mixed Models for Clustered Data

J. De Neve[1,*], S. Vansteelandt[2] and O. Thas[3]

[1] Department of Data Analysis, Ghent University, Ghent, Belgium; Jan.DeNeve@UGent.be
[2] Dpt Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium; Stijn.Vansteelandt@UGent.be
[3] Department of Mathematical Modelling, Statistics and Bio-Informatics, Ghent University, Ghent, Belgium and National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Australia; Olivier.Thas@UGent.be

Abstract: The linear mixed model plays an important role in the analysis of clustered data. In this presentation we propose a class of rank-based semiparametric models for this type of data. The models make use of random effects, but the effect size is measured in terms of the probabilistic index, which is also the effect size parameter on which the Wilcoxon–Mann–Whitney test is based. Our models extend the probabilistic index model [1**?**]. We provide consistent estimators of the model parameters and develop an asymptotic distribution theory, which is validated in a simulation study. The proposed method is also applied in a data example.

Keywords: Mixed models; Probabilistic index model; Semiparametric inference, Wilcoxon–Mann–Whitney test.

References

[1] Thas, O., De Neve, J., Clement, L. and Ottoy, JP. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society: Series B*, **74**, 623–671.

[2] De Neve, J. and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, **110**, 1276–1283.

3.25 Asymptotic theory for frequentist multiple imputation for Cox regression with missing covariate data

F. Eriksson, T. Martinussen

University of Copenhagen, Department of Public Health, Section of Biostatistics, eriksson@sund.ku.dk and tma@sund.ku.dk

Abstract: Incomplete information on explanatory variables is commonly encountered in both observational and randomized studies. A popular approach to deal with partially observed covariates is Rubin's multiple imputation procedure, where a number of completed data sets that can be analyzed by standard complete data methods are obtained by imputing missing values from an appropriate distribution. We describe frequentist (also called improper) multiple imputation for missing covariate data where the outcome of interest is a possibly censored event time. We fill missing values conditional on the observed data from a semi-parametric imputation model in such a way that model congeniality with a Cox regression analysis model is preserved. Inference for the finite-dimensional regression parameter and the infinte-dimensional cumulative baseline hazard parameter can then be based on Cox regression and Breslow's estimator applied to the completed data sets. We show that the proposed estimators are consistent and derive the asymptotic distribution of both the parametric and non-parametric components. We propose two estimators of the asymptotic variance, one that is consistent and another one based on Rubin's estimator that is asymptotically unbiased. Moderate sample size performance of the estimators is investigated via simulation and by application to a real data example.

Keywords: Cox model; Asymptotics; Missing covariate data; Multiple imputation; Survival analysis

3.26 Permuting longitudinal data despite all the dependencies

S. Friedrich[1,*], E. Brunner[2] and M. Pauly[1]

[1] Ulm University, Institute of Statistics; sarah.friedrich@uni-ulm.de, markus.pauly@uni-ulm.de
[2] University Medical Center Göttingen, Institute of Medical Statistics; ebrunne1@gwdg.de

Abstract: In many experiments in the life, social or psychological sciences the experimental units are observed at different occasions, e.g. different time points. This leads to certain dependencies between observations from the same unit and results in a more complicated statistical analysis. Classical repeated measures models assume that the observation vectors are independent with normally distributed error terms and a common covariance matrix for all groups. However, these two assumptions are often not met in practice and may inflate the type-I error rates of the corresponding procedures.

We present a different approach working under covariance heterogeneity and without postulating any specific underlying distribution. For such general models only a few inference procedures are known which typically do not possess good finite sample properties [2]. For example, the asymptotic pivotal Wald-type test statistic (WTS) is known to be very liberal for small sample sizes and repeated measures. We improve its small sample behavior under the null hypothesis by a studentized permutation technique generalizing the results of [4].

The methodology is motivated by a factorial data example on EEG measurements in patients with Alzheimer's disease [1]. In this data set we have small sample sizes, unequal covariance matrices and repeated measurements.

The permutation procedure leads to astonishingly successful results despite all the dependencies in the repeated measures design which is shown in extensive simulation studies. Moreover, the theoretical properties of the method are analyzed and it is applied to the EEG data set.

Keywords: Permutation Tests; Repeated Measures; EEG Data; Quadratic Forms; Longitudinal Data.

References

[1] Bathke, A, Friedrich, S, Konietschke, F, Pauly, M, Staffen, W, Strobl, N and Y. Höller (2015). Using EEG, SPECT, and Multivariate Resampling Methods to Differentiate Between Alzheimer's and other Cognitive Impairments. *Submitted Preprint.*

[2] Brunner, E. (2001). Asymptotic and Approximate Analysis of Repeated Measures Designs under Heteroscedasticity. In: *Mathematical Statistics with Applications to Biometry*, Eds: J. Kunert and G. Trenkler, Josef Eul Verlag, Köln.

[3] Friedrich, S., Brunner, E. and Pauly, M. (2015). Permuting longitudinal data despite all the dependencies. arXiv:1509.05570.

[4] Pauly, M., Brunner, E. and Konietschke, F. (2015). Asymptotic Permutation Tests in General Factorial Designs. *Journal of the Royal Statistical Society - Series B*, **77**, 461–473.

3.27 Wavelet Whittle estimation in multivariate time series models

S. Achard[1] and I. Gannaz[2,*]

[1] CNRS, Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France; sophie.achard@gipsa-lab.grenoble-inp.fr
[2] Université de Lyon, CNRS UMR 5208, INSA de Lyon, Institut Camille Jordan, France; irene.gannaz@insa-lyon.fr

Abstract: Many applications such as finance, geophysics or neuroscience present multivariate time series data. Correlation between time series is an important feature. Yet differences in autocorrelations properties of the processes can induce phase-shifts in estimation. We consider a semiparametric model for multivariate long-range dependent time series. The coupling between time series is characterized by the long-run covariance matrix. The proposed multivariate wavelet-based Whittle estimation is shown to be consistent for the estimation of both the long-range dependence and the covariance matrix. A simulation study illustrates the finite sample behaviour of the estimation. Finally we propose an application to the estimation of a human brain functional network based on MEG data sets. Our study highlights the benefit of the multivariate analysis, namely improved efficiency of estimation of dependence parameters and of long term correlations.

Keywords: Multivariate processes; Long-range dependence; Long-run covariance matrix; Semiparametric estimation; Wavelets.

References

[1] Achard, S. and Gannaz, I. (2015) Multivariate Wavelet Whittle Estimation in Long-range Dependence. *Journal of Time Series Analysis*, doi: 10.1111/jtsa.12170.

3.28 Estimation of the density of random coefficients when the regressors are bounded

C. Gaillac[1] and E. Gautier[2,*]

[1] CREST, 15 Boulevard Gabriel Péri, 92245 Malakoff, France; christophe.gaillac@ensae.fr
[2] Toulouse School of Economics, University of Toulouse Capitole, 21 allée de Brienne, 31000 Toulouse; eric.gautier@tse-fr.eu

Abstract: We consider the estimation of the density of random coefficients in a linear random coefficients model where the random intercept and slopes are independent from p continuously distributed regressors. We address the case where the support of the regressors is a strict subset of $\mathbb{R}^p$. We provide new identification conditions which impose no restriction on the random intercept. These conditions involve the observables, namely the conditional distribution of the outcome given the regressors. We present lower bounds on the minimax risk based on the mean integrated squared error and an adaptive estimator. Rates of convergence range from logarithmic in the sample size to parametric up to a log factor depending on the smoothness class. Finally, we estimate the heterogeneity in price and income elasticities of British households in a linear demand model with random coefficients.

Keywords: Inverse problem; Random coefficients.

3.29 Estimation of a Partially Linear Regression in Triangular Systems

X. Geng[1,*], C. Martins-Filho[1,2] and F. Yao[3,4]

[1] IFPRI, 2033 K Street NW, Washington, USA; x.geng@cgiar.org, c.martins-filho@cgiar.org
[2] Department of Economics, University of Colorado, Boulder, USA; carlos.martins@colorado.edu
[3] Department of Economics, West Virginia University, Morgantown, USA; feng.yao@mail.wvu.edu
[4] School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, China; 201470006@oamail.gdufs.edu.cn

Abstract: We propose a kernel-based estimator for a partially linear regression in a triangular system where endogenous regressors appear both in the nonparametric and linear components of the regression. Compared with alternative estimators currently available in the literature (**?** ; **?**), our estimator has an explicit functional form, is easier to implement, and exhibits better experimental finite sample performance. The estimator is inspired by the control function approach of [3] and was initially proposed by [4]. It explores conditional moment restrictions that make it suitable for additive regression estimation as in [5] and [6]. We establish consistency and $\sqrt{n}$ asymptotic normality of the estimator for the parameters in the linear component of the model and give a uniform convergence rate for the estimator of the

nonparametric component. In addition, for statistical inference, a consistent estimator for the covariance of the limiting distribution of the parametric estimator is provided. We illustrate the empirical viability of our estimation procedure by applying it to the study of the impact of foreign aid and policy on growth of per capita gross domestic product (GDP) in developing countries.

Keywords: Partially linear regression; Endogeneity; Semiparametric instrumental variable estimation.

References

[1] Ai, Chunrong and Chen, Xiaohong (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**, 1795–1843.

[2] Otsu, Taisuke (2011). Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econometric Theory*, **36**, 8-46.

[3] Newey, Whitney K and Powell, James L and Vella, Francis (1999). Nonparametric estimation of triangular simultaneous equation models. *Econometrica*, **67**, 565–603.

[4] Martins-Filho, Carlos and Yao, Feng (2012). Kernel-based estimation of semiparametric regression in triangular systems. *Economics Letters*, **115**, 24–27.

[5] Kim, Woocheol and Linton, Oliver B and Hengartner, Niklaus (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**, 278–297.

[6] Manzan, Sebastiano and Zerom, Dawit (2005). Kernel estimation of a partially linear additive model. *Statistics and Probability Letters*, **72**, 313-322.

3.30 Influence functions for prediction performance in censored data

Thomas A. Gerds and Rikke N. Mortensen

Department of Biostatistics, University of Copenhagen
Department of Health Science and Technology, Aalborg University Hospital

Abstract: To validate a risk prediction model means to assess the performance when the model is applied to new (yet unseen) subjects. Traditional measures of prediction performance are the Brier score and the area under the ROC curve (AUC). However, for this task data are needed. Ideally, these data would be entirely new data collected for the purpose of validating the model. This is called external validation. In practice external validation data will often not be available. Thus, the best one can do to obtain validation data is to split the available data, such that one part of the data is hidden in the model building process and then used to calculate prediction performance. Averaging the prediction performance across repeated splits of the data is called cross-validation. In this talk, I will derive the influence function of the cross-validation estimate of prediction performance in the case where the outcome is the event status at a future time-point and the data can be right censored.

3.31 Minimax wavelet estimation for multisample heteroscedastic nonparametric regression.

M. Giacofci[1,*], S. Lambert-Lacroix[2] and F. Picard[3]

[1] IRMAR, UMR 6625, Université Rennes 2, F-35043, Rennes, France; joyce.giacofci@univ-rennes2.fr
[2] UJF-Grenoble 1/CNRS/UPMF/TIMC-IMAG UMR 5525, Grenoble, F-38041, France; sophie.lambert@imag.fr
[3] LBBE, UMR CNRS 5558 Université Lyon 1, F-69622, Villeurbanne, France; franck.picard@univ-lyon1.fr

Abstract: Functional data analysis has gained increased attention in the past years, in particular in high-throughput biology with the use of mass spectrometry. A new challenge in functional data analysis is the availability of multisample data for which functional ANOVA becomes an appropriate framework. More specifically for spectrometry data, it is now well accepted that the noise corrupting the signal can be divided into a technical white noise added to an important inter-individual variability. In this case, the usual nonparametric regression framework (a deterministic trend corrupted by a

random noise) is no longer appropriate since it does not account for complex noise structures. Functional mixed-effects models [1] appear to be a powerful framework to handle these data, or others, but estimation in these models raises new issues and challenges. In our work, we focus on the estimation of the functional fixed effect. We assume the functional fixed effect to lie in a Besov space. This framework allows us to model curves that can exhibit strong irregularities such as peaks or jumps for instance. The lower bound for the L_2 minimax risk is provided, as well as the upper bound of the minimax rate which is derived by constructing a wavelet-based thresholding estimator for the functional fixed effect. Our approach is illustrated on realistic simulated datasets as well as on experimental data.

Keywords: Functional mixed-effects models; Wavelets; Minimax risk; Besov class.

References

[1] Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, **51**, 10, 4793–4813.

3.32 Nonparametric Regression to Model Shape Variability Using Spherical Coordinates

M. Golalizadeh* and M. Moghimbeygi

Department of Statistics, Tarbiat Modares University, Tehran, Iran; golalizadeh@modares.ac.ir, Meisam.Moghimbeygi@modares.ac.ir

Abstract: One of the popular regression-type methods to make link between some responses and covarites is nonparametric smoothing. Although following such procedure is trivial for the data on Euclidean space, some specific adoptions are required to deal with this topic for the manifold valued data. Directional and shape data are particular examples of manifold valued data (see, e.g. [3] for more details). Constructing nonparametric regression models using these typical data has recently received great attentions. It might be worth to confine on building up regression models on the simple manifold; sphere, in which earlier activity on it dates back to [5]. To recall other activities, we can mention [1], [6] and [4]. In this paper, we use nonparametric smoothing procedure to define a regression-type model on the unit sphere, defined by shape spherical coordinates, with minimizing an suitable risk function. This function gives rise to include interrelationships among angles as well as time changes. Using this capability, we are able to trace, in each time instance, the evolution of the model. This provides us a measure on how smooth the model is. Moreover, our procedures are easy to follow both from theoretical and computational view points. To illustration purpose, we apply our proposed model to analysis the well-known rat skull data, initially analyzed by [2].

