BOOK OF ABSTRACTS

June, 12th - 16th, 2014
Cádiz, Spain

Edited by, Universidad Carlos III & The International Society for NonParametric Statistics (ISNPS)
Depósito Legal: M-16734-2014
ISBN: 978-84-89315-69-3

Welcome

Following the successful First Conference of the International Society for NonParametric Statistics, held in Greece in 2012, we have organized the Second ISNPS Conference in Cádiz, Spain.

The ISNPS was founded in 2010 "to foster the research and practice of nonparametric statistics, and to promote the dissemination of new developments in the field via conferences, books and journal publications." ISNPS has a distinguished Advisory Committee that includes R.Beran, P. Bickel, R. Carroll, D. Cook, P. Hall, R. Johnson, B. Lindsay, E. Parzen, P. Robinson, M. Rosenblatt, G. Roussas, T. SubbaRao, and G. Wahba. The Charting Committee of ISNPS consists of over fifty prominent researchers from all over the world.

The nature of ISNPS is uniquely global, and its international conferences are designed to facilitate the exchange of ideas and recent advances among researchers from all around the world. These conferences are organized in cooperation with established statistical societies, such as the American statistical Association (ASA), the Institute of Mathematical Statistics (IMS) and the International Statistical Institute (ISI).

The aim of this Conference is to put together recent advances and trends in several areas of nonparametric statistics in order to facilitate the exchange of research ideas, promote collaboration among researchers from all over the world and contribute to the further development of the field.

We thank all the institutions and organizations that have sponsored the meeting. We also thank all of you for attending the Conference. We hope you will have an enjoyable and fruitful stay.

Ricardo Cao, Wenceslao González-Manteiga and Juan Romo
Co-Chairs of the Second ISNPS Conference

Sponsors

ASA – American Statistical Association
IMS – Institute of Mathematical Statistics
BERNOULLI – Society for Mathematical Statistics and Probability
JNPS – Journal of Nonparametric Statistics
UC3M - Universidad Carlos III de Madrid

Committees

Executive Committee

Michael Akritas
Soumen Lahiri
Dimitris Politis

Co-Chairs

Ricardo Cao
Wenceslao González-Manteiga
Juan Romo

Advisory Committee

Rudy Beran
Peter Bickel
Ray Carroll
Dennis Cook
Peter Hall
Richard Johnson
Bruce Lindsay
Emanuel Parzen
Peter Robinson
Murray Rosenblatt
George Roussas
Tata SubbaRao
Grace Wahba

Local Committee

Germán Aneiros Pérez
Ana Arribas Gil
Mercedes Conde Amboage
Pedro Delicado Useros
Mario Francisco Fernández
Eduardo García Portugués
Áurea Grané Chávez
Rosa E. Lillo Rodríguez
Salvador Naya Fernández
Beatriz Pateiro López
Ewa Strzalkowska-Kominiak

Charting Committee

Karim Abadir
Moulinath Banerjee
Rainer Dahlhaus
Jan deLeeuw
Piotr Fryzlewicz
Wenceslao González-Manteiga
Trevor Hastie
Nils Hjort
Claudia Klüppelberg
Hira Koul
Jacek Leskow
Regina Liu
Natalia Markovic
Simos Meintanis
Axel Munk
Stathis Paparoditi
Joseph Romano
Simon Sheather
Dag Tjostheim
Slava Vasiliev
Wei Biao Wu
Chunming Zhang

Ian Abramson
Patrice Bertail
Anirban DasGupta
Frederic Ferraty
Subhashis Ghoshal
Marc Hallin
Xuming He
Joel Horowitz
Piotr Kokoszka
Jens-Peter Kreiss
Bing Li
Gabor Lugosi
Eric Matzner-Lober
Karl Mosler
Dan Nordman
Jeff Racine
Juan Romo
Winfried Stute
Alexandre Tsybakov
Philippe Vieu
Qiwei Yao

Anestis Antoniadis
Ricardo Cao
Aurore Delaigle
Jurgen Franke
Irene Gijbels
Jeff Hart
Nancy Heckman
Marie Huskova
Michael Kosorok
Michele La Rocca
Runze Li
Enno Mammen
George McCabe
Hans-Georg Müller
Victor Panaretos
Alfredas Rackauskas
Theofanis Sapatinas
Robert Taylor
Ingrid Van Keilegom
Michael Wolf
Bin Yu

Contents

Abstracts 11

Authors Index... 229

Abstracts

1.01

The Lasso: a brief review and a new significance test

Robert Tibshirani[1], Richard Lockhart[2], Jonathan Taylor[3] and Ryan Tibshirani[4]

[1] *Stanford University*
[2] *Simon Fraser University*
[3] *Stanford University*
[4] *Carnegie Mellon University*

Abstract. *I will review the lasso method and show an example of its utility in cancer diagnosis via mass spectometry. Then I will consider testing the significance of the terms in a fitted regression, fit via the lasso. I will present a novel test statistic for this problem, and show that it has a simple asymptotic null distribution. This work builds on the least angle regression approach for fitting the lasso, and the notion of degrees of freedom for adaptive models (Efron 1986) and for the lasso (Efron et. al 2004, Zou et al 2007). We give examples of this procedure, discuss extensions to generalized linear models and the Cox model, and describe an R language packagefor its computation.*

1.02

From multivariate depth to functional depth

P. Rousseeuw[1,*], M. Hubert[1] and P. Segaert[1]

[1] *Mathematics Department, KU Leuven, Celestijnenlaan 200B, BE-3001 Heverlee, Belgium*
[*] *Corresponding author, peter@rousseeuw.net*

Abstract. *For univariate data, depth is a symmetrized version of ranking. In multivariate data, the halfspace depth of a point is its smallest univariate depth in any projection. There exist many other depth functions for multivariate data, all of which rank the data from the outside inward. Much work went into devising efficient algorithms for computing depth values, depth contours and depth medians (see e.g. Rousseeuw and Struyf 2004). This made it possible to construct the bagplot (Rousseeuw et al., 1999), a bivariate generalization of the boxplot which visualizes the location, scatter, shape and tails of the data. Depth can also be applied to the classification of multivariate data (Li et al., 2012). Recently, notions of depth were introduced for univariate functional data, such as the modified band depth of López-Pintado and Romo (2009). For multivariate curves Claeskens et al. (2014) proposed the multivariate functional halfspace depth (MFHD). It considers a set of multivariate curves on the same time interval, and defines the depth of a curve as the (possibly weighted) integral of its halfspace depth at* *each time point. We will use MFHD for the supervised classification of curves, and compare its performance with that of other classifiers. Their behavior is also studied in the presence of outlying curves.*

Keywords. *Ranking; Halfspace depth; Bagplot; Classification of curves*

1.03

Explicit optimal rules for functional classification

José R. Berrendero[1], Antonio Cuevas[1] and José L. Torrecilla[1,*]

[1] *Universidad Autónoma de Madrid; joser.berrendero@uam.es, antonio.cueva@uam.es, joseluis.torrecilla@uam.es*
[*] *Corresponding author*

Abstract. *Given trajectories generated by two different stochastic processes, we are concerned with the classification problem of identifying the population to which a new observation belongs. In some cases, when the classification problem consists of deciding between two different Gaussian processes with equivalent measures, the optimal (Bayes) classification rule can be explicitly calculated. Some classical results (due to Cameron, Martin, Parzen and Shepp, among others) provide expressions for the Radon-Nikodym derivatives of these measures from which the Bayes rule can be obtained. This Bayes rule usually depends on mean functions and covariance structures which must be empirically estimated. Some proposals are given in order to solve these and other estimation problems in practice. Some connections with variable selection are also studied.*

Keywords. *Functional classification; Bayes rule; Gaussian process; Dimension reduction.*

1.04

The Mahalanobis distance for functional data with applications to classification

P. Galeano[1,*], E. Joseph[1] and R. E. Lillo[1]

[1] *Universidad Carlos III de Madrid; pedro.galeano@uc3m.es, esdras.joseph@uc3m.es, rosaelvira.lillo@uc3m.es*
[*] *Corresponding author*

Abstract. *This paper presents a new semi-distance for functional observations that generalizes the Mahalanobis distance for multivariate datasets. The main characteristics of the functional*

Mahalanobis semi-distance are shown. In order to illustrate the applicability of this measure of proximity between functional observations, new versions of several well known functional classification procedures are developed using the functional Mahalanobis semi-distance. A Monte Carlo study and the analysis of two real examples indicate that the classification methods used in conjunction with the functional Mahalanobis semi-distance give better results than other well-known functional classification procedures.

Keywords. *Classification methods; Functional data analysis; Functional Mahalanobis semi-distance; Functional principal components.*

1.05

Functional principal fitted componens regression using B-spline expansion

AhYeon Park[1,*], Serge Guillas[1]

[1] *University College London; ah.park.09@ucl.ac.uk, s.guillas@ucl.ac.uk*
[*] *Corresponding author*

Abstract. *We consider functional linear regression with a scalar response and a functional covariate. Dimension reduction for this regression can be achieved, either through functional principal components regression (FPCR), or through penalized B-splines regression (Cardot et al., 2003; Goldsmith et al., 2011). In FPCR (Reiss and Ogden, 2007), the subspace of the functional coefficient is restricted to the span of selected principal components, thus the components are chosen without regard to how well they predict the response. We introduce functional principal fitted components regression (FPFCR) that takes into account the response when choosing the components. The model is an extension of multivariate principal fitted components regression (Cook and Forzani, 2008) to functional data. To include the response in the reduction step, we use an inverse approach. We carry out an inverse regression of the functional covariate on the response, and compute the conditional covariance function of the covariate given the response. To estimate the parameters involved in the inverse regression we use maximum likelihood theory. The covariance of the error function in the inverse regression has a general form, so its inverse can be ill-conditioned. We suggest two regularizing methods, truncation (James et al., 2000) and smoothing (Silverman, 1996) and deal with the ill-posed inverse problem. We fit a regression of the scalar responses on the scores of the selected fitted components. We control the roughness of the coefficient function via a penalized B-Splines approach. The performance of the methodology is illustrated via simulations and real data analyses.*

Keywords. *Functional linear models; Functional principal components; Functional principal fitted components; Penalized B-splines; Ill-posed inverse problems*

References

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.

Cook, R. D. and L. Forzani (2008). Principal fitted components for dimension reduction in regression. *Statistical Science* **23**, 485–501.

Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20**, 830–851.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

1.06

Parametrically guided nonparametric density and hazard estimation with censored data

M. Talamakrouni[1,*] I. Van Keilegom[1] and A. El Ghouch[1]

[1] *Université catholique de Louvain; majda.talamakrouni@uclouvain.be, ingrid.vankeilegom@uclouvain.be, anouar.elghouch@uclouvain.be*
[*] *Corresponding author*

Abstract. *The parametrically guided kernel smoother proposed by Hjort and Glad (1995) is a promising nonparametric estimation approach that aims to reduce the bias of the classical kernel density estimator without increasing its variance. In this paper we generalize this method to the censored data case and show how it can be used for density and hazard function estimation. The asymptotic properties of the proposed estimators are established and their performance is evaluated via finite sample simulations.*

Keywords. *Right censoring; Density estimation; Kernel smoothing; Maximum likelihood; Kaplan-Meier estimator.*

References

Hjort, N.L. and Glad, I.K.(1995). Nonparametric density estimation with parametric start. *The Annals of Statistics* **23**, 882–904.