Keywords: Non-Parametric regression; Manifold-valued data; Spherical coordinates; Risk function; Triangulation of shapes.

References

[1] Bhattachraya, R. and Patrangenaru, V. (2002). Nonparametic estimation of location and dispersion on Riemannian manifolds. *Journal of Statistical Planning and Inference*, **108**, 23–35.

[2] Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press. New York.

[3] Mardia, K. V. and Jupp, P. (2000). *Directional Statistics*. John Wiley and Sons. London.

[4] Di Marzio, M., Panzera, A., and Taylor, C. C. (2013). Nonparametric regression for circular responses. *Scandinavian Journal of Statistics*, **40**, 238–255.

[5] Gould, A. L. (1969). A regression technique for angular variates. *Biometrics*, **25**, 683–700.

[6] Monnier, J. B. (2011). Nonparametric regression on the hyper-sphere with uniform design. *Test*, **20**, 412–446.

3.33 Nonparametric estimation of the basic reproductive function for SIR epidemic models

A. Gribinski[1], T. Kerdreux[1] and N. Hengartner[2],⋆

[1] Ecole Polytechnique, Paris. agribinski@aol.fr and thomaskerdreux@gmail.com
[2] Theoretical Biology group, Los Alamos National Laboratory; nickh@email.edu

Abstract: Susciptible-Infected-Recovered (SIR) epidemic models, and their generalizations, provide useful descriptions of epidemic dynamics in terms of coupled differential equations. The basic reproductive function ϱ_t, which is related to the growth rate of the epidemic as a function of time can be derived for these models. For these models, it is common to estimate the model parameters by least squares fit of the models to data, event though the assumption of additive errors on the observations is unrealistic. As a result, the estimates can be biased. Furthermore, sometimes the population mixing assumption implied by the differential equations fail, which also leads to erroneous estimates for the basic reproductive function. This talk presents two contributions to answer the above criticism: First, we present a simple stochastic model that, up to minor discretization error, reproduces the dynamics of epidemics described by the differential equations. Second, we show how to modify the differential equations in a manner that makes explicit use of the reproductive function ϱ_t. Taken together, this enables us to propose a likelihood based nonparametric estimate for the basic reproductive function.

Keywords: Epidemiology; basic reproductive function; SIR.

3.34 Semi-Supervised Approaches to Efficient Evaluation of Model Prediction Performance

J. Gronsbell[1,*] and T. Cai[2]

[1] Harvard University; jgronsbell@fas.harvard.edu, tcai@hsph.harvard.edu

Abstract: In many modern machine learning applications, the outcome is expensive or time-consuming to collect while the predictor information is easy to obtain. Semi-supervised learning (SSL) aims at utilizing large amounts of 'unlabeled' data along with small amounts of 'labeled' data to improve efficiency relative to a traditional supervised approach. Though numerous SSL classification and prediction procedures have been proposed in recent years, no methods currently exist to evaluate the performance of a working model. In the context of risk prediction models derived from electronic medical records (EMRs), we present an efficient two-step estimation procedure for evaluating a binary classifier based on various prediction performance measures in the semi-supervised (SS) setting. In step I, the labeled data is used to obtain a non-parametrically calibrated estimate of the conditional risk function. In step II, SS estimates of the prediction accuracy parameters are constructed based on the estimated conditional risk function and the unlabeled data. We demonstrate that under mild regularity conditions, the proposed estimators are consistent and asymptotically normal. Importantly, the asymptotic variance of the SS estimators is always smaller than that of the supervised counterparts under correct model specification. We also correct for potential overfitting bias in the SS estimators in finite sample with cross-validation and develop a perturbation resampling procedure to approximate their distributions. Our proposals are evaluated through extensive numerical studies and illustrated with two real data analyses of EMR-based studies of rheumatoid arthritis and multiple sclerosis.

Keywords: Semi-Supervised Learning; Model Evaluation; Perturbation Resampling; Re- ceiver Operating Characteristic Curve; Risk Prediction

3.35 Uniform Strong Convergence of the Distribution Function Estimator for Associated, Truncated and Censored data

Z. Guessoum[1,*], A.Tatachak[1]

[1] Lab MSTD, Faculté de mathématiques, USTHB, BP 32 El Alia, Algiers; zguessoum@usthb.dz, atatachak@usthb.dz

Abstract: In this paper we study the strong convergence with rate of the distribution function estimator for a left truncated and right censored model (LTRC) when the lifetime observations form an associated sequence. This extends the results of [1] in i.i.d. case and [2] in α -mixing case. The performance of the estimator are illustrated through simulation studies.

Keywords: LTRC; Association.

References

[1] Gijbels, I. and Wang, J. L. (1997). Strong representations of the survival function estimator for truncated and censored data with applications. *Journal of Multivariate Analysis*, **47**, 210–229.

[2] Liang, H. Y. and De Uña-Álvarez, J. and Iglesias-Pérez, M. D. C. (2012). *Asymptotic properties of conditional distribution estimator with truncated, censored and dependent data. TEST*, **21,4**, 780–810.

3.36 The focused information criterion for high-dimensional data

Thomas Gueuning[1,*] and Gerda Claeskens[1]

[1] ORSTAT & Leuven Statistics Research Center, KU Leuven, Belgium; thomas.gueuning@kuleuven.be, gerda.claeskens@kuleuven.be

Abstract: Many variable selection procedures (such as the AIC, the BIC and the LASSO) generally select one single best model that is used to estimate all parameters of interest related to the data setting. For instance, the same model is often used for performing prediction on a number of new data points, for estimating the variance parameter and for estimating a quantile. Conversely, the focused information criterion (FIC) selects the model that best estimates a particular quantity of interest (the focus) in terms of mean squared error (MSE). The FIC can select different models for different focuses and can produce estimators with small MSE. The current FIC literature is restricted to the low-dimensional case $p < n$. In this paper, we show that the FIC idea can be extended to high-dimensional data ($p > n$). We distinguish two cases: (i) the case where the considered submodel is of low-dimension and (ii) the case where it is of high-dimension. In the former case, we obtain an alternative low-dimensional FIC formula that can directly be applied. In the latter case we use a desparsified estimator that allows us to derive the MSE of the focus estimator. We illustrate the performance of the high-dimensional FIC with a numerical study.

Keywords: Focused information criterion (FIC); High-dimensional data; Desparsified estimator.

3.37 A non-parametric Bayesian approach to decompounding from high frequency data

S. Gugushvili[1,*], F. van der Meulen[2] and P. Spreij[3]

[1] Mathematical Institute, Leiden University, The Netherlands; shota.gugushvili@math.leidenuniv.nl
[2] Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands; f.h.vandermeulen@tudelft.nl
[3] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands; spreij@uva.nl

Abstract: Given a sample from a discretely observed compound Poisson process, we consider non-parametric estimation of the density f_0 of its jump sizes, as well as of its intensity λ_0. We take a Bayesian approach to the problem and specify the prior on f_0 as the Dirichlet location mixture of normal densities. An independent prior for λ_0 is assumed to be compactly supported and possess a positive density with respect to the Lebesgue measure. We show that under suitable assumptions the posterior contracts around the pair (λ_0, f_0) at essentially (up to a logarithmic factor) the $\sqrt{n\Delta}$-rate, where n is the number of observations and Δ is the mesh size at which the process is sampled. The emphasis is on high frequency data, $\Delta \to 0$, but the obtained results are also valid for fixed Δ. In either case we assume that $n\Delta \to \infty$. Our main result implies existence of Bayesian point estimates converging (in the frequentist sense, in probability) to (λ_0, f_0) at the same rate. Simulations complement the theory.

Keywords: Compound Poisson process; Non-parametric Bayesian estimation; Posterior contraction rate; High frequency data.

References

[1] Gugushvili, S., van der Meulen, F., and Spreij, P. (2015). A non-parametric Bayesian approach to decompounding from high frequency data. arXiv:1507.03263.

[2] Gugushvili, S., van der Meulen, F., and Spreij, P. (2015). Nonparametric Bayesian inference for multidimensional compound Poisson processes. *Modern Stochastics: Theory and Applications*, **2**, 1–15.

3.38 Martingale approach for multiple testing and FDR control

A. Janssen

Mathematical Institute, Heinrich-Heine-University of Duesseldorf, Germany, janssena@math.uni-duesseldorf.de

Abstract: Under martingale dependence the false discovery rate (FDR) of various multiple tests can exactly be calculated. The results are key tools in order to discuss finite sample FDR control of these tests. Some of these results are also new when the p-values are independent. It is shown how the famous Benjamin/Hochberg multiple test can be modified. Martingale models are also helpful for step down multiple tests.
The second part of the talk discusses adaptive multiple tests with data dependent critical values. We extend the adaptive multiple test given by Storey. In Heesen and Janssen (2016) it is shown that a large class of adaptive multiple tests allow the finite sample FDR control.
The martingale method is very powerful and has two aspects. This finite sample FDR control can be extended to various dependence concept. On the other hand even new results are partially derived for multiple test based on independent p-values. Some results also hold for set down multiple tests.

Keywords: False discovery rate (FDR); adaptive Benjamini Hochberg methods; Storey test.

References

[1] P. Heesen and A. Janssen (2015) Inequalities for the false discovery rate (FDR) under dependence, *Electron. J. Stat.*, **9**, 679–716.

[2] P. Heesen and A. Janssen (2016) Dynamic adaptive multiple tests with finite sample FDR control, *J. Statist. Plann. Inference*, **168**, 38–51.

3.39 Adaptive Bayesian estimation in indirect Gaussian sequence space models

J. Johannes[1,*], A. Simoni[2] and R. Schenk

[1] Ruprecht-Karls-Universität Heidelberg, Institut für Angewandte Mathematik, Heidelberg, Germany; johannes@math.uni-heidelberg.de
[2] CREST, 15 Boulevard Gabriel Péri, France; simoni.anna@gmail.com

Abstract: In an indirect Gaussian sequence space model lower and upper bounds are derived for the concentration rate of the posterior distribution of the parameter of interest shrinking to the parameter value θ° that generates the data. While this establishes posterior consistency, however, the concentration rate depends on both θ° and a tuning parameter which enters the prior distribution. We first provide an oracle optimal choice of the tuning parameter, i.e., optimized for each θ° separately. The optimal choice of the prior distribution allows us to derive an oracle optimal concentration rate of the associated posterior distribution. Moreover, for a given class of parameters and a suitable choice of the tuning parameter, we show that the resulting uniform concentration rate over the given class is optimal in a minimax sense. Finally, we construct a hierarchical prior that is adaptive. This means that, given a parameter θ° or a class of parameters, respectively, the posterior distribution contracts at the oracle rate or at the minimax rate over the class. Notably, the hierarchical prior does not depend neither on θ° nor on the given class. Moreover, convergence of the fully data-driven Bayes estimator at the oracle or at the minimax rate is established.

Keywords: Bayesian nonparametrics; Hierarchical Bayes; Exact concentration rates; Oracle optimality; Minimax theory.

References

[1] Johannes, J., Simoni, A. and Schenk, R. (2015). Adaptive Bayesian estimation in indirect Gaussian sequence space models. Discussion paper, arXiv:1502.00184.

3.40 Nonlinear spectral analysis via the local Gaussian correlation

L. A. Jordanger[1,*] and D. Tjøstheim[1]

[1] University of Bergen, Norway; Lars.Jordanger@uib.no, Dag.Tjostheim@uib.no

Abstract: Spectral analysis can reveal interesting properties of time series, but it's not able to distinguish dependent, but uncorrelated, time series (e.g. GARCH-models) from white noise.
We define the local Gaussian spectral density by replacing the autocovariances in the spectral density with local Gaussian autocorrelations [1], and investigate if this local approach can be used as a tool when analysing nonlinear time series.
This talk will define the local Gaussian spectral density function, explain how to estimate it, and give an asymptotic result for the convergence properties of the estimate.
An interactive visualisation tool will be used to present a few examples.

Keywords: GARCH-models; Local Correlation; Central Limit Theorem; Graphical Tools.

References

[1] Tjøstheim D., Hufthammer K. O. (2013). Local Gaussian correlation: A new measure of dependence. *Journal of Econometrics*, **172(1)**, 33–48.

3.41 On the locally most powerful sequential rank tests

J. Kalina[1]

[1] Institute of Computer Science of the Academy of Sciences of the Czech Republic; kalina@cs.cas.cz

Abstract: This paper fills the gap of the optimality results for various hypothesis tests based on sequential ranks. While the classical locally most powerful rank statistics are known to be obtained as projections of the locally most powerful (parametric) statistics into the space of linear rank statistics, we show that this does not hold for sequential rank statistics. Thus, the results bring arguments in favor of some of the previously used test statistic based on sequential ranks [2]. We derive the locally most powerful sequential rank test for the hypothesis of randomness against a general alternative, including the two-sample difference in location or regression in location as special cases for the alternative hypothesis. Further, the locally most powerful sequential rank tests are derived for the one-sample problem and for independence of two samples in an analogous spirit as the classical results of [1] for classical ranks. The locally most powerful tests are derived for a fixed sample size. More importantly, we propose a sequential testing procedure based on these statistics of the locally most powerful tests. We illustrate that the sequential test comes to the conclusion more quickly compared to a test based on classical ranks, which is especially appealing if acquiring new observations is expensive.

Keywords: Sequential ranks; Nonparametric tests; Linear rank statistic; Stopping variable.

References

[1] Hájek, J. and Šidák, Z. (1967). *Theory of rank tests.* Academia, Prague & Academic Press, New York.

[2] Lombard F. and Mason D.M. (1985). Limit theorems for generalized sequential rank statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **70**, 395–410.