1.07

The jackknife estimate of covariance under censorship when covariables are present

L. Azarang [1,*], J. de Uña Álvarez[1], and W. Stute[2]

[1] *University of Vigo, Campus de Vigo As Lagoas, Marcosende, s/n, 36310 Vigo, Pontevedra, Spain; leyla.azarang@uvigo.es, jacobo@uvigo.es*
[2] *University of Giessen; winfried.stute@math.uni-giessen.de*
[*] *Corresponding author*

Abstract. *Multivariate Kaplan-Meier integrals were introduced by Stute (1993), for a situation in which a p-variate vector of covariates is paired with the possibly unobserved lifetime. It has been shown that, under mild conditions, a vector of multivariate Kaplan-Meier integrals is asymptotically normal. In this work the covariance between two Kaplan-Meier integrals with covariates is estimated using the Jackknife method. It is shown that the Jackknife estimate of covariance consistently estimates the limit covariance. Application in the scope of the estimation of transition probabilities for multi-state models is discussed.*

Keywords. *Strong consistency; Survival analysis; Censored data; Kaplan-Meier; Illness-death model.*

References

Stute, W. and Wang, J.-L. (1993). The Strong Law under Random Censorship. *The Annals of Statistics* **21**, 1591-1607.

Stute, W. (1993). Consistent Estimation under Random Censorship When Covariables Are Present. *Journal of Multivariate Analysis* **45**, 89-103.

Stute, W. and Wang, J.-L. (1994). The Jackknife Estimate of a Kaplan- Meier Integral. *Biometrika* **81**, 602-606.

Stute, W. (1996b). The Jackknife Estimate of Variance of a Kaplan-Meier Integral. *The Annals of Statistics* **24**, 2679-2704.

Stute, W. (1996a). Distributional Convergence under Random Censorship When Covariables Are Present. *Scandinavian Journal of Statistics* **23**, 461-471.

1.08

Kaplan-Meier estimator based on ranked set samples

E. Strzalkowska-Kominiak [1,*] and M. Mahdizadeh [2]

[1] *Department of Statistics, Universidad Carlos III de Madrid, Spain; ewa.strzalkowska@uc3m.es,*
[2] *Department of Statistics, Hakim Sabzevari University, Iran.*
[*] *Corresponding author*

Abstract. *When quantification of all sampling units is expensive but a set of units can be ranked, without formal measurement, ranked set sampling (RSS) is a cost-efficient alternative to simple random sampling (SRS). See, e.g., Chen et al (2004). In this work, we propose a new Kaplan-Meier estimator for the distribution function based on RSS under random censoring and study its asymptotic properties. We present a simulation study to compare the performance of the proposed estimator and the standard Kaplan-Meier estimator based on SRS. It turns out that RSS design can yield a substantial improvement in efficiency over the SRS design. See, Strzalkowska-Kominiak and Mahdizadeh (2013) for details. Additionally, we apply our methods to a real data set from an environmental study. Finally, we propose a new bootstrap approach in the setup of ranked set sampling under censoring. Similarly, as by Kaplan-Meier estimator under simple random sampling, the bootstrap based confidence intervals for the parameter of* *interest are more accurate than the intervals based on the asymptotic normality.*

Keywords. *Ranked Set Sampling; Random Censorship; Kaplan-Meier estimator.*

References

Chen, Z., Bai, Z., Sinha, B.K. (2004) *Ranked set sampling: theory and applications.* New York: Springer.

Strzalkowska-Kominiak, E., Mahdizadeh, M. (2013). On the Kaplan-Meier estimator based on ranked set samples. *Journal of Statistical Computation and Simulation* (in press)

1.09

Generalized seasonal block bootstrap methos for time series with periodic structure

Anna Dudek[1,*]

[1] *AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland; aedudek@agh.edu.pl*
[*] *Joint work with J. Leśkow, D. Politis and E. Paparoditis*

Abstract. *When time series data contain a periodic component, the usual block bootstrap procedures are not directly applicable. We propose a modification of the block bootstrap—the Generalized Seasonal Block Bootstrap (GSBB)—and show its asymptotic consistency without undue restrictions on the relative size of the period and block size. The consistency of GSBB is shown for the overall mean and seasonal means of periodically correlated (PC) time series. Moreover, we present applicability of GSBB for triangular arrays with growing period.*

Keywords. *Generalized Seasonal Block Bootstrap; Periodic time series; Resampling; Seasonality.*

References

Dudek, A. E., Leśkow, J., Politis, D. and Paparoditis, E. (2014). A generalized block bootstrap for seasonal time series. *J. Time Ser. Anal.* **35**, 89–114.

1.10

The autoregresive-aided block bootstrap

Tobias Niebuhr[1,*], Jens-Peter Kreiss[1] and Efstathios Paparoditis[2]

[1] *Institut für Mathematische Stochastik, Technische Universität Braunschweig, Germany; t.niebuhr@tu-braunschweig.de, j.kreiss@tu-braunschweig.de*

[2] *Department of Mathematics and Statistics, University of Cyprus; stathisp@ucy.ac.cy*
* *Corresponding author*

Abstract. *The bootstrap has been established as a powerful tool in the field of nonparametric statistics. We present a modification of the general block bootstrap procedure, called the autoregressive-aided (AR-aided) block bootstrap, and show its asymptotic validity.*
The proposal consists of two steps. An autoregressive model fit is applied to the time series observations at first; in a second step the estimated residuals are block bootstrapped. Bootstrap validity will be shown to hold under very mild assumptions, namely as long as the time series yields stationarity. Thus bootstrap validity goes far beyond the class of AR processes. Especially, any given dependence structure of the noise (e.g. m-dependence) can be captured correctly. The presented procedure is tailor-made for several models such as discrete-time observations of continuous-time autoregressive moving average processes. A short simulation study will give insight in the procedure's performance and conclude the talk.

Keywords. *Block bootstrap; Dependent noise; Weak ARMA; CARMA process.*

1.11

Booystrapping realized covariance

G. Feng[1,*] and J.-P. Kreiss [1]

[1] *Institut für Mathematische Stochastik, Technische Universität Braunschweig, Germany; g.feng@tu-braunschweig.de, j.kreiss@tu-braunschweig.de*
* *Corresponding author*

Abstract. *Modeling the financial market appropriately is of essential importance when handling stock prices and options. Especially the market's volatility is of great interest. Realized covariance as a consistent estimator for the integrated covariance is often used to measure the financial market's volatility with multivariate high frequency data.*
Starting with a multivariate nonparametric volatility model, we propose a resampling procedure in order to approximate the distribution of the realized covariance. The crucial point for our approach will be shown to be extraction of the correlation structure based on discrete time returns. Asymptotic validity of the proposed resampling procedure will be proved. Furthermore, a simulation study will conclude the talk comparing performances for finite sample properties of traditional methods and our above described approach.

Keywords. *Realized covariance; Nonparametric bootstrap*

References

Barndorff-Nielsen, O and Shephard, N. (2004). Econometric analysis of realised covariation: high frequency based covariance, regression and correlation in financial economics. *Econometrica* **72**, 885–925.

Dovonon, P., Goncalves, S. and Meddahi, N. (2013). Bootstrapping realized multivariate volatility measures. *Journal of Econometrics* **1**, 49–65.

1.12

Bootstrap for nonparametric trend estimation in locally stationary time series

Jonas Krampe[1*], Efstathios Paparoditis[2] and Jens-Peter Kreiss[1]

[1] *Institut für Mathematische Stochastik, TU Braunschweig Pockelsstrasse 14, D-38106 Braunschweig; j.krampe@tu-bs.de, j.kreiss@tu-bs.de*
[2] *Department of Mathematics and Statistics, University of Cyprus, P.O.Box 20537, CY-1678 Nicosia; stathisp@ucy.ac.cy*
[*] *Corresponding author*

Abstract. *Based on the idea of local stationary processes with time varying spectral densities (Dahlhaus, 2012), this work deals with nonparametric trend estimation for locally stationary time series. For a kernel based estimation procedure consistency and asymptotic normality will be shown. In order to obtain an alternative approximation of the distribution of the proposed nonparametric trend estimator a wild bootstrap approach is suggested (Kreiss and Paparoditis, 2014), which correctly mimics the first, second and (to a sufficient extend) also the fourth order structure of the time series. Consistency of the proposed bootstrap procedure for nonparametric trend estimation will be shown and the finite sample size ability of the bootstrap will be demonstrated by simulations.*

Keywords. *bootstrap; locally stationary processes; kernel estimation; trend function*

References

Dahlhaus, R. (2012). Locally Stationary Processes.*Handbook of Statistics 30*, 351-413.

Kreiss, J.-P. and Paparoditis, E. (2014). Bootstrapping Locally Stationary Processes. *J. R. Statist. Soc. B*, to appear

1.13

A boosting algorithm for estimating generalized propensity scores with continuous treatments

Yeying Zhu[1], Donna L.Coffman[2] and Debashis Ghosh[3]

[1] *Department of Statistics and Actuarial Science, University of Waterloo; yeying.zhu@uwaterloo.ca*
[2] *The Methodology Center, Pennsylvania State University; dcoffman@psu.edu*
[3] *Department of Statistics, Pennsylvania State University; ghoshd@psu.edu*

Abstract. *In this talk, we study causal inference with a continuous treatment variable using propensity score-based methods. For a continuous treatment, the generalized propensity score is defined as the conditional density of the treatment level given covariates (confounders). The dose-response function is then estimated by inverse probability weighting, where the weights are calculated from the estimated propensity scores. When the dimension of the covariates is large, the traditional nonparametric density estimation suffers from the curse of dimensionality. Some researchers have suggested a two-step estimation procedure by first modeling the mean function. In this study, we suggest a boosting algorithm to estimate the mean function of the treatment given covariates. In boosting, an important tuning parameter is the number of trees to be generated, which essentially determines the trade-off between bias and variance of the causal estimator. We propose a criterion called average absolute correlation coefficient (AACC) to determine the optimal number of trees. The idea is that the treatment variable and the covariates are supposed to be unconfounded in the weighted pseudo sample. Our simulation results show that the proposed method works better than a simple linear approximation (Robins et al., 2000) or L_2 boosting (Bühlmann and Yu, 2003). The proposed methodology is also illustrated through an obesity study: Early Dieting in Girls study, which examines the influence of mothers' overall weight concern on daughters' dieting behavior.*

Keywords. *Boosting; Distance Correlation; Dose-Response Function; Generalized Propensity Scores; High-dimensional.*

References

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2-loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324–339.

Robins, J.M. and Hernán, M.Á. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

1.14 Nonparametric estimation of a conditional distribution for inter-occurrence times, through waiting times in a cross-sectional sampling

J.A. Cristóbal[1] and J.T. Alcalá[1,*]

[1] *Statistical Methods Department, University of Zaragoza (Spain); cristo@unizar.es; jtalcala@unizar.es*
[*] *Corresponding author*

Abstract. *The goal of this paper is to make inferences about waiting times X between two consecutive events of a stationary renewal process, such as the unemployment times of different individuals in a certain population. We suppose that our data (obtained by cross-sectional sampling) are the backward recurrence times Y from the occurrence of the last event up to a pre-established time, such as the time between the last unemployment entry and the sampling time, along with the corresponding values of a certain set of covariates Z. We deduce a nonparametric estimation of the regression function $E(X|Z)$ only based on our data Y, which are obtained through a multiplicative censoring of the unobservable (and biased) durations X^w. To achieve this goal, we first construct an adjusted Nadaraya-Watson estimator of the conditional* *distribution function of $(Y|Z)$, starting from our data (Hall, Wolf and Yao (1999)). Then we use de Pool Adjacent Violators Algorithm (PAVA) to obtain an estimator of the probability density function $f\,(Y|Z)$. Specifically, we use the Penalized Maximum Likelihood method (Woodroofe and Sun (1993)) to make the estimate consistent in the origin. Finally, we obtain an estimator of the conditional distribution function of $(X|Z)$. We carry out a comprehensive study of simulation with different distributions for the X variable and some types of regression functions, paying special attention to the problem of the automatic choice of the smoothing parameter (Li and Racine (2008)).*

Keywords. *Backward recurrence times; CDF estimation; Penalized Maximum Likelihood method; Pool Adjacent Violators algorithm.*

References

Hall, P., Wolf, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of American Statistical Association* **94**, 154–163.

Li, Q. and Racine, J. (2008) Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics* **26**, 423-434.

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of f(0+) when f is non-increasing. *Stat. Sinica* **3**, 501-515.

1.15

Nonparametric method for estimating the distribution of time to failure of engineering materials

S. Naya[1,*], I. López-de-Ullibarri[1] J. Tarrio-Saavedra[1] and A. Meneses[1]

[1] *Universidade da Coruña, Escuela Politécnica Superior, Campus de Esteiro, s/n, 15403 Ferrol, Spain; salva@udc.es, ilu@udc.es, jtarrio@udc.es, antoniomenesesfreire@hotmail.com*

[*] *Corresponding author*

Abstract. *Estimating the failure-time distribution or long-term performance of components of high-reliability products is particularly difficult. In recent years there has been an upsurge in the use of statistical methods to improve the reliability of engineering materials (Castillo and Fernandez-Canteli, 2009). Following up on the papers of Pinheiro and Bates (1995) and Meeker et al. (1998), we apply a flexible statistical methodology to the study of fatigue models used to estimate the reliability of different materials. We propose a non parametric method for modelling crack growth and estimating failure-time. In this work we present models for the estimation from the physical point of view, devoted to the statistical study and present a spline-based flexible method of estimation of data degradation. We conclude with an application to real data.*

Keywords. *data degradation; reliability data; failure-time.*

References

Castillo, E. and Fernandez-Canteli, A. (2009). *A Unified Statistical Methodology for Modeling Fatigue Damage.* Springer.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4:1**, 12–35.

Meeker, W. Q., Escobar L.A. and Lu J. (1998) Accelerated Degradation Tests: Modeling and Analysis. *Technometrics* **40:2**, 89–99.

1.16

Moving order statistics

López Blázquez, F.[1,*], Salamanca Miño, B. [1]

[1] *Universidad de Sevilla, Spain; lopez@us.es, bsm@us.es;*
[*] *Corresponding author*

Abstract. *In many experiments the observations are obtained sequentially and a window consisting on the last n observations is useful in the statistical analysis. For instance, this is the case of moving averages. Some other statistics may be of interest, so, we consider the case of order statistics. Although moving order statistics (MOS) appear frequently in practice, it seems that their distribution theory is not well-known. Our purpose is to provide a rigorous framework for the study of MOS. Our results can be extended to other situations in which order statistics are obtained from overlapping samples. We also compare our results to some previous works in this area.*

Keywords. *Order statistics; Overlapping samples; Moving maxima and minima*

1.17

Classification based on non-negative matrix factorization

Hong Gu*, Toby Kenney and Yun Cai

Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada;
hgu@dal.ca, tkenney@mathstat.dal.ca, caiy@mathstat.dal.ca

Abstract. *Inflammatory bowel disease (IBD), comprising Crohn's disease (CD) and ulcerative colitis (UC), is emerging as a global health problem. Metagenomics is a key factor in the disease. Our understanding of the metagenomics involved is still very limited. Statistically, it is a very challenging problem to identify the metagenomic community behaviours which are associated*

with the disease states, due to the extremely high dimensionality of the data and relatively much smaller sample sizes.
We apply non-negative matrix factorization to these data to find the typical types for the diseased group and typical types for the healthy group. Based on these typical types, we can effectively reduce the dimensionality of the problem and assemble a suitable supervised learning method for the transformed data. Software is developed to facilitate the interactive exploration of the data using these typical types as references. Such interactive exploration can help to decide whether linear or non-linear methods are suitable as a learning method. Excellent separation can be achieved using our proposed methods for these complex metagenomic data.

Keywords. *Metagenomics;Non-negative Matrix Factorization;Supervised Learning;Data Visualization*

1.18

Inhomogeneous large-scale data: maximin effects and their statistical estimation

Peter Bühlmann[1]

[1] *Seminar for Statistics, ETH Zürich*

Abstract. *Large-scale or "big" data usually refers to scenarios with potentially very many variables (dimension p) and very large sample size n. Such data is most often of "inhomogeneous" nature, i.e., neither being i.i.d. realizations from a distribution nor being generated from a stationary distribution. We propose a new methodology for some class of large-scale inhomogeneous data, in terms of so-called maximin effects which optimize performance in the most adversarial constellation. The advocated procedure is computationally efficient and we provide corresponding statistical accuracy guarantees for scenarios where n and/or p are large.*

1.19

Nonparametric estimation of conditional distribution functions via pre-adjustment techniques and extensions

I. Gijbels

Department of Mathematics and Leuven Statistics Research Center, KU Leuven, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium

Abstract. *A conditional distribution function describes how the distribution of a variable of interest Y changes with the value taken by a covariate X (or a set of covariates). Nonparametric estimation of a conditional distribution function has been studied in, for example, Hall et al. (1999). In Veraverbeke et al. (2014) it is discussed how estimation of a conditional dis-*

tribution function might depend on whether one has any prior knowledge on how the covariate X influences Y. In this talk we discuss various ways of pre-adjustments, that pre-adjust the response observations for 'obvious' effects of the covariate, and as such provide opportunities to an improved estimation of the conditional distribution function. We also discuss links to other recent developments, such as these provided via Swanepoel and Van Graan (2005) and Kiwitt and Neumeyer (2012).

The basic idea of pre-adjustment is applicable to a variety of other estimation settings, such as conditional quantile estimation, conditional density estimation, conditional copula estimation, ... See also Gijbels et al. (2013). Moreover, the methodology can be extended to more complex data settings such as censored data.

A simulation study and real data applications illustrate the performances of the discussed methods, and their use in practice.

This talk is based on joint work with Marek Omelka and Noël Veraverbeke

Keywords. *Conditional distribution function; conditional density and quantile function; pre-adjustment; kernel smoothing.*

References

Gijbels, I., Omelka, M. and Veraverbeke, N. (2013). Estimation of a copula when a covariate affects only marginal distributions. *Manuscript.*

Hall, P., and Wolff, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154–163.

Kiwitt, S. and Neumeyer, N. (2012). Estimating the conditional error distribution in nonparametric regression. *Scandinavian Journal of Statistics*, **39**, 259–281.

Swanepoel, J. W. H. and Van Graan, F. C. (2005). A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics*, **32**, 551–562.

Veraverbeke, N., Gijbels, I. and Omelka, M. (2014). Pre-adjusted nonparametric estimation of a conditional distribution function. *Journal of the Royal Statistical Society, Series B*, 76, 399–438.

1.20

Efficiency in functional nonparametric models with autoregressive errors

S. Dabo-Niang[1], S. Guillas[2] and C. Ternynck[1,*]

[1] *Laboratory EQUIPPE, University Lille 3, Villeneuve d'Ascq, France ; sophie.dabo@univ-lille3.fr, camille.ternynck@univ-lille3.fr*
[2] *Department of Statistical Science, University College London, London, UK ; s.guillas@ucl.ac.uk*
[*] *Corresponding author*

Abstract. *In this talk, a kernel-based procedure of estimation for a nonlinear functional regression is introduced in the case of a functional predictor and a scalar response. More precisely, the explanatory variable takes values in some abstract function space and the residual process is stationary and autocorrelated. The procedure consists in a pre-whitening transformation of* *the dependent variable as it is done in the multivariate context by Xiao* et al. *(2003). The main idea is to transform the original regression model, so that this transformed regression has a residual term that is uncorrelated.The asymptotic distribution of the proposed estimator is established considering that the explanatory variable is an $\alpha-$mixing process, the most general case of weakly dependent variables. Although, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely ("working independence estimator"), it is shown here that the autocorrelation function of the error process has useful information for improving estimators of the regression function. The skills of the methods are illustrated on simulations where the relative efficiency of the proposed efficient estimator over the conventional estimator is given for different values of the auto-regressive parameters.*

Keywords. *Kernel regression - Time series - Pre-whitening - Functional data*

References

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes, Estimation and Prediction.* Springer, second edition.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic processes and their applications* **115(1)**, 155–177.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis.* Springer, second edition.

Xiao, Z., Linton, O., Carroll, R. and Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* **98(464)**, 980–992.

1.21

Estimation in flexible functional models using continuous wavelet dictionaries

G. Claeskens[1], M. Giacofci[2,*], I. Gijbels[2] and M. Jansen[3]

[1] *ORSTAT & Leuven Statistics Research Center, KUL; gerda.claeskens@kuleuven.be*
[2] *Department of Mathematics & Leuven Statistics Research Center, KUL; joycemadison.giacofci@wis.kuleuven.be, irene.gijbels@wis.kuleuven.be*
[3] *Departments of Mathematics and Computer Science, ULB, maarten.jansen@ulb.ac.be*
[*] *Corresponding author*

Abstract. *Owing to the constant evolution of technologies many scientific studies lead nowadays to the collection of large amounts of data. We consider data that consist of set of curves recorded on individuals and that are usually modelled as functional data. In a multi-individual context, curves are often recorded on subject-specific non-equidistant time points and variations in phase (i.e variations in timing of features) or in amplitude (i.e variations in size of features) are commonly encountered. While amplitude variations is a widely studied subject in the functional setting, only little attention has been devoted to the study of phase variations as a proper* *source of information.*

We consider a model derived from the flexible modelling setting proposed by Slaets, Claeskens & Jansen (2011), based on a continuous wavelet representation of curves using overcomplete dictionaries. The use of wavelets allows to consider a wide range of irregular curves while the overcomplete setting offers a good flexibility. Individual variations in phase and in amplitude are modelled by introducing random effects for wavelet scales and locations and for wavelet coefficients. The model can be fit into the class of nonlinear mixed-effects models and follows the idea that the data result from one main pattern with curve-specific deviations in location, scale and amplitude.

In this setting, we propose a new procedure for parameter estimation. Maximum likelihood estimation of the model parameters is performed using an MCEM algorithm, a variant of the EM algorithm. The non-linearity of the mixed model yields the E-step untractable and a Monte-Carlo integration is used for random effects predictions of individual wavelet scales and locations and individual coefficient amplitudes. The performance of the procedure is investigated through a simulation study.