3.42 Admissibility of k-nn type smoothers

P. A. Cornillon[1], A. Gribinski[2], N. Hengartner[3], T. Kerdreux[2*] and E. Matzner-Løber[1]

[1] IRMAR, Univ Rennes 2, France, [2] Polytechnique, France, [3] Los Alamos National Laboratory, USA

Abstract: The aim of the article is to study some transformations of smoothers used in non parametric regression. Symmetric smoother with eigen values in $[0, 1]$ (such as smoothing spline) have nice properties as for example admissibility. We will show that symmetrisation of row stochastic smoother such as geometric mean or arithmetic mean lead to smoother with eigen values bigger than one.

More specifically, we give a systematic method to transform smoothers into admissible estimators. We will focus on k-nearest neighbor type (classical, mutual, symmetric) smoothers and we propose an estimator symmetric with eigen values in $[0, 1]$ which could be evaluated at any points.

Keywords: Nonparametric smoother, k-nn, mutual, row stochastic smoother

3.43 Algorithmic leveraging and elemental estimation

K. Knight[1,*]

[1] University of Toronto; keith@utstat.toronto.edu

Abstract: Algorithmic leveraging [1] is a popular method for approximating least squares estimation by subsampling using the leverage scores (the diagonals of the "hat" matrix) as an importance sampling distribution. We will compare this approach to one in which random subsets of observations are sampled in order to construct elemental estimates that are then averaged in some way to approximate least squares estimates.

Keywords: Algorithmic leveraging; Elemental estimation; Least squares estimation.

References

[1] Ma, P., Mahoney, M. W. and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, **16**, 861-911.

3.44 Moving Average Sieve Bootstrap

J. Krampe[1,*], J. Kreiss[2] and E Paparoditis[2]

[1] TU Braunschweig; j.krampe@tu-bs.de, j.kreiss@tu-bs.de
[2] University of Cyprus; stathisp@ucy.ac.cy

Abstract: For nondeterministic stationary processes with spectral densities, Szegö's factorization of the spectral density can be used to get a moving average (MA) representation of the spectral density, similar to the Wold decomposition. We use these MA coefficients to generate a Wild MA Sieve Bootstrap, which is able to mimic the second order structure of the underlying process. If the underlying process is linear, the MA Sieve Bootstrap is able to mimic to the necessary extent the fourth order moment structure, which occurs in statistics like the empirical autocovariance. We prove that the proposed bootstrap procedure asymptotically works for the mean and a general class of statistics, which contains the empirical autocovariance and autocorrelation, among others. Furthermore, we compare it to the linear process bootstrap and the AR-Sieve bootstrap. Its finite sample performance is investigated by means of simulations.

Keywords: stationary time series, bootstrap, moving average representation

3.45 Nonparametric estimation of the intensity function from indirect Poisson point process observations

Martin Kroll[*]

[*]Universität Mannheim; martin.kroll@math.uni-mannheim.de

Abstract: We consider the statistical inverse problem of estimating the intensity function of a Poisson point process on $[0, 1)$ from a size n sample of noisy observations. The additive noise is considered modulo $[0, 1)$ and its distribution assumed to be unknown: to ensure identifiability and enable inference we require the availability of an additional size m sample from the noise distribution. In the first part, we take up a minimax point of view and study the rate of convergence for the estimation of the intensity with respect to weighted mean integrated squared error under smoothness assumptions on the intensity and the noise distribution. We derive minimax lower bounds with respect to the sample sizes n and m and suggest an orthogonal series estimator containing an additional cut-off addressing to the missing knowledge of the noise distribution. This estimator is shown to attain the minimax lower bounds and thus to be rate optimal. However, it depends on the smoothness characteristics of the intensity and the noise distribution. Thus, in the second part, we introduce an adaptive estimator of the intensity which automatically adapts to unknown smoothness. In

a preliminary step, we define a partially-adaptive estimator depending only on the smoothness of the noise distribution but not on the one of the intensity. Building on this, we define a fully-adaptive estimator which does not rely on any prior smoothness assumptions. The adaptive estimators are shown to attain the optimal rates in a wide range of scenarios.

Keywords: Poisson point process; Intensity function; Statistical inverse problem; Minimax; Adaptive estimation.

3.46 Shape-constrained uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider

M. Kuusela[1,*] and P. B. Stark[2]

[1] École Polytechnique Fédérale de Lausanne, EPFL Switzerland; mikael.kuusela@epfl.ch
[2] Department of Statistics, University of California, Berkeley, US; stark@stat.berkeley.edu

Abstract: The high energy physics unfolding problem is an important statistical inverse problem arising in data analysis at the Large Hadron Collider at CERN. The problem arises in making nonparametric inferences about a particle spectrum from measurements smeared by the finite resolution of the particle detectors. Existing unfolding methodology has major practical limitations stemming from ad hoc discretization and regularization of the problem. As a result, confidence intervals derived using the current methods can have significantly lower coverage than expected. In this work, we regularize the problem by imposing physically justified shape constraints (positivity, monotonicity and convexity). We quantify the uncertainty by constructing a nonparametric confidence set for the true spectrum consisting of all those spectra that satisfy the shape constraints and that predict observations within an appropriately calibrated level of fit to the data. Projecting that set produces simultaneous confidence intervals for all functionals of the spectrum, including averages within bins. The confidence intervals have guaranteed frequentist finite-sample coverage in the important and challenging class of unfolding problems with steeply falling particle spectra. We demonstrate the efficacy of the method using simulations designed to mimic the unfolding of the inclusive jet transverse momentum spectrum at the Large Hadron Collider. The shape-constrained intervals provide usefully tight conservative confidence intervals, while the conventional methods suffer from severe undercoverage.

Keywords: Poisson inverse problem; Finite-sample coverage; High energy physics; Fenchel duality; Semi-infinite programming.

3.47 On a robust and nonparametric estimation of the covariance matrix in the portfolio selection problem

Henry Laniado[1,*], Sergio Botero-Botero[2]

[1] School of Mines, Universidad Nacional de Colombia, Medellín and Department of Mathematical Sciences, Universidad Eafit, Medellín; hlaniaor@unal.edu.co, hlaniado@eafit.edu.co
[2] School of Mines, Universidad Nacional de Colombia, Medellín; sbotero@unal.edu.co

Abstract: The mean-variance model for portfolio optimization, proposed by [2] has become in the starting point for modern portfolio theory. This can be considered as one of the first models that introduces the concept of diversification. The model considered by Markowitz states that an investor should hold a portfolio on the set of risk and return couples that cannot be improved simultaneously, this set is named efficient frontier. To implement Markowitz, we needs to know the first and second distributional moments of stock returns, which are in practice unknown. Hence, they are estimated with sample information usually through the maximum likelihood estimators, but but according to [1] this kind of estimators bring a lot of estimation error in special in the estimation of mean of returns and thus they perform poorly out of sample due to estimation error. In this talk, we show a simple technique for finding portfolios with better performance out of sample. This new methodology just is focused for solving the classical minimum variance problem by using a modified version of the covariances matrix. We propose both a robust and nonparametric estimation of the covariances matrix of stock returns based on replacing the covariances by a functions that depend on other robust coefficient of dependency, for example, Kendall tau or Spearman. We show on simulated and empirical data that our approach works well out of sample in terms of Sharpe ratio. Finally, we will present the main conclusions and some future research lines whose work is currently underway

Keywords: Portfolio selection; Covariance matrix; Robust estimation; Nonparametric estimation.

References

[1] DeMiguel V., Nogales, F.J., (2009). Portfolio Selection with Robust Estimation. *Operations Research*, **7**, 1–18.

[2] Markowitz, H., (1952). Portfolio selection. *The Journal of Finance*, **7**, 77–91.

3.48 Nonparametric Data Analysis for the Market Size Prediction to Estimate R&D benefits in the Korean Feasibility Studies

Hyunsook Lee[1,*]

[1] Korea Institute of S&T Evaluaton and Planning; hlee@kistep.re.kr

Abstract: Feasibility study has been utilized in Korean budgeting process in order to enhance the budget efficiency. Predicting the market size of target products after the R&D program for a technology life cycle plays a key role in deciding the budget allocation plausibility of the newly proposed government R&D program where innovation sometimes brings a jackpot. Despite its critical role in benefit estimation, resources and adaptable growth curve models to predict a technology related market size are generally limited and often estimated benefits are biased. Information from relevant or similar markets can be incorporated to reduce uncertainties in estimating benefits from the government R&D investments. Although no mathematical scientific procedure has been introduced, adapting external information benefits market prediction and benefit estimation more in a feasible and scientific way. In this presentation, methodologies like kernel density estimation(KDE) and data depth(DD) are applied to data from KOSIS(KOrean Statistical Information Service) and KITA(Korea International Trade Association). KDE assists to understand the properties of manuafacturing markets and to refine target markets for benefit estimation in feasibility studies. DD describes characteristics of target technology in manufacturing markets with sales, exports, imports, and associated variables characterizing the nation's industry. The knowledge acquired from KDE and DD will be applied to model selection and curve estimation. Such process will be demonstrated with a recently finalized feasiblity study about machinery.

Keywords: Kernel Density Estimation; Data Depth; Market Prediction; Feasibility Studies, R&D Policy

3.49 Goodness-of-fit test for multistable Lévy processes

R. Le Guével[1]

[1] ronan.leguevel@univ-rennes2.fr

Abstract: Multistable processes, that is, processes which are, at each "time", tangent to a stable process, but where the index of stability varies along the path, have been recently introduced as models for phenomena where the intensity of jumps is non constant. We will present how to estimate two functions of interest of theses models, and we will explain how we can obtain a statistical test in order to decide if a data set comes from a stable process or a multistable one.

Keywords: Nonstationary models; Locally stable processes; Locally self-similar processes; Jump processes; Localisable processes

3.50 Bandwidth Selection for Smoothing Sparse and Non-Sparse Functional Data with Covariate Adjustments

D. Liebl[1]

[1] University of Bonn; dliebl@uni-bonn.de

Abstract: This paper deals with the nonparametric estimation of the conditional mean and covariance function of a stationary time series of weakly dependent random functions with covariate-adjustments. As in the context of sparse functional data, it is assumed that only the noisy discretization points of a random function are observable. Estimation is done using classical multivariate local linear estimators.

By using a double asymptotic we consider all cases from sparsely to densely sampled discretization points per function and therefore take into account the vague cases typically found in applications. We show that the asymptotic first- and second-order variance terms of the estimators can switch places depending on the asymptotic scenario. This has a surprising effect on the solution of the involved multiple bandwidth selection problem. By contrast to all classical

bandwidth results, it becomes optimal to choose the bandwidths *anti*-proportionally to each other as soon as the number of observed discretization points per function is diverging at a certain (relatively slow) rate. We further show that ignoring these results and using the classical bandwidth expressions instead can lead to a diverging asymptotic mean integrated squared error.

Our research is motivated by the problem of estimating and testing the differences in the electricity prices before and after Germany's abrupt nuclear phaseout after the nuclear disaster in Fukushima Daiichi, Japan, in mid-March 2011.

Keywords: Functional data analysis; Local linear estimation; Multiple bandwidth selection; Time series analysis

3.51 Measuring the Algorithmic Convergence of Random Forests via Bootstrap Extrapolation

M. E. Lopes[1]

[1]UC Davis, Department of Statistics; melopes@ucdavis.edu

Abstract:

When making predictions with a voting rule, a basic question arises: "What is the smallest number of votes needed to make a good prediction?" In the context of ensemble classifiers, such as Random Forests or Bagging, this question represents a tradeoff between computational cost and statistical performance. Namely, by paying a larger computational price for more classifiers, the prediction error of the ensemble tends to improve and become more stable. Conversely, by using fewer classifiers and tolerating some variability in accuracy, it is possible to speed up the tasks of training and making new predictions. In this paper, we propose a bootstrap method to quantify this tradeoff for the methods of Bagging and Random Forests. To be specific, suppose the training dataset is fixed, and let the random variable Err_t denote the prediction error of a randomly generated ensemble of $t = 1, 2, \ldots$ classifiers. (The randomness of Err_t comes only from the algorithmic randomness of the ensemble.) Working under a "first order model" of Random Forests, we prove that the centered law of Err_t can be consistently estimated via our proposed method as t diverges. As a consequence, this result offers practitioners a guideline for choosing the smallest number of base classifiers needed to ensure that the algorithmic fluctuations are negligible (e.g. the variance of Err_t being less than a given threshold). Lastly, we explain how the technique of "extrapolation" can be used to substantially reduce the computational cost of resampling.

Keywords: Random Forests; Bagging; Bootstrap; Extrapolation; Time-Data Tradeoffs

3.52 Adaptive deconvolution of survival function on the nonnegative real line

G. Mabon[1,2*]

[1] CREST - ENSAE, *3 avenue Pierre Larousse*, 92245 Malakoff, France
[2] Université Paris Descartes, *45 rue des Saints-Pères*, 75006 Paris, France; gwennaelle.mabon@ensae.fr

Abstract: We consider the problem of adaptive estimation in the convolution model when both random variables are nonnegative. The goal is to recover the survival function of the target random variables when the error distribution is known. We want to emphasize that the estimation of the survival function does not rely on the estimation of the density. This issue can be seen as a classical statistical deconvolution problem which has been tackled in many cases using Fourier-type approaches. Nonetheless, in the present case the random variables have the particularity to be $\mathbb{R}^+$-supported. Knowing that, we propose a new angle of attack by building a projection estimator with an appropriate Laguerre basis. We present a data driven strategy for selecting a relevant projection space for the $\mathbb{L}^2$-risk based on a penalized criterion à la Birgé and Massart (1997). Our procedure achieves faster convergence rates than Fourier methods for estimating Gamma type functions. The procedure is illustrated with simulated data.

Keywords: Deconvolution; Adaptive estimation; Survival analysis; Laguerre basis.

References

[1] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation, In *Festschrift for Lucien Le Cam*, pages 55-87, Springer. New-York.