Keywords. *Wavelets; Overcomplete Dictionaries; Nonlinear Mixed Effects Models; EM algorithm*

References

Slaets, L., Claeskens, G. and Jansen, M. (2011). Flexible Modelling of Functional Data using Continuous Wavelet Dictionaries. In F.Ferraty (ed.) *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics , 297–300. Physica-Verlag HD

1.22

Homogeneity test for functional data based on depth measures

Ramón Flores, Rosa Lillo and Juan Romo

rflores@est-econ.uc3m.es, romo@est-econ.uc3m.es
lillo@est-econ.uc3m.es
Corresponding author: Ramón Flores

Abstract. *We deal with the problem of testing homogeneity of samples in the context of functional data analysis. We propose hypothesis tests that use four different statistics to measure distance between samples. The statistics are based on different depth measures recently defined in the literature. A Monte Carlo study and the analysis of real examples indicate that the efficiency of the proposed methods is very acceptable when confronted with shape or magnitude perturbations.*

Keywords. *Functional data; Depth measures; Homogeneity tests.*

1.23

On the optimal allocation of components in parallel-series and series-parallel systems

Henry Laniado [1,*], Jiantian Wang [2]

[1] *Departamento de Estadística, Universidad Carlos III de Madrid, 28911, Leganés, Spain; hlaniado@est-econ.uc3m.es.com; @gmail.com*
[2] *Department of Mathematics, Kean University, Union, NJ, 07083, USA; jwang@kean.edu*
[*] *Corresponding author*

Abstract. *We study the problem of where and how allocate the components in two-parallel-series and two-series-parallel systems in order to optimize the reliability in some stochastic sense. In the literature this problem has been solved considering the usual stochastic order and recently in Laniado and Lillo (2014), assuming two types of components, obtained the result in both, the hazard rate order and reversed hazard rate order. In this talk we show a stronger result by considering the likelihood rate order.*

Keywords. *Parallel-series system; Series-parallel system; Allocation policy; Stochastic order; Proportional hazard rate models.*

References

Laniado, H., Lillo, R.E., (2014). Allocation policies of redundancies in two-parallel-series and two-series-parallel systems. *IEEE Transactions on Reliability* **63**, 223–229.

1.24

A lack-of-fit test of quantile regression models with multiple covariates

Mercedes Conde-Amboage[1], César Sánchez-Sellero[1,*] and Wenceslao González-Manteiga[1]

[1] *Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain; mercedes.amboage@usc.es, cesar.sanchez@usc.es, wenceslao.gonzalez@usc.es*
[*] *Corresponding author*

Abstract. *We propose a new lack-of-fit test for quantile regression models with multiple covariates. The test is based on the cumulative sum of residuals with respect to unidimensional linear projections of the covariates. The test is then adapting the ideas of Escanciano (2006) to cope with multiple covariates, to the test proposed by He and Zhu (2003). To approximate the critical values of the test, a wild bootstrap mechanism is used which is similar to that proposed* *by Feng, He and Hu (2011). An extensive simulation study was carried out that shows the good properties of the new test, particularly when the dimension of the covariate is high. The test can also be applied and performs well under heteroscedastic regression models. The test is illustrated with real data about economic growth of 161 countries. The data set is available in the R package quantreg, under the name barro.*

Keywords. *Quantile regression; Lack-of-fit testing; Multiple regression.*

References

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric theory* **22**, 1030–1051.

Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika* **22**, 995–999.

He, X. and Zhu, L.-X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association.* **98**, 1013–1022.

1.25

A bandwidth selector for nonparametric quantile regression

Mercedes Conde-Amboage[1] and César Sánchez-Sellero[1,*]

[1] *Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain; mercedes.amboage@usc.es, cesar.sanchez@usc.es*
[*] *Corresponding author*

Abstract. *In the framework of quantile regression, local linear smoothing techniques have been studied by several authors, particularly by Yu and Jones (1998). The problem of bandwidth selection was addressed in the literature by the usual approaches, such as cross-validation or plug-in methods. Most of the plug-in methods rely on restrictive assumptions on the quantile regression model in relation to the mean regression, or on parametric assumptions. Here we present a plug-in bandwidth selector for nonparametric quantile regression, that is defined from a completely nonparametric approach. To this end, the curvature of the quantile regression function and the integrated sparsity (inverse of the conditional density) are both nonparametrically estimated. The new bandwidth selector is shown to work well in different scenarios, particularly when the conditions commonly assumed in the literature are not satisfied.*

Keywords. *Quantile regression; Bandwidth; Nonparametric regression.*

References

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association.* **93**, 228–237.

1.26

Model selection via bayesian information criterion for quantile regression models

E. R. Lee[1], H. Noh[2], * and B.U. Park

[1] *University of Mannheim, Germany; silverryuee@gmail.com*
[2] *Sookmyung Womens University, Korea; word5810@gmail.com*
[3]*Seoul National University, Korea.*
** Corresponding author*

Abstract. *Bayesian Information Criterion (BIC) is known to identify the true model consistently as long as the predictor dimension is finite. Recently, its moderate modifications have been shown to be consistent in model selection even when the number of variables diverges. Those works have been done mostly in mean regression, but rarely in quantile regression. The best known results about BIC for quantile regression are for linear models with a fixed number of variables. In this paper, we investigate how BIC can be adapted to high-dimensional linear quantile regression and show that a modified BIC is consistent in model selection when the number of variables diverges as the sample size increases. We also discuss how it can be used for choosing the regularization parameters of penalized approaches that are designed to conduct variable selection and shrinkage estimation simultaneously. Moreover, we extend the results to structured nonparametric quantile models with a diverging number of covariates. We illustrate our theoretical results via some simulated examples and a real data analysis on human eye disease.*

Keywords. *high-dimension; linear quantile regression; nonparametric quantile regression; model selection consistency; regularization parameter selection; shrinkage method.*

1.27

Nonparametric estimators of extreme value index based on averaged regression quantiles

Jan Picek

Department of Applied Mathematics, Technical University of Liberec, Czech Republic; jan.picek@tul.cz

Abstract.
The problem of estimating the so-called extreme value index, which determines the behavior of the distribution function in its upper tail, has received much attention in the literature, see e.g. (de Haan and Ferreira, 2006) and references cited there. More attention has been paid to estimators that are based on a certain number of upper order statistics.
However, one of the challenging ideas of the recent advances in the field of statistical modeling of extreme events has been the development of models with time-dependent parameters or more

generally models incorporating covariates. Therefore, in the present contribution we aim at extending the general result given in Drees (1998) to linear regression.

The contribution deals with estimators of extreme value index based on weighted averaged regression α-quantile in the linear regression model. Jurečková and Picek (2014) showed asymptotic equivalence to the α-quantile of the location model. The weighted averaged quantiles can be seen as a possible generalization of the quantile idea. Following Drees (1998) we consider a class of smooth functionals of the tail quantile function as a tool for the construction of estimators in the linear regression context. Pickands and probability weighted moments estimators are illustrated on simulated and climatological data.

Keywords. *Linear regression model; averaged regression quantile; extreme value index.*

References

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory, An Introduction*, Springer, New York.

Drees, H. (1998). On Smooth Statistical Tail Functionals. *Scandinavian Journal of Statistics* **25**, 187–210.

Jurečková, J. and Picek, J. (2014). Averaged regression quantiles. *Contemporary Developments in Statistical Theory*, Springer Proceedings in Mathematics & Statistics, Volume 68, 203–216.

1.28

Distribution free independence tests between two point processes

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

[1] *Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France; Melisande.Albert@unice.fr, Patricia.Reynaud-Bouret@unice.fr*

[2] *Univ. Nice Sophia Antipolis, CNRS, LPMC, UMR 7336, 06100 Nice, France; yann.bouret@unice.fr*

[3] *Univ. européenne de Bretagne, IRMAR; magalie.fromont@univ-rennes2.fr*

Abstract. *Considering an i.i.d. sample from the joint distribution of a pair of point processes observed on a given time period, we address the question of detecting dependence between the two underlying marginal point processes. This question is motivated by correlation studies of spike trains in neuroscience (see Tuleau-Malot et al. (2013), Grün et al. (2010) or Pipa and Grün (2003)). Because of the large debate on models for spike train analysis, our aim is to propose independence tests that are free from the joint distribution.*

We mainly follow Romano (1989) who proposed independence tests in $\mathbb{R}^d$ based on bootstrap and permutation (see also Hoeffding (1952)) approaches. Here, due to the nature of our objects (namely point processes), our test statistics are much more complex, and cannot be seen as empirical processes evaluated on particular families of events. Therefore, we need to push further Romano's arguments to obtain for instance, even in this case, the convergence of the distributions of both bootstrapped and permuted statistics to the true distribution of the statistic under the independence assumption.

We will first present the motivations of our independence tests from the neuroscience point of *view. Then, we will underline the main variations w.r.t independence tests in $\mathbb{R}^d$ when proving that they are of the asymptotic or exact desired level and also consistent against particular alternatives.*

Keywords. *Point processes; Independence tests; Permutation tests; Bootstrap methods; Multiple testing.*

References

Grün, S., Diesmann, M., and Aertsen, A.M. (2010). Unitary Events Analysis. *In Analysisof Parallel Spike Trains*, Grün, S., and Rotter, S., Springer Series in Computational Neuroscience.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics* **23**, 2, 169–192.

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

Romano, J.P. (1989). Bootstrap and Randomization Tests of some Nonparametric Hypotheses. *The Annals of Statistics* **17**, 1, 141-159.

1.29

Consistent and powerful nonparametric tests for dependence

Andrey Feuerverger[1]

[1] *University of Toronto; andrey@utstat.toronto.edu*

Abstract. *Modern applications and current volumes of data require new approaches to the problems of testing for dependence. Such needs arise in financial engineering contexts, copula modeling, and in many other areas where subtle dependence structures may be an issue. Such applications call for tests which have demonstrably high power, and which are consistent against all alternatives to independence. It turns out that tests constructed carefully in the Fourier domain have these desirable properties; they also turn out have suggestive and unexpectedly interesting functional forms. The emphasis of the talk will be on basic ideas and on how they can be extended to develop tests applicable to diverse dependence testing contests.*

Keywords. *Consistency; Dependence; Nonparametric; Testing.*

1.30

Conditional empirical copula process for time series

Félix Camirand Lemyre[1], Jean-François Quessy[2,*] and Taoufik Bouezmarni[1]

[1] *Département de mathématiques, Université de Sherbrooke, Québec, Canada; Felix.Camirand.Lemyre@usherbrooke.ca, taoufik.bouezmarni@mat.usherbrooke.ca*
[2] *Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Trois-Rivières, Canada; jean-francois.quessy@uqtr.ca.*
[*] *Corresponding author*

Abstract. *The dependance structure between two random variables might be influenced by some covariate. To model conditional dependance two conditional copula estimators were proposed in* **?** *and their asymptotic properties were studied in* **?** *in the case of i.i.d data. This presentation is concerned about some asymptotic properties of these two estimators in the case of mixing data.*

Keywords. *Bootstrap ; Conditional copula; Empirical copula process ; Strong mixing ; Weak convergence.*

References

Gijbels, Irène and Veraverbeke, Noël and Omelka, Marel. (2011). Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis* **55** 1919–1932.

Veraverbeke, Noël and Omelka, Marek and Gijbels, Iréne. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics* **38** 766–780.

1.31

A von Mises approach for the small sample distribution of the trimmed mean

A. García-Pérez

[1] *Departamento de Estadística, I. O. y C. N., Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040-Madrid, Spain; agar-per@ccia.uned.es*

Abstract. *The von Mises approach (Serfling, 1980), is a very useful tool to transfer a known value of a functional at a model distribution G to the unknown value of this functional at another close distribution F, if we are able to integrate the Hampel's influence function, or the Tail Area Influence Function (TAIF) if the tail probability functional is considered.*

Only saddlepoint approximations of the TAIF were used in this approximation (see García-Pérez (2008), and the references therein) but in a recent paper, García-Pérez (2012), a closed-form expression of the TAIF was obtained, and an iterative procedure used for not only very close distributions, improving in this way the accuracy and the applicability of the von Mises approach.

In this context, a new analytic approximation of the small sample distribution of the trimmed mean is obtained, that is accurate, easy to apply and where the elements involved on it have a straightforward interpretation. This allows, for instance, a better choice of the trimming fraction in a test based on the trimmed mean.

Keywords. *Trimmed mean; Point estimation; Hypotheses testing; Robustness.*

References

García-Pérez, A. (2008). Approximations for F-tests which are ratios of sums of squares of independent variables with a model close to the normal. *Test* **17**, 350–369.

García-Pérez, A. (2012). A linear approximation to the power function of a test. *Metrika* **75**, 855–875.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons. New York.

1.32

Sparse robust graphical models

Hyonho Chun[1,*] Myung Hee Lee[2]

[1,*] *Purdue University; chunh@purdue.edu;*
[2] *Colorado State University; mhlee@stat.colostat.edu;*
[*] *Corresponding author*

Abstract. *A graphical model, which can be used for analyzing biological datasets of gene expressions, proteins or minerals, is an approach that describes conditional relationships among multiple variables. These conditional relationships can be conveniently identified by the zero entries of an inverse covariance matrix under the Gaussian assumption, called Gaussian graphical models (GGM)s. The GGMs are very useful, when datasets are assumed to follow Gaussian distributions. However, in the presence of outliers or unknown contaminations, the Gaussian assumption is hardly met, and one needs an approach that is robust to such deviations. For this reason, we propose a graphical model approach that is robust to the distributional assumption, and we do this via applying a set of sparse quantile regression models. This is a very natural semi-parametric extension to the GGM approach in which the least squares regression coefficients are used for finding the conditional relationships. Later, we show that the quantile regression coefficients bear information on the conditional relationships. We then demonstrate the advantages of our approach using simulation studies under various non-Gaussian scenarios and apply our method to an interesting real biological dataset, where a considerable amount of the dataset is truncated, illustrating the usefulness of our robust approach in a real setting.*

Keywords. *Robustness; Semi-parametric models; Gaussian graphical models; Contamination; Quantile Regression*

1.33

(Robust) multivariate mode estimation

T. Kirschstein[1,*], S. Liebscher [1,$], G.C. Porzio[2] and G. Ragozini[3]

[1] *Martin-Luther-University; thomas.kirschstein@wiwi.uni-halle.de, steffen.liebscher@wiwi.uni-halle.de*
[2] *University of Cassino; porzio@eco.unicas.it*
[3] *Federico II University of Naples; giragoz@unina.it*
[*] *Corresponding author*
[$] *Presenter*

Abstract. *To this day, there are only very few contributions on estimating the mode in the multivariate case. In general, there are two ways to tackle this problem, either indirectly, by searching for the maximum in the distribution density (see e.g. Abraham et al., 2003), or directly, by searching the smallest interval containing a certain proportion of points. An example of an estimator following the latter approach can be found in Sager (1979). In the multivariate case, searching for an "interval" means searching for a "body" of minimal size (usually minimum volume), which is – even today – a non-trivial task. Therefore, at the time Sager proposed his estimator its applicability was severely hampered due to the unavailability of efficient algorithms to determine a minimum volume subset.*

In this talk we present an algorithm to efficiently compute minimum volume sets, which makes Sager's estimator usable in practice. Furthermore, we discuss the choice of several parameters of the algorithm, for which up to now only rules of thumb existed and which have – as is shown – a great influence on the resulting estimator's properties (e.g., if properly chosen, Sager's estimator possesses a maximum finite sample breakdown point). Overall performance of the estimator is assessed by means of a simulation study which compares it to another direct estimator just recently proposed in Hsu and Wu (2013).

Keywords. *Convex hull; Robust location estimation; Skewed distributions; Subset selection.*

References

Abraham, C., Biau, G., and Cadre, B. (2003). Simple estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, 31(1):23–34.

Hsu, C.-Y. and Wu, T.-J. (2013). Efficient estimation of the mode of continuous multivariate data. *Computational Statistics & Data Analysis*, 63(0):148–159.

Sager, T. W. (1979). An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366):329–339.

1.34

Nonparametric robust regression estimates

V. Simakhin[1] and O. Cherepanov[2]

[1] *Kurgan State University, Kurgan, Russia; sva_full@mail.ru*
[2] *Kurgan State University, Kurgan, Russia; ocherepanov@inbox.ru*

Abstract. *The paper deals with the regression task of the form*

$$y = m(x, \Theta) + \varepsilon$$

where $m(x, \theta)$ is an known function, $\Theta = (\theta_1, \ldots, \theta_q)$ is a vector of unknown parameters. Let take $g_1(x)$ and $g_2(\varepsilon)$ as an a priory distribution density of x and ε, respectively. Robust nonparametric estimates of Θ are synthesized by the weighted maximum likelihood method (Simakhin, 2006). The ψ-functions of the estimates have the following form

$$\Psi_i(x, y, \Theta) = \left(\frac{\partial}{\partial\theta_i} g_2(y - m(x, \Theta))\right) g_2^l(y - m(x, \Theta)) g_1^l(x), i = \overline{1, \ldots, q}.$$

Parzen-Rosenblat univariate density estimator and it's modified versions were used for estimate $g_1(x)$ and $g_2(\varepsilon)$. The radical parameter l determines robust properties of the estimate. The bootstrap method has been used to estimate value of radical parameter. Results of investigations demonstrate high efficiency of proposed estimates under symmetrical and asymmetrical outliers.

Keywords. *Regression; Semiparametric; Robust; Adaptive*

References

Simakhin, V. (2006). Nonparametric robust regression estimate. *Proceedings SPIE* **vol. 6522**, 130–139.

1.35

Subsampling method for periodically stationary sequences with heavy tails and long memory

Jacek Leśkow[1], Elżbieta Gajecka-Mirek[2,*]

[1] *Institute of Mathematics, Cracow University of Technology, Krakow, Poland; jleskow@pk.edu.pl*
[2] *Department of Economics, State Higher Vocational School, Nowy Sacz, Poland; egajecka@gmail.com*
[*] *Corresponding author*

Abstract. *New and a more general concept of weak dependence introduced by Doukhan in 1999 (Doukhan, 2008) gives tools for the analysis of statistical procedures with very general data generating processes.*

One of such statistical procedures is subsampling (Politis, 1999). The advantage of subsampling is its insensitivity to the form of the asymptotic distribution. For independent data and stationary time series subsampling procedures are well investigated. Our research are focused on periodically stationary time series.

In the presentation we will introduce the model which simultaneously be dealing with three features of time series: periodic non-stationarity, heavy tails and long memory. The model investigated in the presentation can be considered as an extension of the results by Politis (2011).

Without knowledge about existence of non-degenerated asymptotic distribution for estimated parameters we can't use subsampling method. In the presentation we will investigate the joint asymptotic behavior of the vectors of sample means and the sample variances.

The main goal of the presentation is to apply subsampling method to estimate the vector of the means. Weak dependence conditions allow to achieve positive results.

As a motivation to study our model we will present real data from the European Energy Market. We will show how our model can be used to estimate periodic mean of such data set.

Keywords. *Heavy-tails; Long Memory; Weak Dependence; Periodic Correlation; Subsampling.*

References

Doukhan P., Dedecker J., Lang G., Leon J. R., Louhichi S., Prieur C. (2008). *Weak Dependence: With Examples and Applications.* Springer-Verlag.

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

Leśkow J., Lenart L., Synowiecki R. (2008). Subsampling in estimation of autocovariance for PC time series. *J. Time Ser. Anal.* **29**, 9995–1018.

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.

Politis D.N., Romano J.P., Wolf M. (1999). *Subsampling.* Springer-Verlag. New York.

1.36

A directional multivariate value at risk

Raúl Torres[1], Henry Laniado[2] and Rosa E. Lillo[3]

[1,2] *Departamento de Estadística, Universidad Carlos III de Madrid, 28911, Leganés, Spain; ratorres@est-econ.uc3m.es, henry.laniado@uc3m.es*

[3] *Departamento de Estadística, Universidad Carlos III de Madrid, 28903, Getafe, Spain; rosaelvira.lillo@uc3m.es*

Abstract. *The traditional measure of risk in an assets portfolio is the VaR (Value at Risk) due to its good properties and easy interpretation. However, only a few references are devoted* *to the generalization of this concept to the multivariate context. In this work, we introduce the definition of a multivariate financial risk measure MRVaR based on the directional extremality quantile notion recently introduced by (Laniado et al., 2012). The directions in the definition of the MRVaR can be chosen by the investor according to her/his risk preferences. We state the main properties of this MRVaR, the non-parametric estimation and a robustness analysis. We also show the advantage of using this MRVaR with respect to other multivariate VaR introduced in the recent literature.*

Keywords. *Multivariate risks; Value at risk; Extremality.*

References

Laniado, H., Lillo, R., Pellerey, F. and Romo, J. (2012). Portfolio selection through an extremality stochastic order. *Insurance: Mathematics and Economics* **51**, 1-9.

1.37

Lower bounds to the accuracy of tail index estimation

S.Y.Novak[1]

[1] *Middlesex University, London NW44BT, UK; S.Y.Novak@mdx.ac.uk*

Abstract. *We suggest a simple method of deriving minimax lower bounds to the accuracy of nonparametric statistical inference on heavy tails, and present lower bounds to the mean squared error (MSE) of tail index, tail constant and extreme quantiles estimators from a sample of heavy-tailed random variables. The results indicate that the MSE of a robust estimator depends in a specific way on the sample size, the tail index and the tail constant, revealing the corresponding information functional.*

Keywords. *Lower bounds; heavy tails.*

References

Novak S. Y. (2011) *Extreme value methods with applications to finance*. London: Chapman & Hall/CRC Press. ISBN 9781439835746

1.38

High-dimensional autocovariance matrices, linear process bootstrap and optimal linear prediction

Dimitris N. Politis[1]

[1] *University of California–San Diego, USA; dpolitis@ucsd.edu*

Abstract. *Given data $X_1, ..., X_n$ from a stationary time series, the prime objective is consistent estimation of the $n \times n$ Toeplitz autocovariance matrix. Under short range dependence conditions, convergence rates are established for a flat-top tapered version of the sample autocovariance matrix and its inverse. Two applications will be discussed in detail: (a) a new method for time series resampling, the so-called Linear Process Bootstrap, and (b) optimal linear prediction using the full sample.*

[*This talk will present results from joint work with Tim McMurry and Carsten Jentsch*]

Keywords. *Autocorrelation; Bootstrap; Prediction; Time Series.*

1.39

Bootstrapping sample quantiles of discrete data

C. Jentsch[1,*], A. Leucht[1] and T. Niebuhr[2]

[1] *Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany; cjentsch@mail.uni-mannheim.de, leucht@uni-mannheim.de*
[2] *Institut für Mathematische Stochastik, Technische Universität Braunschweig, Pockelsstraße 14, 38106 Braunschweig, Germany; t.niebuhr@tu-braunschweig.de*
[*] *Corresponding author*

Abstract. *Sample quantiles are consistent estimators for the true quantile and satisfy central limit theorems (CLTs) if the underlying distribution is continuous. If the distribution is discrete, the situation is much more delicate. In this case, sample quantiles are known to be not even consistent in general for the population quantiles. In a motivating example, we show that Efron's bootstrap is not consistent in general for sample quantiles even in the discrete independent and identically distributed (i.i.d.) data case. To overcome this bootstrap inconsistency, we provide two different and complementing strategies.*

In the first part of this paper, we prove that the i.i.d. m-out-of-n bootstrap is consistent for sample quantiles in the discrete data case. As the corresponding bootstrap confidence intervals tend to be conservative due to the discreteness of the true distribution, we propose randomization techniques to construct bootstrap confidence sets of asymptotically correct size.