[2] Mabon, G. (2014). *Adaptive deconvolution on the nonnegative real line.* preprint MAP5 2014-33.

3.53 Change-points and Shape Constrains in LASSO Regularized Nonparametric Regression

M. Maciak[1,*] and I. Mizera[2]

[1] Dpt Probability and Mathematical Statistics, Charles University in Prague, Czech Republic; maciak@karlin.mff.cuni.cz
[2] Department of Mathematical Sciences, University of Alberta, Edmonton, Canada; mizera@ualberta.ca

Abstract: Estimating various types of structural changes in some unknown dependence structure is an important aspect in regression modelling approaches especially when dealing with more complex data where sudden changes are expected.
Our estimation method is motivated by some machine learning ideas and using recent developments in statistic, post-selection inference especially, we propose a fully data driven estimation approach based on LASSO regularization techniques where the unknown model and existing change-points in the model are estimated all at once. The estimation approach a-priori considers all possible model alternatives which makes the method suitable in situations where no knowledge on the number or position of change-points is given in advance.
The estimation approach also allows for different structures of change-points to be considered. In addition, some restrictions with respect to the overall shape of the unknown regression function can be implemented in a relatively straightforward way.
We will discuss some theoretical results and practical examples and applications too.

Keywords: Change-points; Shape constrains; Nonparametric regression; LASSO;

3.54 Bayesian Matrix Completion under General Sampling Distribution

T.Tien Mai* & P. Alquier

CREST, ENSAE, University Paris Saclay *TheTien.Mai@ensae.fr

Abstract: Bayesian methods for low-rank matrix completion with noise have been shown to be very efficient computationally. While the behaviour of penalized minimization methods is well understood both from the theoretical and computational points of view in this problem, the theoretical optimality of Bayesian estimators have not been explored yet. In this work, we propose a Bayesian estimator for matrix completion under general sampling distribution. We also provide an oracle inequality for this estimator. This inequality proves that, whatever the rank of the matrix to be estimated, our estimator reaches the minimax-optimal rate of convergence (up to a log term). We end the paper with a short simulation study.
This work was done when the first author was at UCD and INSIGHT, Ireland

Keywords: Matrix completion, PAC-Bayesian Analysis, Oracle inequality

References

[1] Mai, T. Tien and Alquier, P. (2015). A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electron. J. Statist.*, **9**, 823–841.

3.55 Asymptotic equivalence for pure jump Levy processes with unknown Levy density and Gaussian white noise

Ester Mariucci

LJK, Univ. Joseph Fourier Grenoble

Abstract: The aim of this paper is to establish a global asymptotic equivalence between the experiments generated by the discrete (high frequency) or continuous observation of a path of a L evy process and a Gaussian white noise experiment observed up to a time T, with T tending to 1. These approximations are given in the sense of the Le Cam distance, under some smoothness conditions on the unknown L evy density. All the asymptotic equivalences are established by constructing explicit Markov kernels that can be used to reproduce one experiment from the other.

Keywords: Nonparametric experiments, Le Cam distance, asymptotic equivalence, Levy processes.

3.56 Scale and Space Inference of Nonhomogeneous Poisson Process Intensities

M.L. Gámiz[1], A.J. López-Montoya[2], M.D. Martínez-Miranda[3,*] and R. Raya-Miranda[4]

[1] University of Granada; mgamiz@ugr.es, ajlopez@ugr.es mmiranda@ugr.es. rraya@ugr.es

Abstract: The main purpose of this work is to provide an effective graphic tool to explore the underlying characteristics of the intensity of a nonhomogeneous Poisson process (NHPP), in order to detect constant or monotonic patterns, or possible trend changes. To this goal we develop an extension of the graphical tool SiZer Map introduced by Chaudhuri and Marron (1999). We develop SiZer Map considering local linear kernel estimators for the intensity function and its first derivative, where the bandwidth parameter is the viewing scale and the considered time interval is the localization space. The shape characteristics of the intensity function are distinguished from those which are merely an artifact of the sampling variability in the data, using confidence intervals for the first derivative. We consider pointwise and simultaneous confidence intervals, using the normal limiting distribution and the bootstrap method described by [2]. Our proposal is illustrated using real datasets, consisting on the time occurrences of events recurrent in time; and its performance is evaluated through an extensive simulation study.

Keywords: SiZer Map; Intensity; Bootstrap; Simultaneous Inference

References

[1] Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structure in curves. *Journal of the American Statistical Association*, **94**, 807–823.

[2] Cowling, A., Hall, P. and Phillips, M.J. (1996). Bootstrap confidence regions for the intensity of a Poisson point process. *Journal of the American Statistical Association*, **91**, 1516–1524.

3.57 Nonparametric hypothesis testing of the human microbiome using evolutionary trees

M. Mincheva[1], H. Li[2] and J. Chen[3]

[1] Temple University, mincheva@temple.edu
[2]University of Pennsylvania, hongzhe@upenn.edu
[3] Mayo Clinic, chen.jun2@mayo.edu

Abstract: Thanks to next-generation sequencing technologies, researchers can gain access to millions of DNA sequences from a single experiment. A major area of interest has been the analysis of the human micro biome, which represents the total genome of the microorganisms living in the human body. The human body contains about 10^{13} human cells and 10^{14} bacterial cells, so the microbiome is frequently perceived as an extended human genome. Different parts of the body have distinct bacterial compositions, that are closely related to the presence of diseases, such as inflammatory bowel disease, peripheral vascular disease, asthma and hypertension. That is why, developing statistical procedures for comparison and identification of microbial diversity can be essential in the early diagnosis or curing of these and many other diseases. In this project, we develop a two-sample nonparametric test statistic to test the hypothesis that the overall microbial compositions of two samples are different. We also establish its theoretical properties and nice asymptotic F-distribution. None of the existing methods establish exact distributions and rely only on permutation algorithms (PERMANOVA). We demonstrate through simulations and real data example the superiority of the type 1 error and computational efficiency of our method over the rest of them.

Keywords: Phylogenetic tree; Nonparametric test statistic; Human microbiome; Permanova; F-distribution.

3.58 Non-parametric lower bounds and information functions

S.Y. Novak[1,*]

[1] Middlesex University London, SST, The Burroughs NW44BT, UK; S.Novak@mdx.ac.uk

Abstract: We introduce the notions of information index and information function and present a non-parametric generalisation of the Fréchet–Rao–Cramér inequality. A regularity condition involves the Hellinger or the χ^2 distance. We show that if the information index is larger than two, then unbiased estimators do not exist. Suppose one wants to estimate a quantity of interest, a_P, from a sample $X_1, ..., X_n$ of i.i.d. observations from an unknown distribution P. Let d_H^2 and d_χ^2 denote Hellinger and χ^2 distances, respectively. We introduce an information index ν that indicates how "rich" or "poor" the class $\mathcal{P}$ is. We show that the best possible accuracy of estimation is $O(n^{-1/\nu})$. In the case of a regular parametric family of distributions $\nu = 2$. "Irregular" parametric families of distributions may obey the regularity conditions with $\nu < 2$. Parametric subfamilies of non–parametric classes typically obey the regularity conditions with $\nu > 2$, cf. [1]. We present a lower bound, which indicates that the accuracy of estimation is determined by the information index and the information function. A remarkable fact is that there are no unbiased estimators with finite second moment if the regularity conditions hold with $\nu > 2$. Namely, if a regularity condition holds and $\sup\limits_{P\in\mathcal{P}} \mathbb{E}_P\|\hat{a}_n - a_P\|^2 < \infty$, then an arbitrary estimator $\hat{a}_n$ is necessarily biased. On the other hand, the accuracy of non-parametric in terms of the Hellinger distance might be of order $1/n$, as suggested by a lower bound to $\sup_{P\in\mathcal{P}} \mathbb{E}_P d_H^2(\hat{a}_n; a_P)$.

Keywords: Non-parametric lower bounds; information index; information function.

References

[1] Novak S.Y. (2011). *Extreme value methods with applications to finance*. Taylor & Francis/CRC Press. London. ISBN: 9781-43983-5746.

3.59 Estimating multivariate and conditional density functions using local Gaussian approximations

H. Otneim[1*] and D. Tjøstheim[1]

[1] University of Bergen; hakon.otneim@math.uib.no, dag.tjostheim@math.uib.no

Abstract: It is well known that the nonparametric kernel density estimator breaks down quickly as the dimension increases. In this talk, we will present a procedure for multivariate density estimation that makes use of the local likelihood framework, but with crucial adaptations so that we circumvent the curse of dimensionality without introducing too much modeling error. This method, that is based on local Gaussian approximations, can be used for conditional density estimation in a natural way. Simulation experiments and applications to real data will illustrate the finite sample performance of the estimator, and asymptotic results will be presented as well.

Keywords: Density estimation; Local likelihood; Multivariate; Curse of dimensionality

3.60 permute: A Python Package for Randomization Inference

K. Ottoboni[1,*], J. Millman[2] and P.B. Stark[1]

[1] Department of Statistics, UC Berkeley; kellieotto@berkeley.edu, pbstark@berkeley.edu
[2] Division of Biostatistics, UC Berkeley; millman@berkeley.edu

Abstract: Software packages for randomization inference are few and far between.

This forces researchers either to rely on specialized stand-alone programs or to use classical statistical tests, which may require implausible assumptions about their data-generating process. As Python gains popularity as a language for carrying out data analysis from start to finish, the absence of a package for randomization inference presents a severe limit to users' statistical capabilities. We present `permute`, the first (to our knowledge) comprehensive Python package for randomization inference. We illustrate the program's capabilities with three examples:

- a randomized experiment comparing the student evaluations of teaching for male and female instructors [1]
- a study of the association between salt consumption and mortality at the level of nations
- an assessment of inter-rater reliability for a series of labels assigned by multiple raters to video footage of children on the autism spectrum

We discuss future plans for `permute` and the role of software development in Statistics.

Keywords: Software; Permutation tests; Python; Two-sample problem; Inter-rater reliability

References

[1] MacNell, L. and Driscoll, A. and Hunt, A. N. (2014). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, 1–13.

3.61 Robust testing for superiority between two regression curves

J.C. Pardo-Fernández[1,*] and G. Boente[2]

[1] Departamento de Estatística e Investigación Operativa, Universidade de Vigo, Spain; juancp@uvigo.es
[2] Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Argentina; gboente@dm.uba.ar

Abstract: In this talk we will focus on the problem of testing the null hypothesis that the regression functions of two populations are equal versus one-sided alternatives under a general nonparametric homoscedastic regression model. To protect against atypical observations, the test statistic is based on the residuals obtained by using a robust estimate for the regression function under the null hypothesis. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root-n local alternatives. A Monte Carlo study is performed to compare the finite sample behaviour of the proposed tests with the classical one obtained using local averages. A sensitivity analysis is carried on a real data set.

Keywords: Hypothesis testing; Nonparametric regression models; Robust inference; Smoothing techniques.

3.62 FLAME: a Functional Linear Adaptive Mixed Estimation for high dimensional functional-on-scalar regression

A. C. L. Parodi[1], M. Reimherr[2,*]

[1] MOX - Department of Mathematics, Politecnico di Milano, Italy; alicecarla.parodi@polimi.it
[2] Statistics Department, Pennsylvania State Univeristy, PA, USA; mreimherr@psu.edu

Abstract: Functional-on-scalar regression is a primary problem in Functional Data Analysis; this work introduces a new method for estimation and variable selection with a large number of predictors. Recently a number of techniques have been developed, such as FS-LASSO ([1]), AFS-LASSO ([2]) and FPCA based estimation ([3]), but none fully exploits the smoothness of the underlying parameters. Here we introduce a method, called FLAME, which achieves simultaneous variable selection, estimation and smoothing. Specifically, the coefficients are embedded in an Hilbert space which can be different from the space of the data. Here we choose a Reproducible Kernel Hilbert Space, so that the identification of a proper kernel allows us to tune the smoothness of the estimators. A coordinate descent algorithm is adopted and the use of C++ and R guarantees computational power. Simulations will be presented to highlight the efficacy of our method over existing ones. Time permitting, asymptotic theory will also be discussed.

Keywords: FLAME; functional-on-scalar regression; RKHS.

References

[1] Barber, R F, Reimherr, M and Schill, T (2015). *The functional-on-scalar lasso with applications to longitudinal GWAS*. Under revision.

[2] Fan, J and Reimherr, M (2015). *High-Dimensional Adaptive Function-on-Scale Regression*. Under revision.

[3] Chena, Y, Goldsmitha, J, Ogdena T (2015). *Variable Selection in Function-on-Scalar Regression*. Under revision.

3.63 A stochastic process approach to multilayer neutron detectors

V. Pastukhov[1], D. Anevski[1*] and R. H. Wilton[2]

[1] Department of Mathematical Statistics, Lund University, Sweden; pastuhov@math.lth.se, dragi@maths.lth.se
[2] European Spallation Sourse, Lund, Sweden; john.doe@email.edu; richard.hall-wilton@esss.se

Abstract: We discuss the feasibility of statistical determination of a neutron wavelength and energy spectrum in the new generation of neutron detectors. The data from the multi-grid detector consists of counts of the number of absorbed neutrons, along the sequence of the detector cells. First, we consider the unimodal incident beam which is assumed to be a Poisson process. Using the Maximum Likelihood (ML) estimator we discuss its asymptotic properties. Next, we generalise this result for the case of the spectrum of a multimodal Poisson beam. The last part is the ultimate challenge and is dedicated to the estimation the continuous spectrum of the incident neutron beam, under the assumption of monotonicity.

Keywords: Poisson process, monotonicity, asymptotic distribution, neutron scattering

3.64 Nonparametric and Semiparametric Inference for Signal Symmetries

Mirosław Pawlak

Department of Electrical and Computer Eng., University of Manitoba, Canada; Miroslaw.Pawlak@umanitoba.ca

Abstract: Symmetry plays an important role in signal/image understanding and recognition. This paper formulates the problem of assessing reflectional symmetries of a signal/image function observed in the presence of noise. Rigorous nonparametric statistical tests are developed for testing image invariance under reflections. The symmetry relation is expressed as the restriction for Fourier coefficients with respect to a class of radial orthogonal functions. Therefore, our test statistics are based on checking whether the estimated radial coefficients approximately satisfy those restrictions. We derive the asymptotic distribution of the test statistics under both the hypothesis of symmetry and under fixed alternatives. We also examine the semi-parametric problem of estimating parameters of a given type of signal/image symmetry, e.g., estimating the axis of reflectional symmetry. The issue of semi-parametric efficiency is also addressed.

Keywords: Symmetry detection, symmetry estimation, radial polynomials, limit distribution, degree of symmetry.

3.65 Nonparametric Union-Intersection Approach in Multivariate Permutation Tests

Fortunato Pesarin[1] and Luigi Salmaso[2*]

[1] Department of Statistical Sciences, University of Padova, Italy; pesarin@stat.unipd.it
[2] Department of Management and Engineering, University of Padova, Italy; luigi.salmaso@unipd.it

Abstract: The Union-Itersection (UI) principle for multivariate testing has quite a long story since [**?**]. This approach assumes that the hypotheses H_0 and H_1 can be equivalently written as $H_0 \equiv \bigcap_{k=1}^{K} H_{0k}$ and $H_1 \equiv \bigcup_{k=1}^{K} H_{1k}$, and that the global test is carried out by combining a suitable list of partial tests T_k, each specific for H_{0k} versus $H_{1k}, k = 1, \ldots, K$. When they are known, the provided UI solutions are generally coincident with those obtained by likelihood techniques in the conditions for the latter. But outside that setting no further problems have found solutions in closed form. In this respect [**?**] says "The crux of the problem is however to find the distribution theory for the maximum of these possibly correlated statistics. Unfortunately, this distribution depends on the unknown F, even under the null hypothesis. An easy way to eliminate this impasse is to take recourse to the permutation distribution theory". A proposal, however, which is far from being easy to achieve if one wishes to manage the underlying dependence (which can be much more complex than linear) by using suitable estimators of all coefficients, the number and type of which are usually unknown. It has a general solution when it is possible to handle such a dependence in a nonparametric way. This goal is obtained within the conditional testing principle by conditioning on the whole data set. A principle which, when under H_0 the whole data set $\mathbf{X}$ is sufficient for the underlying distribution F, provides for exact permutation solutions even in multivariate settings and constrained alternatives. The related methods are based on the so-called "NonParametric Combination (NPC) of dependent permutation tests" [**? ? ?**]. The principal goal of present talk is discussing on main properties of the UI-NPC methodology and presenting some of its multivariate applications.

Keywords: Nonparametric combination; Permutation tests; Multivariate permutation tests; Union-intersection principle.

3.66 Conditional least squares and copulae for panel data

M. Pešta[1,*] and O. Okhrin[2]

[1] Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics; michal.pesta@mff.cuni.cz
[2] Dresden University of Technology, School of Transportation, Institute of Economics and Transport; ostap.okhrin@tu-dresden.de

Abstract: A general time series model, that allows for modeling the conditional mean and variance, is proposed for panel data. The time series innovations are not considered to be independent. Conditional least squares are used to estimate model parameters [1]. Consistency of the estimates is proved. The copula approach is applied for modeling the dependence structure, which facilitates to make predictions. Moreover, semiparametric bootstrap is employed in order to estimate the predictions' distribution. The consistency of the parameter estimates provides validity for bootstrapping. Real data examples from non-life insurance [2] are provided as an illustration of the potential benefits of the presented approach.

Keywords: Panel data; Conditional least squares; Dependency modeling; Copula; Semiparametric bootstrap.

References

[1] Pešta, M. and Okhrin, O. (2014). Conditional least squares and copulae in claims reserving for a single line of business. *Insurance: Mathematics and Economics*, **56**(1), 28–37.

[2] Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance.* John Wiley & Sons. Hoboken, NJ.

3.67 An Efficient Analysis of Change Points via Swarm Intelligence

L.L.H. Chang[1], H.H. Chan[1] and F.K.H. Phoa[1*]

[1] Institute of Statistical Science, Academia Sinica; fredphoa@stat.sinica.edu.tw

Abstract: Evolutionary algorithm has received numerous attentions in recent statistics researches. In this paper, we propose an efficient method via swarm intelligence, namely SIBCP, for analyzing composite functions that consists of multiple change points. Instead of simply applying the standard SIB framework, SIBCP introduces a new operation called VARY that allows the adjustment of number of change points being included during the optimization. Numerical results show that SIBCP successfully points out the location of change points accurately by using a small number of iterations, and the deviation between the fitted model and the true function is small.

Keywords: Change Points; Swarm Intelligence.

3.68 A Non-Parametric Inferential Framework for Domain Selection in Functional Data Analysis

A. Pini[1,*], L. Spreafico[2], S. Vantini[1], and A. Vietti[2]

[1] MOX - Department of Mathematics, Politecnico di Milano; alessia.pini@polimi.it, simone.vantini@polimi.it
[2] ALPs - Alpine Laboratory of Phonetics and Phonology – Free University of Bozen-Bolzano; lorenzo.spreafico@unibz.it, alessandro.vietti@unibz.it

Abstract: In the framework of null hypothesis significance testing for functional data we propose a procedure able to select intervals of the domain imputable for the rejection of a null hypothesis. An unadjusted p-value function and an adjusted one are the output of the procedure, namely, Interval-Wise Testing [1]. Depending on the sort and level α of type-I error control, significant intervals can be selected by thresholding the two p-value functions at level α. We prove that the unadjusted (adjusted) p-value function *point-wise* (*interval-wise*) controls the probability of type-I error and it is *point-wise* (*interval-wise*) consistent. To enlighten the gain in terms of interpretation of the phenomenon under study, we applied the Interval-Wise Testing to the analysis of tongue profiles. Specifically, we analyze a dataset of tongue shapes recorded for a study on Tyrolean, a German dialect spoken in South Tyrol [2]. Data are composed of 160 tongue profiles on the sagital plane of five variants of uvular /R/ recorded from one native speaker of Tyrolean. The five groups of curves corresponds to five different manners of articulation: vocalized /R/, approximant, fricative, tap, and trill. To identify the significant differences between the five groups, the Interval-Wise Testing is performed on the data curves (tongue position), on their first derivative (tongue slope), and on their second derivative (tongue concavity). Interval-Wise Testing allows group comparisons in terms of adjusted p-value functions, which may result in a more informative and detailed representation of the regions of the tongue where a significant difference is located.

Keywords: Interval-Wise Testing; Permutation Methods; Functional Data Analysis; Phonetics

References

[1] Alessia Pini and Simone Vantini (2015). Interval-Wise Testing for Functional Data *MOX-report 30/2015* Politecnico di Milano.

[2] Vietti, Alessandro and Spreafico, Lorenzo (2015). An ultrasound study of the phonetic allophony of Tyrolean /R/, ICPhS 2015 Proceedings.

3.69 Joint sparse graphical models for brain imaging data with different coarseness levels

Eugen Pircalabelu[1], Gerda Claeskens[1] and Lourens J. Waldorp[2]

[1] KU Leuven, ORSTAT and Leuven Statistics Research Center
[2] University of Amsterdam, Department of Psychological Methods

Abstract: We propose a method for estimating brain networks from fMRI datasets that do not all contain measurements on the same set of regions. For certain datasets, some of the regions have been split in smaller subregions, while others have not been split. This gives rise to the framework of mixed scale measurements and the purpose is to estimate sparse undirected graphical models.

The resulting graphical models combine information from several subjects using the data available for all coarseness levels, overcome the problem of having data on different coarseness levels and take into account that dependencies exist between a coarse scale node and its finer scale nodes, since finer scale nodes are obtained by splitting coarser ones. Our procedure is directed towards estimating effects between split and unsplit regions, since this offers insight into whether a certain large ROI is constructed by aggregating homogeneous or heterogeneous parts of the brain. To overcome the problem of mixed coarseness levels, the expand and the reduce algebraic operators are used throughout the procedure. To estimate a sparse undirected graph, the alternating direction method of multipliers (ADMM) algorithm is used and to ensure similarity of graphs across coarseness levels, the procedure uses the fused graphical lasso and group graphical lasso penalties for certain block submatrices and a classical lasso penalty for the remaining submatrices. The method results in estimating graphical models for each subject and coarseness level in the analysis, referred to as within level edges, and identifies possible connections between a large region and its subregions, referred to as between level edges. We also investigate zooming-in and out procedures to assess the evolution of edges across the coarseness scales. The applicability of the method we propose goes beyond fMRI data, to other areas where data on different scales are observed and where the joint estimation of undirected graphs that resemble each other is desired. Moreover, the method avoids the tedious task of selecting one coarseness level for carrying out the analysis and produces interpretable results at all available levels. Empirical and theoretical evaluations illustrate the usefulness of the method.

Keywords: graphical models,fMRI

3.70 Minimax estimation of Hölder functions: a complexity analysis

P. Morkisz[1] and L. Plaskota[2,*]

[1] AGH University of Science and Technology, Krakow, Poland, morkiszp@agh.edu.pl
[2] University of Warsaw, Warsaw, Poland, leszekp@mimuw.edu.pl

Abstract: We study the minimax L^p estimation of d-variate regression functions $f : [0,1]^d \to \mathbb{R}$ that are r times differentiable and rth derivatives are Hölder continuous with exponent ϱ. The estimators use observations $y_i = f(x_i) + e_i$ with independent noise $e_i \sim \mathcal{N}(0, \sigma_i^2)$. Let $1 \le p < \infty$ and $s = r + \varrho$. It is known that for the n-point uniform (non-sequential) design and fixed variances σ^2 the minimax L^p error is proportional to $\max(n^{-s}, (\sigma^2/n)^{2/(2s+d)})$. The question is whether this error level can be beaten by using sequential (adaptive) observations with varying n and variances σ_i^2. This corresponds to the nonparametric heteroscedastic regression with sequential design. We show that the answer is negative. Specifically, assume that only such designs are allowed that for each f the expected value of $\sum_i \sigma_i^{-2}$ is at most N, where N is interpreted as the design cost. Then for any estimator using such a design the expected error is at least $CN^{-s/(2s+d)}$. In addition, a new version of a (non-local) piecewise polynomial estimator is constructed whose error achieves the lower bound. We conclude that the ε-complexity of the problem is proportional to $(1/\varepsilon)^{2+d/s}$.

Keywords: Nonparametric estimation; Hölder classes; Sequential design; Complexity

3.71 Higher Criticism - An Alternate Interpretation

T. Porter[1*] and M. Stewart[1]

[1] The University of Sydney, Australia; thomas.porter@sydney.edu.au, michael.stewart@sydney.edu.au

Abstract: The [2] Higher Criticism statistic (HC) and the [1] statistic (BJ) have been suggested as non-parametric tests of homogeneity for various two component mixtures; both possess a certain optimal lower-order asymptotic power property (attaining the "detection boundary" studied in [3] and [2]) which other popular tests such as the Kolmogorov-Smirnov test do not possess. This talk identifies an alternate interpretation for both HC and BJ in the context of mixture detection, the various implications of this disclosure and how some deficiencies of HC and BJ may be improved upon.

Keywords: Higher criticism; Berk-Jones; Sparse mixture detection

References

[1] Berk, R. Jones, D. (1979), Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Probability Theory and Related Fields*, **47(1)**, 47–59.

[2] Donoho, D. and Jin, J. (2004), Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, **32(3)**, 962–994.

[3] Ingster, Y. I. (1999), Minimax detection of a signal for ℓ_n^p-balls. *Mathematical Methods of Statistics*, **7**, 401–428.

3.72 A Monte Carlo evaluation of the performance of two new tests for symmetry

C. Pretorius[1,*] and J.S. Allison[1]

[1] North-West University, Potchefstroom, South Africa; charl.pretorius@nwu.ac.za, james.allison@nwu.ac.za

Abstract: We propose two new tests for symmetry based on well-known characterisations of symmetric distributions. The performance of the new tests is evaluated and compared to that of other existing tests by means of a Monte Carlo study. All tests are carried out in a regression setup where we test whether the error distribution in a linear regression model is symmetric. It is found that the newly proposed tests perform favourably compared to the other tests.

Keywords: Characterisation of symmetry; Empirical characteristic function; Goodness-of-fit; Symmetry

3.73 An Evaluation of Kolmogorov-Smirnov Style Statistics for Bivariate Circular Data

R. Quill[1,2,*], J.J. Sharples[1,2], L.A. Sidhu[1] and J. Piantadosi[3]

[1] School of Physical, Environmental and Mathematical Science, UNSW Canberra, Australia rachael.quill@student.adfa.edu.au
[2] Bushfire and Natural Hazards Cooperative Research Centre, Australia
[3] University of South Australia, Australia

Abstract: The sensitivity of Kolmogorov-Smirnov (K-S) style tests to changes in distribution structures, such as modal location, size or shape, is not well-known. In order to interpret distribution comparison results in terms of real-world changes, it is necessary to better understand these sensitivities. In this study, univariate and bivariate Normal distributions are simulated and manipulated for comparison using the K-S statistical test. Thresholds for significance testing are defined in terms of distribution structures, such as shifts in modal mean, density or spread, and thresholds for significance in bivariate distribution comparisons are analysed in relation to equivalent thresholds found in the univariate case. Similarly, thresholds for significance will be defined for Kuiper's test using simulated univariate wrapped Normal distributions. A further investigation into an extension of Kuiper's test to the bivariate case, equivalent to that of the extended K-S test, will be considered and evaluated through analysis of univariate and bivariate thresholds for significance testing.