In the second part, we consider a continuous modification of the cumulative distribution function and make use of mid-quantiles studied in Ma et al. (2011). Contrary to ordinary quantiles and due to continuity, mid-quantiles lose their discrete nature such that they can be estimated consistently and asymptotic normality can be achieved. We generalize the limiting results obtained in Ma et al. (2011) to the time series case. However, as the mid-quantile function fails to be differentiable, classical i.i.d. or block bootstrap methods do not lead to satisfactory results and we show that m-out-of-n variants are required here as well.

The finite sample performances of both approaches are illustrated in a simulation study by comparing coverage rates of bootstrap confidence intervals.

Keywords. *Bootstrap inconsistency; Count processes; Mid-distribution function; m-out-of-n bootstrap; Integer-valued processes.*

References

Ma, Y., Genton, M. G. and Parzen, E. (2011). Asymptotic properties of sample quantiles of discrete distributions. *Ann. Inst. Stat. Math.* **63**, 227–243.

1.40

Approximating moments by nonlinear transformations, with an application to resampling from fat-tailed distributions

K. M. Abadir[1], A. Cornea[2,*]

[1] *Imperial College London; k.m.abadir@imperial.ac.uk*
[2] *University of Exeter; A.Cornea@exeter.ac.uk*
[*] *Corresponding author*

Abstract. *We derive expansions of $E(x)$ in terms of the moments of a transformation of x, in a more general context than Taylor expansions. Apart from the intrinsic interest in such a fundamental relation that links the moments of a variate and its nonlinear transformations, our results can be used in practice to approximate $E(x)$ by the low-order moments of a transformation which can be chosen to give a good approximation for $E(x)$. We generalize the approach of bounding the terms in expansions of characteristic functions, and use it to derive an explicit and accurate bound for the remainder in the moment expansion. We illustrate one of the implications of our method by providing accurate bootstrap confidence intervals for the mean of a fat-tailed distribution with an infinite variance, in which case currently-available bootstrap methods are either asymptotically invalid or unreliable in finite sample.*

Keywords. *Expansion of functions and remainder's bound; Stable laws; Moment approximations; Bootstrap; Complex analysis.*

1.41

Sequential Subsampling and Applications

N. Steland

Institute of Statistics, RWTH Aachen University; steland@stochastik.rwth-aachen.de

Abstract. *Subsampling is a resampling technique that has been shown to be consistent for a wide class of time series under weak assumptions, see Politis, Romano and Wolf (1999). By drawing on general consistency theorems obtained for subsampling, one can adopt the methodology to subsample sequential decision procedures leading to sequential subsampling. We discuss this issue, its application to inferential procedures with complex limiting, and illustrate it by analyzing real data streams, see Rafajłowicz and Steland (2014)*

Keywords. *Change-point; Invariance Principle, Keyword2. Include up to five keywords separated by semi–colons starting with capital letters.*

References

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer. New York.

Rafajłowicz, E.R. and Steland, A. (2014). Decoupling change-point detection based on characteristic functions: Methodology, asymptotics, subsampling and application. *Journal of Statistical Planning and Inference*, in press.

1.42

Regularization procedure of signal estimation in multiplicative models

A. Dobrovidov[1,*] and L. Markovich[1]

[1] *Institute of Control Sciences of Russian Academy of Sciences; dobrovidov@gmail.com, dobrovid@ipu.ru*

Abstract. *At the previous ISNPS conference (Halkidiki'2012) the report was presented where the problem of filtering of signal with an unknown distribution from the mixture with a noise had been solved by using nonparametric kernel techniques (Dobrovidov et al., 1012). In this novel paper a nonlinear multiplicative observation model with non-gaussian signal and noise is considered. The main feature of the problem is that the support of the distributions entering in the evaluation functional is a positive semi-axis. Therefore, the classical methods of nonparametric estimation with symmetric kernels are not applicable because of great estimate bias. We have to apply asymmetric kernel functions of the gamma-kernel type Song Xi Chen*

(2000). The equation of the optimal filtering contains a statistic in the form of a logarithmic derivative of a multidimensional density of observations. The convergence theorem for non-parametric estimate of the logarithmic derivative by dependent data is proved. Moreover, we build a data-driven bandwidths for the gamma-kernel and its derivative. To have a stable estimates we propose a regularization procedure with data-driven optimal regularization parameter. Such approach leads to an automatic algorithm of non-parametric filtration in multiplicative observation models. Similar filtering algorithms can be used, for instance, in problems of volatility estimation in models of financial mathematics.

Keywords. *Problem of filtering;Multiplicative observation model;Nonparametric estimates;Regularization*

References

Dobrovidov, A. V., Koshkin, G.M. and Vasiliev, V.A.(2012). *Non-parametric state space models* Kendrick press, USA.

Song Xi Chen(2000). Probability density function estimation using gamma kernels.*Ann. Inst. Statist. Math.* **54**, 471-480.

1.43

Kalman filter using nonparametric functional estimators

G. Koshkin[1] and V. Smagin[2]

[1] *Tomsk State University, Tomsk, Russia; kgm@mail.tsu.ru,*
[2] *Tomsk State University, Tomsk, Russia; vsm@mail.tsu.ru*

Abstract. *The paper deals with the Kalman filtering algorithm for a class of systems with unknown additive inputs. Such classes include object models with possible failures, and also the models of controlled processes with unknown disturbances. The known methods of calculating estimates of the state vector are based on algorithms that use the estimators of an unknown perturbation (see Astrom and Eykhoff (1971)–Hsien (2010)). In Astrom and Eykhoff (1971), there are considered the extension algorithms of the states space requiring complete information on the model of this input. In this paper, as in Gillijns and Moor (2007)–Hsien (2010), additional information on the models of an unknown input is not required. For a discrete object with an unknown input, we propose the modification of the Kalman filtering algorithm, using the nonparametric estimators of distributions functionals of the observed process according to Dobrovidov* et al. *(2012). An example, illustrating the effectiveness of the proposed algorithm, is given in comparison with the known algorithms from Gillijns and Moor (2007)–Hsien (2010).*

Keywords. *Kalman filter; Nonparametric estimator; Unknown input.*

Research supported by the Russian Foundation for Basic Research, grants no. 13-08-00744, 13-08-98027, and 13-08-01015.

References

Astrom, K. and Eykhoff, P. (1971). System identification — A survey. *Automatica* **7**, 123–162.

Gillijns, S. and Moor, B. (2007). Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica* **43**, 111–116.

Smagin, S.V. (2009). Filtering in linear discrete systems with unknown perturbartion. *Optoelectronic, Instrumentation and Data Processing* **45**, **6**, 513–519.

Hsien, C.-S. (2010). On the optimal of two-stage Kalman filter for systems with unknown input. *Asian Journal of Control* **12**, **4**, 510–523.

A. Dobrovidov, G. Koshkin and V. Vasiliev (2012). *Non-Parametric State Space Models.* Heber, UT 84032, USA. Kendrick Press, Inc.

1.44

Estimating a distribution function for censored and dependent data

Dimitris Ioannides [1]

[1] *University of Macedonia-Thessaloniki-Greece, dimioan@uom.gr.*

Abstract. *In some studies a series of dependent and censored failures times are observed. The failure times have a common marginal distribution function and a kernel estimator for it can be constructed. Under some regularity conditions the asymptotic normality of our proposed estimator is obtained. Point wise confidence intervals for the survival function is developed.*

Keywords. *Distribution function; Censored data; Dependent data.*

References

Bagkavos, D. and Ioannides, D. (2012). Smooth Confidence Intervals for the Survival function under right censoring. *Electron. J. Statist.* **6**, 843-860.

1.45

Measures of asymmetry based on a necessary/necessary and sufficient conditions for symmetry

D. Bagkavos[1,*], P. N. Patil[2] and A.T.A. Wood[3]

[1] *Accenture, Greece; dimitrios.bagkavos@gmail.com,* [2] *Mississippi State University, U.S.A,* [3] *University of Nottingham, U.K.*
[*] *Corresponding author*

Abstract. *It is common practice to make assertions about the symmetry or asymmetry of a probability density function based on coefficients of skewness. Since most coefficients of skewness are designed to be zero for a symmetric density, they do, overall, provide an indication of symmetry. However, skewness, as opposed to asymmetry, is primarily influenced by the tail behavior of a density function. Therefore, coefficients of skewness do not reliably calibrate asymmetry in the density curve. Here are presented two measures of symmetry (a weak and a strong one) based on necessary / necessary and sufficient condition for a continuous probability density function to be symmetric. It results to a coefficient of asymmetry, for a continuous probability density function on the scale of -1 to 1. We show through examples that the proposed measures do an admirable job of capturing the visual impression of asymmetry of a continuous density function. Further, we discuss when to use which as well as their implementation in practice and conclude by real world dataset illustrations.*

Keywords. *Asymmetry measure; Correlation; Nonparametric; Skewness*

1.46

Scalable machine learning on massive social networks

Qirong Ho[1], Eric P. Xing[1], Junming Yin[1]

[1] Carnegie Mellon University; qho+@cs.cmu.edu, epxing@cs.cmu.edu, junmingy@andrew.cmu.edu

Abstract. *Today's social, communication, and WWW networks easily contain millions or even billions of nodes, and copious amounts of side information (context) such as text, attribute, temporal, image and video data. A thorough analysis of such a network must consider both the graph and the associated side information, with inference algorithms that can cope with the size of such data. Much of the existing work on statistical network modeling and inference remain limited to small-scale problems unsuitable for realistic application. Towards the goal of rich analysis on societal-scale networks, we develop new strategies for network representation, statistical modeling, inference algorithm, and distributed multi-machine programming that, together, ensure massive scalability to large networks. In this talk, I will present some recent results from these efforts, and demonstrate how a mixed-membership model formalism is effectively enabled to distill community structure and personal interests from a networks with over 100 million nodes on just a few cluster machines in a few hours.*

Keywords. Statistical Network Model; Machine Learning; Parallel Inference; Social Network; Mixed Membership Model.

References

Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*(2):251–276.

Bottou, L. (2004). Stochastic learning. *Advanced Lectures on Machine Learning*, 146–168.

Carman M., Crestani, F., Harvey, M., & Baillie, M. (2010). Towards query log based personalization using topic models. In proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10), 1849–1852.

Gopalan, P, Mimno, D., Gerrish, S., Freedman, M. & Blei, D. (2012). Scalable inference of overlapping communities. *Advances in Neural Information Processing Systems* **25**, 2258–2266.

Granovetter, M. The strength of weak ties. (1973). *American Journal of Sociology*, **78**(6):1360–1380.

Ho, Q., Parikh, A. & Xing, E. (2012). A multiscale community blockmodel for network exploration. *Journal of the American Statistical Association*, **107**(499).

Ho, Q., Yin, J., & Xing, E. (2012). On triangular versus edge representations — towards scalable

1.47

Pseudo-likelihood methods for community detection in large sparse networks

Arash A. Amini[1], Aiyou Chen[2], Peter J. Bickel[3] and Elizaveta Levina[1]

[1] *Department of Statistics, University of Michigan; aaamini@umich.edu, elevina@umich.edu*
[2] *Google, Inc; aiyouchen@google.com*
[3] *Department of Statistics, University of California, Berkeley; bickel@stat.berkeley.edu*

Abstract. *We consider the problem of community detection in a network, that is, partitioning the nodes into groups that, in some sense, reveal the structure of the network. Many algorithms have been proposed for fitting network models with communities, but most of them do not scale well to large networks, and often fail on sparse networks. We present a fast pseudo-likelihood method for fitting the stochastic block model, a well-known model for networks with communities, as well as a variant that allows for an arbitrary degree distribution by conditioning on degrees. We provide empirical results showing that the algorithms perform well under a range of settings, including on very sparse networks, and illustrate on the example of a network of political blogs. We also present spectral clustering with perturbations, a method of independent interest, which works well on sparse networks where regular spectral clustering fails, and use it to provide an initial value for pseudo-likelihood. We discuss theoretical results showing that pseudo-likelihood provides consistent estimates of the communities under mild conditions on the starting value, for the case of a block model with two communities. Time permitting, we give some insights as to why perturbations help with spectral clustering on sparse networks.*

Keywords. *community detection; network; pseudo-likelihood.*

1.48

Universality of the stochastic blockmodel

Sofia C. Olhede[1,*] & Patrick J. Wolfe[1]

[1] *Department of Statistical Science, University College London; s.olhede@ucl.ac.uk, p.wolfe@ucl.ac.uk*
[*] *Corresponding author*

Abstract. *The stochastic blockmodel has become ubiquitous in statistical network analysis. We discuss recent results showing that for a large class of network models the blockmodel can be considered a universal representation even under model misspecification, like a histogram (Wolfe, 2013). Blocks of edges play the role of histogram bins, and community sizes that of histogram bandwidths or bin sizes. Just as standard histograms allow for varying bandwidths, different blockmodel estimates can all be considered valid representations of an underlying probability model, subject to bandwidth constraints.*
We show that under these constraints, the mean integrated square error of the network histogram tends to zero as the network grows large, and we provide methods for optimal bandwidth selection–thus making the blockmodel a universal representation, see Olhede (2013). With this insight, we discuss the interpretation of network communities in light of the fact that many different community assignments can all give an equally valid representation of the network. To demonstrate the fidelity-versus-interpretability tradeoff inherent in considering different numbers and sizes of communities, we show an example of detecting and describing new network community microstructure in political weblog data.

Keywords. *Community detection, exchangeable random graphs, graphons, statistical network analysis, stochastic block-models*

References

Wolfe, P. J. & Olhede, S. C. (2013). Nonparametric Graphon Estimation. *arXiv:1309.5936*, technical report.

Olhede, S. C. & Wolfe, P. J.(2013). Network histograms and universality of blockmodel approximation. *arXiv:1309.5936*, technical report.

1.49

Estimating latent variable densities for exchangable network models

Sharmodeep Bhattacharyya[1,*], Peter J. Bickel[1] and Patrick J. Wolfe[2]

[1] *University of California, Berkeley; sharmo@stat.berkeley.edu, bickel@stat.berkeley.edu*
[2] *University College London; p.wolfe@ucl.ac.uk*
[*] *Corresponding author*

Abstract. *Exchangeable network models provide a general non-parametric class of models for unlabeled random graphs. The main component of the exchangeable network models is latent variable density or graphon equivalence class as defined in Lovasz and Szegedy (2006). Recently there has been some focus in statistics on estimation of latent variable densities and its use as an exploratory tool for network data analysis (Wolfe and Olhede (2013), Airoldi et.al. (2013), Latouche and Robin (2013)). We propose an unified framework under which the exchangeable network models are approximated by stochastic block models and latent variable densities are estimated using piece-wise constant or histogram-type estimators derived from the fitted block models. We show that if the latent variable density has some smoothness properties and the fitted block models, where the fitting can be done by any method including spectral method, variational or profile likelihood, satisfy certain consistency properties, then we can have consistent estimators for the latent variable densities under suitable metrics. We derive rates of convergence for the fitted block models estimating the latent variable density for appropriate metrics. We also propose a cross-validation method using subgraph counts and their smooth functions to choose the size of the block model approximating the latent variable density. A simulation study and illustration on real networks will also be provided.*

Keywords. *Statistical Network Analysis; Exchangeable Network Models; Graphon; Spectral Methods; Stochastic Block Models*

References

Airoldi, Edoardo M and Costa, Thiago B and Chan, Stanley H (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems* , 692–700.

Latouche, Pierre and Robin, Stéphane (2013). Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a w-graph model. *arXiv preprint arXiv:1310.6150.*

Lovász, László and Szegedy, Balázs (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, **96:6** , 933–957.

Wolfe, Patrick J and Olhede, Sofia C (2013). Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936* .

1.50

Dynamic of communities in social networks

Emmanuel Viennet

Université Paris 13, Sorbonne Paris Cité, L2TI, F-93430, Villetaneuse, France
emmanuel.viennet@univ-paris13.fr

Abstract. *Modern online social network platforms generate huge amount of data and pose new challenges to data mining techniques. This kind of data is heterogeneous, mixing texts, images, videos and relational links (interactions between users, friends relationships). Analyzing efficiently this data is important for a lot of applications, e.g. social network community management, recommendation systems, marketing. One problem of particular importance is the study of communities in the networks, that is, subsets of nodes with strong interactions. A lot of work has been devoted to the design and analysis of community detection algorithms(Fortunato, 2010; Ngonmang et al., 2012). However, several important problems are not yet solved satisfactorily, such as taking in account simultaneously the relations and the attributes(Dang and Viennet, 2012) or the network's dynamic.*

Until recently, most research on community detection in complex network have only considered static networks: a snapshot of the network is taken at a particular time and the communities are computed on the constructed graph.

In this communication, we present some recent research in community detection for dynamic networks. We define the community prediction problem: knowing the evolution of the network until the time-step t, can we predict the communities at the time-step $t + 1$? We propose a general approach for communities prediction based on a machine learning model predicting interaction in social networks. We show that simple models allows to predict the behavior of the users. Evaluations on academic datasets (DBLP, Facebook Walls) and on a real world application are presented.

Keywords. *Social Networks, Communities*

References

Dang, T. A. and Viennet, E. (2012). Community detection based on structural and attribute similarities. In *International Conference on Digital Society (ICDS)*, pages 7–14. ISBN: 978-1-61208-176-2.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Ngonmang, B., Tchuente, M., and Viennet, E. (2012). Local communities identification in social networks. *Parallel Processing Letters*, 22(1). 16 pages.

1.51

Nonparametrics for network degree structure

Sofia C. Olhede[1,*] & Patrick J. Wolfe[1]

[1] *Department of Statistical Science, University College London; s.olhede@ucl.ac.uk, p.wolfe@ucl.ac.uk*
[*] *Corresponding author*

Abstract. *We derive the sampling properties of random networks based on weights whose pairwise products parameterize independent Bernoulli trials. This enables an understanding of many degree-based network models, in which the structure of realized networks is governed by properties of their degree sequences. We provide exact results and large-sample approximations for power-law networks and other more general forms. This enables us to quantify sampling variability both within and across network populations, and to characterize the limiting extremes of variation achievable through such models. Our results highlight that variation explained through expected degree structure need not be attributed to more complicated generative mechanisms.*

Keywords. *Exchangeable random graphs, graphons, network motif, statistical network analysis*

References

Olhede, S. C. & Wolfe, P. J. (2013). Degree-based network models. *arXiv:1211.6537*, technical report.

Wolfe, P. J. & Olhede, S. C. (2013). Nonparametric Graphon Estimation. *arXiv:1309.5936*, technical report.

1.52

“Patchwork” bootstrapt for inference on random networks

Yulia R. Gel[1,*], Vyacheslav Lyubchich[2] and Lilia Leticia Ramirez Ramirez[3]

[1] *University of Texas at Dallas, USA and University of Waterloo, Canada; ygl@utdallas.edu*
[2] *University of Waterloo, Canada; vlyubchi@uwaterloo.ca*
[3] *ITAM, Mexico; lilialeticia.ramirez@itam.mx*
[*] *Corresponding author*

Abstract. *In this talk we discuss an alternative new nonparametric “patchwork” resampling approach to network inference that may be viewed as an adaptation of block bootstrap for time series (Politis and Romano (1993)) and re-tiling for spatial data (Hall et al. (1986)) to random* *networks. We focus on uncertainty quantification for network mean degree using a "patchwork" nonparametric bootstrap, under the assumption that network is exchangeable but its degree distribution and order are unknown. We develop a data-driven crossvalidation methodology for selecting an optimal "patch" size. This "patchwork" bootstrap methodology further extends the ideas developed by Thompson et al. (2014). However, the new method is shown to be substantially less computationally expensive and more robust to unknown or misspecified network order. We illustrate the new "patchwork" bootstrap procedure by simulations and applications to airline alliances and Wikipedia activity data.*

Keywords. *Random networks; nonparametric resampling; block bootstrap; re-tiling; snowball sampling*

References

Hall, P., Horowitz, J. L. and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82**, 561–574.

Politis, D.N. and Romano, J.P. (1993). The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Thompson, M.E., Ramirez Ramirez, L. L., Lyubchich, V., and Gel, Y.R. (2014). Using bootstrap for statistical inference on random graphs. *Submitted.*

1.53

A non-parametric approach to random graph models with local dependence

Michael Schweinberger[1,*], Mark S. Handcock[2]

[1] *Department of Statistics, Rice University, Houston, TX, USA; michael.schweinberger@rice.edu*
[2] *Department of Statistics, University of California, Los Angeles, CA, USA; handcock@stat.ucla.edu*
[*] *Corresponding author*

Abstract. *Dependent phenomena, such as relational, spatial, and temporal phenomena, tend to be characterized by local dependence in the sense that units which are close in a well-defined sense are dependent. In contrast to spatial and temporal phenomena, however, relational phenomena tend to lack a natural neighborhood structure in the sense that it is unknown which units are close and thus dependent. Owing to the challenge of characterizing local dependence and constructing random graph models with local dependence, many conventional exponential-family random graph models induce strong dependence and are not amenable to statistical inference. We take first steps to characterize local dependence in random graph models, inspired by the notion of finite neighborhoods in spatial statistics and M-dependence in time series, and show that local dependence endows random graph models with desirable properties which make them amenable to statistical inference. We discuss statistical inference both when neighborhood structure is observed and when it is unobserved. In the absence of observed neighborhood structure, we take a Bayesian view and express the uncertainty about the neighborhood structure by specifying a prior on a set of suitable neighborhood structures based on Dirichlet process priors. We* *present simulation results and applications to two real-world networks with ground truth.*

Keywords. *Random graph models; Exponential families; Dirichlet processes; Local Dependence; Weak Dependence.*

1.54

Identifying skeleton curves in noisy data: a Bayesian approach to principal curves

H. Jankowski[1], L. Stanberry[2]

[1] *Department of Mathematics and Statistics, York University; hkj@yorku.ca*
[2] *Seattle Children's Research Institute; larissa.stanberry@seattlechildrens.org*

Abstract. *We present a nonparametric Bayesian principal curve method to reconstruct the centerline in noisy data. We illustrate the method on simulated two and three dimensional data and also apply it to recover the centerline of the colon in a single photon emission computed tomography image.*

1.55

Recent progress in nonparametric curve estimation with group testing data

Aurore Delaigle[1]

[1] *University of Melbourne*

Abstract. *Group testing is a method employed when collecting data on a Bernoulli variable Y, where, instead of observing the value of Y for each individual in a sample, the individuals are pooled in J groups of sizes $n_1, \ldots, n_J$; and only the maximum of the Y-values of the individuals within each group is observed. More specifically, let Y_{ij} denote the value of Y for the ith individual in the jth group. In the group testing setting, instead of observing Y_{ij} $(i = 1, \ldots, n_j;$ $j = 1, \ldots, J)$, we observe*

$$Y_j^* = \max_{i=1,\ldots,n_j} Y_{ij} \quad (j = 1, \ldots, J)\,.$$

This technique was originally introduced in infectious disease studies, to reduce the cost and increase the speed of data collection. Often, Y is the result of a blood or urine test, typically a test for an infectious disease, and $Y = I(\text{test is positive})$, where I denotes the indicator function. There, $Y_j^* = 1$ if one or more individuals within the jth group test positive. We consider new developments for nonparametric estimation of the curve $m(x) = P(Y = 1|X = x)$, where X is a covariate or a vector of covariates.