In application, the directional response of prevailing winds to changes in the landscape at the surface are represented as joint circular distributions. To understand the impacts of physical features such as vegetation or topography on wind fields, K-S style tests can be used to compare wind response distributions observed across complex terrain. The above simulation study will not only allow the interpretation of such test results in terms of wind direction changes experienced on the ground, but will also evaluate the robustness of extended K-S style tests when applied to bivariate circular data.

Keywords: Kolmogorov-Smirnov; Kuiper's Test; Bivariate; Circular Data

3.74 Asymptotic properties of U-processes with convex loss and weighted Lasso penalty

W. Rejchel

Nicolaus Copernicus University, Toruń and University of Warsaw, Poland; wrejchel@gmail.com

Abstract: Model selection is an important challenge while working with data sets containing many predictors. In many practical problems (from genetics or biology) finding a (small) subset of relevant predictors is as important (or even more) as accurate estimation or prediction. There are many methods trying to solve this problem, for instance empirical risk minimization with the Lasso penalty [5]. In the talk I consider the case that the empirical risk is a U-statistic. Most of papers describing Lasso estimators are restricted to linear or generalized linear models. The use of U-processes (families of U-statistics) allows us to go beyond this scenario and study more complex models. For instance, [3] proposed a nonparametric approach to a wide class of regression problems. Another application is the pairwise ranking problem [1] that is popular in statistical learning. The main result describes properties of minimizers of U-processes with the Lasso. Namely, if the weights in the penalty are appropriately chosen, then the procedure is model selection consistent and asymptotically normal estimator of nonzero parameters. The example of such algorithm is the adaptive Lasso [6]. Convexity of the loss is the key regularity assumption. First, it allows us to avoid local minimum problems while minimizing U-processes with the penalty. Moreover, it strongly influences on argumentation used to prove results that is related to [4**?**]. Theoretical results are completed by experiments on simulated and real data sets.

Keywords: Weighted Lasso penalty; Model selection consistency; Oracle property; Ranking problem; U-process.

References

[1] Clémençon, S. and Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *Annals of Statistics*, **36**, 844–874.

[2] Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, **22**, 1993–2010.

[3] Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics*, **35**, 303–316.

[4] Niemiro, W. (1992). Asymptotics for M-estimators defined by convex minimization. *Annals of Statistics*, **20**, 1514–1533.

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Royal Statistical Society, Series B*, **58**, 267–288.

[6] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

3.75 Comparison of permutation methods in presence of nuisance variables

O. Renaud[1] and J. Frossard[1]

[1] Methodology and Data Analysis, Dept. of Psychology, University of Geneva; Olivier.Renaud@unige.ch, Jaromil.Frossard@unige.ch

Abstract: For linear models, permutation tests are well known when there is only one explanatory variable. In that simple case, responses can be permuted since data are exchangeable under the null hypothesis. Testing the effect of a given (explanatory) variable in presence of additional ones is more complex, and the literature contains many proposals using a permutation scheme, see e.g. Anderson and Ter Braak (2003), Kherad-Pajouh and Renaud (2010) and Winkler et al. (2014). We will first review them and put them under the same framework. It appears that many of the proposed methods can be viewed as a projection and we will discuss their geometrical interpretation. A large simulation study that compares all these methods is then presented. The effects of various deviations from the parametric assumptions are evaluated. We also present `lmpermutation`, a new `R` package that provides all the discussed methods. Often researchers in neurosciences want to compare signals – e.g. electroencephalogram (EEG) – at each time point in different conditions. Frequently, there are several experimental conditions, implying the presence of nuisance variables when testing one particular effect. In such cases, a multiple comparison strategy must be used. We will discuss several approaches and assess which permutation methods allow for a multivariate extension and can deal with a multiple comparison procedure. We will also present a simulation study that evaluates all the methods in various settings.

Keywords: Multiple comparison; Projection; Full and reduced Residuals; Signal

3.76 Nonparametric estimation of the hazard rate in a multiplicative censoring model

G. Chagny[1,*], F. Comte[2] and A. Roche[3]

[1] LMRS, Univ. Rouen; gaelle.chagny@univ-rouen.fr
[2] MAP5, Univ. Paris Descartes; fabienne.comte@parisdescartes.fr
[3] CEREMADE, Univ. Paris Dauphine; angelina.roche@dauphine.fr

Abstract: We address the problem of hazard rate estimation for the nonnegative random variable X of interest in the multiplicative censoring model $Y = XU$ [3], where U has a uniform distribution on $[0; 1]$, and is independent of X. Only a sample distributed like Y is observed, and thus recovering features of the distribution of X can be considered as an inverse problem. Although the problem of estimating the density f and the survival function F of the target variable in a nonparametric way has already been studied by several authors (**?** and references therein), we choose not to estimate the hazard $h = f/F$ by the ratio of two estimators. In the spirit of [2], a collection of estimator is defined, by minimizing an original regression-type contrast function, involving an estimator for the density of Y, over a collection of linear models. Nonasymptotic upper-bounds are proved for the quadratic risks of the estimates. The consequence is twofold: convergence rates are derived under smoothness assumptions on the function h, and the form of the bias-variance trade off should lead to a penalised model selection procedure - adaptive estimation in this setting is still a work in progress. Simulation experiments illustrate the method.

Keywords: Hazard rate; Minimum of contrast estimation; Multiplicative censoring model.

References

[1] Brunel, E. and Comte, F. and Genon-Catalot, V. (2015) Nonparametric density and survival function estimation in the multiplicative censoring model, *TEST*, to be published, preprint `https://hal.archives-ouvertes.fr/hal-01122847`.

[2] Plancade, S. (2011) Model selection for hazard rate estimation in presence of censoring, *Metrika* **74**, no. 3, 313–347.

[3] Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, **76**, no. 4, 751–761.

3.77 Excess risk concentration in least-squares regression with heteroscedastic noise

F. Navarro[1], A. Saumard[1,*]

[1] CREST, ENSAI, Campus de Ker-Lann, France; fabien.navarro@ensai.fr, adrien.saumard@ensai.fr

Abstract: The nineties and the beginning of the 21st century have witnessed the development of a nonasymptotic theory of statistical learning. More precisely, general upper bounds have been obtained for the excess risk of M-estimators (a.k.a. Empirical Risk Minimizers), at a fixed sample size and in a very general framework (see for instance [1] and references therein).
An exiting new direction in this line of research consists in deriving the concentration behavior of the excess risk of an M-estimator, with the aim of obtaining nonetheless sharp upper bounds but also lower bounds for the corresponding excess risk ([2], [3], [4]). In particular, all these concentration results are based on some representation formulas for the excess risk, the latter being identified as the maximizer of some drifted local suprema of the underlying empirical process.
We will present some generalizations of these representation formulas to any functional of a M-estimator and provide new concentration results in the context least-square regression with heteroscedastic and random design. Furthermore, we will derive optimal upper and lower bounds for the excess risk in this context.

Keywords: M-estimation; Regression; Excess risk; Concentration inequality; Empirical process.

References

[1] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems.* volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint- Flour.

[2] Saumard, A. (2012). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression *Electron. J. Statist.*, **6**, 579–655.

[3] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *The Annals of Statistics*, **42**(6), 2340–2381.

[4] van de Geer, S. and Wainwright, M. J. (2016). On concentration for (regularized) empirical risk minimization. preprint.

3.78 Catching-up, leapfrogging, and falling-back in economic growth — A nonparametric approach

H. Haupt[1], J. Schnurbus[1,*] and W. Semmler[2]

[1] University of Passau, Germany; harry.haupt@uni-passau.de, joachim.schnurbus@uni-passau.de
[2] Bielefeld University, Germany; New School, New York; semmlerw@newschool.edu

Abstract: Classical growth convergence regressions fail to account for various sources of heterogeneity and nonlinearity. Recent contributions advocating nonlinear dynamic factor models remedy those problems by allowing for club-specific convergence paths. Unfortunately and similar to statistical clustering methods, those results are sensitive to choices made in the clustering mechanism. In this paper we improve existing clubbing algorithms while providing an economic rationale for time-varying heterogeneity of number, size, and composition of convergence clubs. We propose a nonparametric strategy for tackling neglected heterogeneity and nonlinearity jointly while alleviating the problem of underspecification of growth convergence regressions. This strategy, which rests on club-specific transition paths derived from a nonlinear dynamic factor model, allows estimation of convergence effects on club and even country level. The proposed approach is illustrated using a current Penn World Table data set. We find empirical evidence for leapfrogging and falling-back of countries over time. Guise and degree of nonlinearities in convergence regressions also differ substantially over time. Furthermore, some countries on club-based convergence paths exhibit, in contrast to their fellow club members, insignificant convergence effects.

Keywords: Growth dynamics; Club convergence; Mixed kernel regression.

3.79 A local-polynomial Chow-type test

Michael Scholz[1,*]

[1] Department of Economics, University of Graz, Universitätsstraße 15/F4, 8010 Graz, Austria; Michael.Scholz@uni-graz.at

Abstract: We use the local-polynomial estimation framework to include categorical variables in nonparametric estimation. The aim of this exercise is to construct a nonparametric Chow-type test for the significance of the categorical predictors. We neither use discrete support kernels for the categorical covariates as, for example, [3], [4], or [1] nor apply a frequency-based (non-smoothing) nonparametric test of regression constancy over subsamples as, for example, [2]. Instead since we directly include the mix of continuous and categorical variables in our setting, we can apply all of the developed powerful tools for local-polynomial estimation in a straightforward way. We analyse the problem briefly from the theoretical perspective, study the finite sample behavior, and present as well an empirical illustration on time series data.

Keywords: Discrete variables; Nonparametric smoothing; Hypothesis testing; Local-polynomial estimation

References

[1] Hall, P., Li, Q., and Racine, J. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics* **89**, 784–789.

[2] Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics* **103**, 307–344.

[3] Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**, 99–130.

[4] Racine, J., Hart, J., and Li, Q. (2006). Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models. *Econometric Reviews* **25**, 523–544.

3.80 Bayesian nonparametric inference for the quantile function

C. Scricciolo*

Dpt of Economics, University of Verona, Polo Universitario Santa Marta, Italy; catia.scricciolo@univr.it

Abstract: A Bayesian nonparametric analysis of the quantile function is performed. First, the direct problem is considered. The posterior law of the quantile process corresponding to a prior for the population distribution function which is a normalized random measure with independent increments is shown to weakly converge to a Gaussian process and the result is applied to construct confidence bands for the quantile function. Secondly, Bayesian adaptive quantile estimation in deconvolution problems with unknown error distribution is considered. The objective is to estimate quantiles from indirect observations that are additively corrupted by unknown error measurements. Quantile estimation in deconvolution is an instance of nonlinear functional estimation in ill-posed inverse problems. We pursue the analysis for mildly ill-posed problems, namely, when the error distribution has characteristic function that decays polynomially fast. We propose a method to derive posterior contraction rates for the inverse problem from those of the direct problem. Credibility regions are constructed whose frequentist coverage is studied.

Keywords: Bernstein-von Mises theorem; Credibility regions; Quantile function estimation.

3.81 Automatic Signal Extraction for Stationary and Non-Stationary Time Series by Circulant SSA

J. Bógalo[1], P.Poncela[2,*] and E. Senra[1]

[1] Universidad de Alcalá, SPAIN; juan.bogalo@edu.uah.es, eva.senra@uah.es
[2] European Comission, Joint Research Centre (JRC), ITALY; pilar.poncela@jrc.ec.europa.eu

Abstract: Singular Spectrum Analysis (SSA) is a nonparametric, and therefore model free, tecnique for signal extraction in time series. However, it requires the intervention of the analyst to identify the frequencies associated to the underlying components. We propose an automatic version of SSA based on the properties of circulant matrices. We extend our new variant of SSA, Circulant SSA, to the nonstationary case. Through several sets of simulations, we show the good properties of our approach: it is fast, automatic and produces strongly separable elementary components by frequency. Finally, we apply Circulant SSA to the Industrial Production of six countries. We use it to deseasonalize the series and illustrate that it also reproduces a cycle in accordance to the dated recessions from the OECD.

Keywords: Circulant matrices, Signal extraction, Singular spectrum analysis, Time series, Toeplitz matrices.

3.82 Advances in Quantile Regression for Vector Responses

P. Boček[1], M. Hallin[2], D. Hlubinka[3], Z. Lu[4], D. Paindaveine[2], M. Šiman[1,*]

[1] The Institute of Information Theory and Automation of the Czech Academy of Sciences, Czech Republic; bocek@utia.cas.cz, siman@utia.cas.cz
[2] ECARES, Université Libre de Bruxelles Belgium; dpaindav@ulb.ac.be, mhallin@ulb.ac.be
[3] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic; hlubinka@karlin.mff.cuni.cz
[4] School of Mathematical Sciences, University of Southampton, UK; Z.Lu@soton.ac.uk

Abstract: Ideal statistical regression methods should be able to provide maximum information under minimal assumptions. Quantile regression for scalar responses comes close to the ideal because it can disclose, model, and predict all the output conditional probability distributions. Its extension for multivariate responses should be therefore highly welcome and suitable for handling complex real-life relations and dependencies. Two types of such generalizations, using convex polyhedral shapes (see [1], [2] and [3]) or ellipsoids (see [5] and [4]) for all the conditional quantile regions, are presented, summarized, illustrated, and compared here, and the very recent progress in their research is highlighted, including their various trend, robust and nonparametric modifications and software implementation.

Keywords: Quantile regression; Multiple-output regression; Multivariate quantile; Data depth; Regression depth

References

[1] Hallin, M. and Paindaveine, D. and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Annals of Statistics*, **38**, 635–669.

[2] Paindaveine, D. and Šiman, M. (2011). On directional multiple-output quantile regression. *Journal of Multivariate Analysis*, **102**, 193–212.

[3] Hallin, M. and Lu, Z. and Paindaveine, D. and Šiman, M. (2015). Local bilinear multiple-output quantile/depth regression. *Bernoulli*, **21**, 1435–1466.

[4] Hallin, M. and Šiman, M. (2016). Elliptical multiple-output quantile regression and convex optimization. *Statistics & Probability Letters*, **109**, 232–237.

[5] Hlubinka, D. and Šiman, M. (2016). On parametric elliptical regression quantiles. Submitted.

3.83 Identification of causal effects in recursive settable and its testability

J. Smits[1] and G. Mellace[2]

[1] Department of Humanities, Social and Political Sciences, ETH Zurich; smits@nadel.ethz.ch,
[2] Department of Business and Economics, University of Southern Denmark

Abstract: This article makes five separate but related contributions. First, it is shown by various examples that identified parameters (under standard assumptions) in the econometrics literature, therein often referred to as "causal effects", can in fact not be given a causal interpretation. This concerns both the structural equations and the potential outcome framework literatures. It is argued that the misinterpetation stems from intermingling of statistical objects and statistical inference with causal nomenclature, without causality and causal effects having been properly defined mathematically. Second, the paper redefines various treatment effect parameters within the Settable System (SS) framework by [1], an extension and refinement of the Pearl Causal Model (PCM), which we argue to be the most suitable language for this purpose. Third, we state and prove sufficient conditions for certain treatment effect parameters to be given a causal interpretation. These conditions consist of the standard, stochastic assumptions, with additional causal assumptions. Fourth, we derive testable implications of these causal assumptions in the case of instrumental variables estimation and propose a statistical test. Fifth and finally, we present Monte Carlo result on the finite sample properties of this test and apply it to an empirical example related to the returns to college education in the US.

Keywords: Causal inference; Settable systems; Instrumental variables.

References

[1] Chalak, K. and White, H. (2012). Causality, conditional independence, and graphical separation in settable systems. *Neural Computation*, **24(7)**, 1611–1668.

3.84 Lasso-type estimators for non-parametric mixed-effects models: application to high-dimensional data from a vaccine clinical trial for HIV

P. Soret[1,2,3,4,*], C. Meza[6], M. Avalos[1,2,3], K. Bertin[6] and R. Thiébaut[1,2,3,4,5]

[1] Univ. Bordeaux, ISPED, F–33000 Bordeaux, France
[2] INRIA SISTM Bordeaux – Sud–Ouest, F–33405 Talence, France
[3] INSERM, Centre INSERM U1219–Epidémiologie–Biostatistique, F–33000 Bordeaux, France
[4] Vaccine Research Institute (VRI), F–94000 Créteil, France [5] CHU de Bordeaux, F-33000 Bordeaux, France
[6] CIMFAV–Facultad de Ingeniería, Univ. de Valparaíso, Valparaíso, Chile * perrine.soret@isped.fr

Abstract: The penalization of likelihoods by L1–norms has become a relatively standard technique for high-dimensional data when the assumed models are based on n independent and identically distributed observations. These techniques should improve prediction accuracy (since regularization leads to variance reduction) together with interpretability (since sparsity identifies a subset of variables with strong effects). Computationally, these penalties are attractive and their theoretical properties have been intensively studied during the last years.

Several authors have recently suggested analyzing high-dimensional clustered or longitudinal data using L1–penalization methods in mixed effects models. These approaches are mostly developed for variable selection purposes in linear and generalized linear mixed effects models and also, but less extensive, in parametric nonlinear mixed effects models.
Only a few works have considered the problem of selecting nonlinear functions using L1–penalization methods in non-parametric mixed effects models, with additive or nonadditive predictors. Nonlinear functions are approximated by a linear combination of smooth functions (spline, wavelet or Fourier basis functions) possibly combined with more irregular functions (spiky basis functions). The resulting estimator depends only on a relatively small number of variables and/or a relatively small number of basis functions [1].
In this study we illustrate the interest of such approaches in the analysis of the DALIA-1 longitudinal trial [2]. Eighteen HIV infected patients received vaccine injections at weeks 0, 4, 8 and 12. Antiretroviral treatment was interrupted at week 24. The patients were followed up to week 48, leading to 14 repeated measures per subject. Our aim was to predict the evolution of viral loads (continuous response) from the about 260 gene sets (predictors). The incorporation of the temporal effect is a key point to reach accurate predictions.

Keywords: genomics; longitudinal data; complex data; machine learning

References

[1] Arribas-Gil, A. and Bertin, K. and Meza, C. and Rivoirard, V. (2012). *LASSO-type estimators for Semiparametric Nonlinar Mixed-Effects Models Estimation*, Statistics and Computing, **24 (3)**, 443-460.

[2] Lévy, Y. and Thiébaut, R. and Montes, M. and Lacabaratz, C. and Sloan, L. and King, B. et al. (2014). *Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load.* European journal of immunology, **44 (9)**, 2802-2810.

3.85 Unhappy with semi-parametrics? A replication study of a life-satisfaction analysis

Setareh Ranjbar and Stefan Sperlich

Geneva School of Economics and Management, University of Geneva

Abstract: Semi-parametric techniques were introduced as a way of circumventing the curse of dimensionality but providing some sort of functional form flexibility. In econometrics they became popular because of their ability to incorporate prior knowledge. The elements for which the economic theory is not specific about, can be passed in good conscience to the non-parametric part. Although the original idea was to keep the theoretically unspecified part non-parametric but model the parts where economic theory does guide the empirical researcher, the tradition in practise was that researchers tended to model parametrically the parts they wanted to interpret. More recent is the trend to let unspecified exactly the part of the model that is the part of interest. For example, in studies of life-satisfaction researchers are particularly interested in studying the impact of age whose correct (parametric) specification has therefore been in the centre of controversy. Consequently, a seemingly attractive remedy was to estimate the impact of age non-parametrically, while keeping the rest of the model parametrically. Having discrete responses y_i, one tries to explain these by some observed individual characteristics $x_i \in R^d$ and age_i via a GLM

$$E[Y|x, age] = G(\eta(x, age)) = G(\beta_0 + x'\beta_1 + age\beta_a) \ .$$

In a correctly specified model the link function corresponds to the error distribution in the latent variable model. However, often G is set equal to identity for convenience by offering various justifications for this choice. When Y can take considerably more than just two values, then a necessary prerequisite for making this a reasonable choice is that either Y is a cardinal variable or the index function $\eta(\cdot)$ is sufficiently flexible. The semi-parametric model

$$E[Y|x, age] = G(\eta(x, age)) = x'\beta_1 + m(age) \tag{2}$$

could deem to be an enticing compromise. This was done quite recently for example in different studies on the overall life-satisfaction (LS). When they estimated model (1) they found a clear indication for the mid-life crises at the age of around 50. We will show, that this hollow of $m(\cdot)$ at age 50 can equally well be caused by either a natural but non-linear link $G(\cdot)$ and non-linearities of the impact of x, respectively. By "natural" we mean that people's responses regress to the mean, i.e. they tend not to give the most extreme values – who wants to rank himself as totally unsatisfied or totally satisfied? This simple shift toward unequal scaling can be captured by a proper choice of $G(\cdot)$, ignoring it can result in a cubic shaped estimate of $m(\cdot)$.

Keywords: GLM modeling, semi-parametrical modeling

3.86 Minimax-Optimal Distribution Regression

Z. Szabó[1,*], B. Sriperumbudur[2], B. Póczos[3] and A. Gretton[1]

[1] Gatsby Unit, University College London; zoltan.szabo@gatsby.ucl.ac.uk, arthur.gretton@gmail.com
[2] Department of Statistics, Pennsylvania State University; bks18@psu.edu
[3] Machine Learning Department, Carnegie Mellon University; bapoczos@cs.cmu.edu

Abstract: We focus on the distribution regression problem (DRP): we regress from probability measures to Hilbert-space valued outputs, where the input distributions are only available through samples (this is the 'two-stage sampled' setting). Several important statistical and machine learning problems can be phrased within this framework including point estimation tasks without analytical solution (such as hyperparameter or entropy estimation) and multi-instance learning. However, due to the two-stage sampled nature of the problem, the theoretical analysis becomes quite challenging: to the best of our knowledge the only existing method with performance guarantees to solve the DRP task requires density estimation (which often performs poorly in practise) and the distributions to be defined on a compact Euclidean domain. We present a simple, analytically tractable alternative to solve the DRP task: we embed the distributions to a reproducing kernel Hilbert space and perform ridge regression from the embedded distributions to the outputs. Our main contribution is to prove that this scheme is consistent in the two-stage sampled setup under mild conditions (on separable topological domains enriched with kernels): we present an exact computational-statistical efficiency tradeoff analysis showing that the studied estimator is able to match the *one-stage* sampled minimax-optimal rate. This result answers a 17-year-old open question, by establishing the consistency of the classical set kernel [**?** **?**] in regression. We also cover consistency for more recent kernels on distributions, including those due to [**?**]. The practical efficiency of the studied technique is illustrated in supervised entropy learning and aerosol prediction using multispectral satellite images.

Keywords: Two-Stage Sampled Distribution Regression; Kernel Ridge Regression; Mean Embedding; Multi-Instance Learning; Minimax Optimality

3.87 Asymptotic confidence bands in the Wicksell's problem

J. Wojdyła[1] and Z. Szkutnik[1,*]

[1] Faculty of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland;
jwojdyla@agh.edu.pl, szkutnik@agh.edu.pl

Abstract: Work on construction of confidence bands for a density function of directly observed i.i.d. data started with the seminal paper by [1]. In the last decade, asymptotic nonparametric confidence bands have been constructed in some inverse problems, like density deconvolution, inverse regression with a convolution operator and regression with errors in variables, with the pioneering contribution of [2]. There seems to be, however, no such construction for practically important inverse problems of stereology. We partially fill this gap by constructing a kernel-type estimator for the density of squared radii in the stereological Wicksell's problem, along with corresponding asymptotic uniform confidence bands and an automatic bandwidth selection method. Let us recall that the Wicksell's problem of stereology consists in unfolding the distribution of random radii of balls randomly placed in an opaque medium and only observed as circles on a random plane slice through the medium, and that the density of the observed circles radii is related to the density of balls radii via an Abel integral equation. Following [1] and [2], we construct asymptotic confidence bands that are based on strong approximations and on a limit theorem for the supremum of a stationary Gaussian process. The performance of the new procedures is also investigated in a simulation experiment and demonstrated with some real astronomical data related to M62 globular cluster and obtained in an observation process described with a very similar Abel equation [3].

Keywords: Abel integral equation; Ill–posed inverse problem; Kernel methods; Nonparametric curve estimation; Stereology.

References

[1] Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

[2] Bissantz, N., Dümbgen, L., Holzmann, H. and Munk, A. (2007). Non-parametric confidence bands in deconvolution density estimation. *J. Roy. Statist. Soc. B*, **69**, 483–506.

[3] Sen, B. and Woodroofe, M. (2012). Bootstrap confidence intervals for isotonic estimators in a stereological problem. *Bernoulli*, **18**, 1249–1266.

3.88 A Robustly Adjusted Boxplot for Functional Data

N. Tarabelloni[1,*], F. Ieva[2]

[1] MOX – Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano; nicholas.tarabelloni@polimi.it
[2] Department of Mathematics "F. Enriques", Università degli Studi di Milano; francesca.ieva@unimi.it

Abstract: The problem of outlier detection in high dimensional settings is a crucial point for a number of statistical analyses. If not properly identified, outliers may lead to model misspecification, biased parameter estimation and incorrect results, especially in those contexts where the number of available statistical units is lower than the number of parameters (for example, Functional Data Analysis). In this paper we introduce a robust, adjusted version of the functional boxplot, which is the most common tool adopted to perform outlier detection in Functional Data Analysis. It is based on the use of statistical depth measures, a convenient, nonparametric tool in the analysis of functional data. A crucial element of the functional boxplot is the inflation factor of the fences, controlling the proportion of observations flagged as outlier. After an overview of the methods and tools currently available in the literature, we will describe a robust and nonparametric method to compute a data-driven value for such inflation factor. In doing so, we will make use of robust estimators of variance-covariance operators and of the corresponding eigenvalues and eigenfunctions. Two simulation studies are proposed to give direct insights into the use of the proposed functional boxplot, and to test both the robustness and accuracy of robust variance-covariance estimators, together with the performances of the functional boxplot in recognising truly outlying observations. R codes are freely available upon request.

Keywords: Functional Data Analysis; Outlier Detection; Functional Depths; Functional Boxplot; Robust Estimators.

3.89 Polymer creep deformation estimates using a flexible approach based on Time/Temperature Superposition principle

A. Meneses[1], J. Tarrío-Saavedra[2,*], S. Naya[2], C. Gracia-Fernández[3] and J. López-Beceiro[2]

[1] Universidad Nacional de Chimborazo (Ecuador); antoniomenesesfreire@hotmail.com
[2] Escuela Politécnica Superior, Universidade da Coruña (Spain); jtarrio@udc.es, salva@udc.es, jorge.lopez.beceiro@udc.es
[3] TA Instruments; CGracia@tainstruments.com

Abstract: The aim of this work is to provide a flexible methodology based on semiparametric statistics to estimate the strain in materials due to creep. Creep phenomenon is the time dependent process by means materials are continuously deformed when a constant stress is applied. It takes place at room temperature in viscoelastic materials such as polymers. In this study, the creep behavior of two polymers based on epoxy resin and glass fiber has been obtained in a dynamic mechanical analyzer as the compliance (strain divided by the applied stress) depending on time. Due to the nature of the studied materials (amorphous), the time/temperature superposition (TTS) physical principle can be applied. The procedure can be summarized in the following steps: (1) Creep compliance curves are obtained $J(t)$. They are evaluated at different temperatures. (2) Each J(t) curve is fitted using a B-spline model by the R TTS package to obtain smoot estimates. (3) Numerical differentiation is performed. (5) The reference temperature at which estimates out of the experimental time range is chosen: master curve. (6) Horizontal and vertical shifts of curves are obtained by shifting the compliance curve derivatives. The L1 distance is assumed. The master curve is obtained by an iterative process. (7) The resulting master curve is smoothed using b-splines basis. This procedure can provide softer and more accurate estimates of material viscoelastic properties.

Keywords: Polymer science; Dynamic Mechanical Analysis; Time/Temperature Superposition; Spline regression; B-splines.

References

[1] Naya, S., Meneses, A., Tarrío-Saavedra, J., Artiaga, R., López-Beceiro, J., Gracia-Fernández, C. (2013). New method for estimating shift factors in time–temperature superposition models. *Journal of thermal analysis and calorimetry*, **113**, 453–460.

[2] Meneses, A., Naya, S., Tarrío-Saavedra, J. (2015). TTS: Master Curve Estimates Corresponding to Time-Temperature Superposition. R package version 1.0. http://CRAN.R-project.org/package=TTS

3.90 On the selection of window length in Singular Spectrum Analysis of a time series.

Poornima Unnikrishnan[1]and V. Jothiprakash[1]

[1] Indian Institute of Technology,Bombay,India; poornimaunni@iitb.ac.in, vprakash@iitb.ac.in

Abstract: Time series analysis and modelling has gained a lot of attention among various fields of research in the recent past. Though, stochastic models such as ARIMA, ARMA etc and data mining techniques such as ANN, Genetic programming etc are commonly being used for real-world time series analysis, both are having their own limitations. Hence, there is a need for a better time series analysis technique that will in-cooperate the advantages of conventional modelling techniques and exclude their disadvantages. Singular Spectrum Analysis (SSA) is a promising non parametric time series modelling technique that has proved to be successful in data pre-processing in diverse application fields (de Menezes et al., 2014; Elsner and Tsonis, 1996; Golyandina and Zhigljavsky, 2013; Rodrıguez-Aragón and Zhigljavsky, 2010). It is based on multivariate statistics and works within the internal structure of the time series. SSA involves decomposition of the time series into various components based on a preassigned window length and elimination of noise by reconstructing the required time series using significant components. Singular Value Decomposition (SVD) of the trajectory matrix of time series is the core stage of SSA in which the number and form of decomposed components depend on the window length utilized for decomposition. Thus, appropriate selection of window length is critical in effective time series modelling. In this study, we have presented the method of SSA in time series analysis in detail and a sensitivity analysis of window length is carried out based on a daily rainfall time series.

Keywords: Singular Spectrum Analysis; Window length; Singular Value Decomposition; Sensitivity analysis

References

[1] de Menezes M.L and Souza R.C and Pessanha J.F.M (2014). Combining Singular Spectrum Analysis and Par(p) Structures to Model Wind Speed Time Series. *J. Syst. Sci. Complex.* **27**, 29–46. doi:10.1007/s11424-014-3301-8

[2] Elsner J.B. and Tsonis A.A (1996). *Singular Spectrum Analysis - A New Tool in Time Series Analysis.* Plenum press, Newyork.

[3] Golyandina N. and Zhigljavsky A (2013). *Singular Spectrum Analysis for Time Series.* Springer, London,Newyork. doi:10.1007/978-3-642-34913-3

[4] Rodrıguez-Aragón L.and Zhigljavsky A.(2010). Singular spectrum analysis for image processing. *Stat. Interface* **3**, 419–426.

3.91 Joint diagonalisation of scatter operators: Hilbertian Fourth Order Blind Identification and other applications

Germain Van Bever[1], B. Li[2], H. Oja[3], R. Sabolova[1] & F. Critchley[1]

[1]MCT Faculty, The Open University, Milton Keynes; germain.van-bever@open.ac.uk, radka.sabolova@open.ac.uk, f.critchley@open.ac.uk
[2]Penn State University; bxl9@psu.edu
[3]Turku University; hannu.oja@utu.fi

Abstract: With the increase in measurement precision, functional data is becoming common practice. Relatively few techniques for analysing such data have been developed, however, and a first step often consists in reducing the dimension via Functional PCA, which amounts to diagonalising the covariance operator. Joint diagonalisation of a *couple* of scatter functionals has proved useful in many different setups, such has Independent Component Analysis (ICA), Invariant Coordinate Selection (ICS), etc. The main part of this talk consists in extending the Fourth Order Blind Identification procedure to the case of data on a separable Hilbert space. In the finite-dimensional setup, this procedure provides a matrix Γ such that ΓX has independent components, if one assumes that the random vector X satisfies $X = \Psi Z$, where Z has independent marginals and Ψ is an invertible mixing matrix. When dealing with distributions on Hilbert spaces, two major problems arise: (i) the notion of "marginals" is not naturally defined and (ii) the covariance operator is, in general, non invertible. These limitations are tackled by reformulating the problem in a coordinate-free manner and by imposing natural restrictions on the mixing model. The proposed procedure is shown to be Fisher consistent and affine invariant. A sample estimator is provided and illustrated on simulated and real datasets. Using the joint diagonalisation

of different scatter operators, we will hint the possible extensions to functional ICS, as well as, in the functional time series context, functional Second Order Blind Identification. This work is supported by the EPSRC grant EP/L010429/1.

Keywords: Invariant Coordinate Selection; Functional Data; Symmetric Component Analysis; Independent Component Analysis.

References

[1] J.-F. Cardoso, Source Separation Using Higher Moments *Proceedings of IEEE international conference on acoustics, speech and signal processing* 2109-2112.

[2] D. Tyler, F. Critchley, L. Dumbgen and H. Oja, Invariant Co-ordinate Selection *J. R. Statist. Soc. B.*, 2009, **71**, 549–592.

[3] J. Ramsay and B.W. Silverman *Functional Data Analysis* 2nd edn. Springer, New York, 2006.

[4] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, A blind source separation technique using second-order statistics, *IEEE Trans. Signal Processing*, vol. 45, pp. 434–444, 1997.

3.92 Locally Stationary Functional Time Series

A. van Delft[1] and M. Eichler[1]

[1] Maastricht University; a.vandelft@maastrichtuniversity.nl, m.eichler@maastrichtuniversity.nl

Abstract: Inference methods for functional data have received a lot of attention the last few years. So far, the literature on functional time series has focused on processes of which the probabilistic law is either constant over time or constant up to its second-order structure. Especially for long stretches of data it is desirable to be able to weaken this assumption. This paper introduces a framework that allows for meaningful statistical inference of functional data of which the dynamics change over time. That is, we extend the locally stationary setting as developed by [1] to functional time series and establish a class of processes that have a functional time-varying spectral representation. We introduce the notion of a time-varying spectral density operator and derive its uniqueness. Time-varying functional ARMA processes are investigated and shown to be locally stationary according to our definition. Additionally, we demonstrate that based on a functional version of the segmented periodogram, the time-varying spectral density operator can be consistently estimated and study its asymptotic law. In particular, a functional central limit theorem is derived.

Keywords: Functional data analysis; Locally stationary processes; Spectral analysis.

References

[1] Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications*, **62**, 139-168.

3.93 Independent component analysis for tensor-valued data

J. Virta[1,*], B. Li[2], K. Nordhausen[1] and H. Oja[1]

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

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

Keywords: FOBI; Kronecker structure; Matrix-valued data; Multilinear algebra

3.94 Using history matching for prior choice

David J. Nott[1,*] and Xueou. Wang[2]

[1] National University of Singapore; standj@nus.edu.sg
[2] National University of Singapore; a0095911@u.nus.edu

Abstract: It can be important in Bayesian analyses of complex models to construct informative prior distributions which reflect knowledge external to the data at hand. Nevertheless, how much prior information an analyst is able to use in constructing a prior distribution will be limited for practical reasons, with checks for model adequacy and prior-data conflict an essential part of the justification for the finally chosen prior and model. This paper develops computationally efficient numerical methods for exploring reasonable choices of a prior distribution from a parametric class, when prior information is specified in the form of some limited constraints on prior predictive distributions, and where these prior predictive distributions are analytically intractable. The methods developed may be thought of as a novel application of the ideas of history matching, a technique developed in the literature on assessment of computer models. We illustrate the approach in the context of logistic regression and sparse signal shrinkage prior distributions for high-dimensional linear models.

Keywords: Approximate Bayesian computation, Bayesian inference, Device of imaginary results, History matching, Prior elicitation.

3.95 Bayesian variable selection for mixed effect models with nonignorable dropout

Mingan Yang

Graduate School of Public Health, San Diego State University, San Diego, California, U.S.A.; mingany@yahoo.com

Abstract: In this article, we use Bayesian nonparametric approach for joint variable selection of both fixed and random effects in presence of nonignorable dropout. We integrate both shrinkage prior, which is expressed as a scale mixture of normal distributions, and the mixed G-prior in our approach. By this way, we greatly improve efficiency and accuracy. In particular, we show that our approach greatly decreases bias in estimation and selection for nonignorable missing data. A stochastic search Gibbs sampler is implemented for variable selection. We further illustrate the proposed approach using simulated data and a real example.

Keywords: Variable selection; mixed effects model; Bayesian model selection; Stochastic search.

3.96 Lag Selection and Model Validation in Nonparametric Autoregressive Conditional Heteroscedastic Models

A. Z. Zambom[1,*] and S. Kim[2]

[1] Loyola University Chicago; azambom@luc.edu
[2] Miami University Ohio; kim20@miamioh.edu

Abstract: In several time series data sets, simple linear models may not adequately fit the data. However, many of the popular alternative, such as nonlinear time series models, use pre-specified parametric functions based on previous knowledge from specific applications. Thus, the aim of this paper is two-fold: for a nonparametric autoregressive conditional heteroscedastic model, to propose a consistent lag selection procedure; and given a parametric form specified a priori, to develop a model validation test. Both methods are based on a new powerful test statistic developed to check the influence of a lag on the conditional mean function. The asymptotic properties of this test statistic are investigated under the null and local alternative hypotheses. Extensive simulation studies suggest that the proposed lag selection method and model validation procedure outperform existing methods for nonlinear models.

Keywords: ANOVA, Conditional heteroscedastic model, Nonlinear model, Nonparametric lag selection, Nonparametric model validation.

3.97 A Scale-invariant L^2-norm Based Two-sample Test for High-dimensional Data

J. Zhang[1,*] and L. Zhang[2]

[1] Department of Statistics and Applied Probability; stazjt@nus.edu.sg
[2] Department of Statistics and Applied Probability; liangz@u.nus.edu

Abstract: In this work, we propose and study a scale-invariant L^2-norm based test for two-sample high-dimensional problems where there are fewer observations than the dimension so that the classical Hotelling two-sample test is no longer applicable. For two-sample high-dimensional problems, several tests have been proposed in the literature and they all approximate the null distributions of their test statistics by a normal distribution although the associated null distributions may not be asymptotically normal. To overcome this problem, we propose to approximate the null distribution of our test statistic using the well-known Welch-Satterthwaite χ^2-approximation. It turns out that our test is adaptive to the shape of the underlying null distribution of the proposed test and has a good size controlling. Simple methods are given for estimating the parameters of the approximation distribution ratio-consistently. The approximate and asymptotic powers of the proposed test are investigated. Simulation studies and real data applications show that the proposed test has a better size controlling than one of the existing tests while their powers are comparable unless their sizes are not comparable.

Keywords: High-dimensional data; Scale-invariant L^2-norm based test; χ^2-type mixtures; Welch-Satterthwaite χ^2-approximation.

3.98 Alternatives for Ghosal-Ghosh-Vaart priors

Bas Kleijn[1] and Yanyun Zhao[2,*]

[1] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Netherlands; B.Kleijn@uva.nl
[2] Wenlan School of Business, Zhongnan University of Economics and Law, China; yyunzhao@gmail.com

Abstract: The study of the asymptotic frequentist behaviors of the posterior distribution usually includes sufficiency of prior mass in sharpened Kullback-Leibler neighbourhoods described in Ghosal et al. (2000). In this article, we try to relax this condition to accommodate a wide range of priors. To that end, we formulate an alternative rates-of-posterior-convergence theorem, based on the approach proposed in Kleijn (2015). The aim is to strengthen model conditions and gain flexibility in the choice for a prior, while maintaining optimality of the posterior rate of convergence. Additionally, the general results are illustrated through the examples in the areas of nonparametric estimation, semiparametric inference and survival analysis.

Keywords: Posterior convergence rate; Kullback-Leibler neighbourhoods; Ghosal-Ghosh-Vaart priors; Support boundary estimation.

References

[1] Ghosal, S. ,Ghosal, J. and van der Vaart, A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, **28**, 500–531.

[2] Kleijn, B. (2005). Criteria for Posterior consistency, (submitted for publication).

3.99 Probabilities of misclassification given by a L^2-norm-based classifier for functional data

Tianming Zhu[1,*] and Jin-Ting Zhang[2]

[1] National University of Singapore; zhu_tianming@u.nus.edu
[2] National University of Singapore; stazjt@nus.edu.sg

Abstract: In recent years, many methods dealing with functional data classification have been developed rapidly. However, few works focus on studying the probabilities of misclassification. We consider a L^2-norm-based classifier for functional data. We first study the probabilities of misclassification for the two known populations with equal covariance case. We show that when the populations are Gaussian processes, the distribution of the criterion is normal distribution.

We further extend this method to the case when parameters are unknown. However, in this case, the distribution of the criterion is extremely complicated. Therefore, we provide the asymptotic distribution of the criterion and the asymptotic expansion of the probabilities of misclassification to term of order n^{-2}. Finally, we consider a more general case, that is, the covariance operators of two populations are different. The performance of our classifier is further demonstrated with a simulation and a real data example.

Keywords: Functional data, Supervised classification, L^2-norm, Misclassification error, Asymptotic distribution.

Index