Keywords. *Bandwidth; Grouped data; Local polynomial estimator; Pooled data.*

1.56

Nonparametric optimal tests for structure detection in multivariate data

G. Claeskens[1*], J.-M. Freyermuth[1], F. Autin[2] and C. Pouet[2]

[1] ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium; gerda.claeskens@kuleuven.be, Jean-Marc.Freyermuth@kuleuven.be
[2] Université d'Aix-Marseille 1 - L.A.T.P. 39, rue F. Joliot Curie, 13453 Marseille Cedex 13, France; autin@cmi.univ-mrs.fr, pouet@cmi.univ-mrs.fr
* Corresponding author

Abstract. *We construct hypothesis tests for a multivariate (d-variate) nonparametric regression model. A test on the atomic dimension, δ, of the function, that is, the maximal degree of interaction between the d variables, has multiple interesting applications such as testing for additivity, meaning δ equal to one. Testing for model simplification by leaving out a variable is considered too. The test statistics are based on hyperbolic wavelet coefficients. Such tests are near optimal for detecting functions from anisotropic smoothness classes. Simulations and a test on fMRI data show the tests' behaviour.*

Keywords. *Anisotropy; atomic dimension; hyperbolic wavelets; hypothesis test; minimax rate.*

1.57

Multivariate plug-in bandwidth selection for density and density derivative estimation

J.E. Chacón[1,*], T. Duong[2] and M.P. Wand[3]

[1] Departamento de Matemáticas, Universidad de Extremadura; jechacon@unex.es
[2] LSTA, Université Pierre et Marie Curie – Paris 6; tarn.duong@gmail.com
[3] School of Mathematical Sciences, University of Technology; Matt.Wand@uts.edu.au
* Corresponding author

Abstract. *An abstract of 250 words or less must be included, summarizing the purpose, methodology, results and conclusions. Formulas should not be included in the abstract, but references are allowed if required. References should be included as Ehrenberg (1982) or as (Lamport, 1986).*

Keywords. *Keyword1; Keyword2. Include up to five keywords separated by semi–colons starting with capital letters.*

References

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

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

1.58

Partial distance covariance

Gábor J. Székely[1] and Maria L. Rizzo[1]

[1] *National Science Foundation, Alexandria, VA USA*
[2] *Bowling Green State University, Bowling Green, OH USA; mrizzo@bgsu.edu*
[*] *Corresponding author*

Abstract. *Distance covariance and distance correlation are scalar coefficients that characterize independence of random vectors in arbitrary dimension. Properties, extensions, and applications of distance correlation have been discussed in the recent literature, but the problem of defining the partial distance correlation has remained an open question of considerable interest. The problem of partial distance correlation is more complex than partial correlation partly because the squared distance covariance is not an inner product in the usual linear space. For the definition of partial distance correlation we introduce a new Hilbert space where the squared distance covariance is the inner product. After presenting a brief overview of distance covariance, we define the partial distance correlation statistics, population coefficients, and methods for inference. Our intermediate results also provide an unbiased estimator of squared distance covariance, and a neat solution to the problem of distance correlation for dissimilarities rather than distances.*

Keywords. *Nonlinear dependence, Multivariate independence, Partial distance correlation; Energy statistics; Dissimilarities*

1.59

A fourier analysis of extremal events

Y. Zhao[1,*] and T. Mikosch[2]

[1] *Ulm University; zywmar@gmail.com*
[2] *University of Copenhagen; mikosch@math.ku.dk*
[*] *Corresponding author*

Abstract. *The extremogram is an asymptotic correlogram for extreme events constructed from a regularly varying strictly stationary sequence. Correspondingly, the spectral density generated from the extremogram is introduced as a frequency domain analog of the extremogram. Its empirical estimator is the* **extremal periodogram***. The extremal periodogram shares numerous asymptotic properties with the periodogram of a linear process in classical time series analysis: the asymptotic distribution of the periodogram ordinates at the Fourier frequencies have a similar form and smoothed versions of the periodogram are consistent estimators of the spectral density. By proving a functional central limit theorem, the* **integrated extremal periodogram** *can be used for constructing asymptotic tests for the hypothesis that the data come from a strictly stationary sequence with a given extremogram or extremal spectral density. A numerical method, the* **stationary bootstrap***, can be applied to the estimation of the integrated extremal periodogram.*

Keywords. *Extremogram; Regular variation; Periodogram; Stationary Bootstrap*

1.60

Multivariate density estimation by local Gaussian approximations

H. Otneim[1*] and D. Tjøstheim[1]

[1] *University of Bergen, Norway; hakon.otneim@math.uib.no, dag.tjostheim@math.uib.no*
[*] *Corresponding author*

Abstract. *The curse of dimensionality precludes the nonparametric kernel density estimator when we have moderate or small sample sizes and the dimension of our data exceeds three or four. We can counter this problem by imposing some simplifying structure on the underlying density, but without specifying a full parametric model such as the Gaussian distribution. Indeed, the multivariate Gaussian distribution is particularly easy to fit to data since the expectations and variances are estimated from the corresponding marginal observation vectors, and the correlations are estimated using each corresponding pair. The curse of dimensionality is therefore not an issue, but the risk of model misspecification certainly is. We introduce the locally Gaussian multivariate density estimator which fits the multivariate Gaussian distribution using local likelihood. In order to avoid the curse of dimensionality we extend the simple global estimation to the local case, meaning that we estimate the marginal parameter functions based on their corresponding data vector, and the correlation functions based on the corresponding data pairs respectively. The resulting density estimate is much more flexible than the global parametric fit and remains unaffected by the curse of dimensionality. The risk of misspecification is hard to quantify, but promising simulation results will be presented along with theory and examples.*

Keywords. *Local Likelihood; Local Gaussian; Multivariate Density Estimation; Curse of Dimensionality.*

1.61

Measuring dependence with local Gaussian correlation

Dag Tjøstheim

Abstract. *The Pearson correlation is the most used dependence measure in statistics. It has several weaknesses, and really only works very well for Gaussian variables. In this talk I introduce a local Gaussian correlation by approximating a bivariate density locally by a bivariate Gaussian density. The correlation coefficient of the approximating Gaussian is taken as the local Gaussian correlation. I will give some theoretical properties of this dependence measure and present a number of applications to independence testing, copula description and recognition, financial contagion, and measuring asymmetry of financial returns.*

References

Tjøstheim and Hufthammer, J. Econometrics (2013)

Berentsen and Tjøstheim, Statistics and Computing (2014)

Berentsen, Støve, Tjøstheim and Nordbø (2014)

Støve, Tjøstheim and Hufthammer (2014), J. Empirical Finance

Støve and Tjøstheim, in Nonlinear Time Series Econometrics, Oxford (2014)

Berentsen, Kleppe and Tjøstheim (2014), J. Statistical Software.

1.62

A difference based approach to the semiparametric partial linear multivariate model

N1. Lawrence D. Brown1[1], N2. Michael Levine[2,*] and N3. Lie Wang [3]

[1] *Department of Statistics, University of Pennsylvania, lbrown@wharton.upenn.edu*
[2] *Department of Statistics, Purdue University mlevins@purdue.edu*
[3] *Department of Mathematics, MIT liewang@math.mit.edu*
[*] *Corresponding Author*

Abstract. *A semiparametric partial linear model with the nonparametric component defined on a multivariate Euclidean space is considered. Compared to the regular semiparametric partial linear model, this case has not received a lot of attention in the literature. We estimate the linear component of the model using a difference based approach. The estimator of the* *nonparametric component is then constructed using residuals as transformed data. Both the estimator of the linear component and the estimator of the nonparametric component asymptotically perform as well as if the other component were known. Moreover, the estimator of the linear component is asymptotically efficient while the estimator of the nonparametric component is asymptotically rate optimal. A test for linear combinations of the regression coefficients of the linear component is also developed. All of the procedures considered are easy to implement. Numerical performance of the procedure is studied using the simulated data. This is joint work with Lawrence D. Brown and Lie Wang.*

Keywords. *Difference-based method; semiparametric partial linear multivariate model; asymptotic efficiency; asymptotic optimality; linear component testing.*

1.63

A conditional empirical likelihood approach to combine sampling design and population level information

Sanjay Chaudhuri[1,*], Mark Handcock[2] and Michael Rendall[3]

[1] *Department of Statistics and Applied Probability, National University of Singapore, Singapore; sanjay@stat.nus.edu.sg,*
[2] *Department of Statistics, University of California, Los Angeles ; handcock@ucla.edu*
[2] *Department of Sociology, University of Maryland, College Park; mrendall@umd.edu*
[*] *Corresponding author*

Abstract. *Inclusion of available population level information in statistical modelling is known to produce more accurate estimates than those obtained only from the random samples. However, a fully parametric model which incorporates both these information may be computationally challenging to handle. Empirical likelihood based methods can be used to combine these two kinds of information and estimate the model parameters in a computationally efficient way. In this article we consider methods to include sampling weights in an empirical likelihood based estimation procedure to augment population level information in sample-based statistical modeling. Our estimator uses conditional weights and is able to incorporate covariate information both through the weights and the usual estimating equations. We show that under usual assumptions, with population size increasing unbounded, the estimates are strongly consistent, asymptotically unbiased and normally distributed. Moreover, they are more efficient than other probability weighted analogues. Our framework provides additional justification for inverse probability weighted score estimators in terms of conditional empirical likelihood. We give an application to demographic hazard modeling by combining birth registration data with panel survey data to estimate annual first birth probabilities.*

Keywords. *Empirical likelihood; Complex surveys; Sampling design; Population level information*

1.64

Resampling assessment of the scale of geophysical and environmental process models

Mark S. Kaiser, Daniel J. Nordman*

Department of Statistics, Iowa State University; mskaiser@iastate.edu, dnordman@iastate.edu
* *Corresponding author*

Abstract. *Models of geophysical or environmental processes inherently involve a concept of scale, either temporal or spatial or both. Although the concept of scale is pervasive, there have been few attempts to define or quantify it, and scientists in specific disciplines may have quite different notions of what the term means. For empirical stochastic models, scale often relates to the resolution at which data are available. For deterministic models, scale may be related to the extent of time/space needed for a process model to cover a range of dynamics allowed by the model. A probabilistic definition of a "structured process" is introduced, based on a concept of scale at which a process may vary. A diagnostic quantity, based on the block bootstrap, is then proposed as a tool for quantifying the scale of such processes, where the block bootstrap is applied to reconstruct empirical distributions. Often, the block bootstrap is known to be asymptotically valid over a range of block lengths. However, for scale-structured processes, this is not the case. In fact, the block bootstrap re-creations exhibit a maximal level of variability when the block length used matches the scale of the underling process. Hence, the bootstrap has certain asymptotic behaviors that make it useful for determining when resampling blocks have become large enough to preserve the dynamic structure of the process model, which quantifies our concept of scale. Use of the diagnostic is motivated though simulation and used to compare regional climate models.*

Keywords. *Block Bootstrap; Environmental Statistics; Process Structure; Regional Climate Models.*

1.65

On statistical testing for spatio-temporal stationary random fields

Y. Yajima[1] and Y. Matsuda[2]

[1] *Univeristy of Tokyo; yajima@e.u-tokyo.ac.jp*
[2] *Tohoku University matsuda@econ.tohoku.ac.jp*

Abstract. *We propose a new test statistic for spatio-temporal stationary random fields and derive its asymptotic properties. The observations are given at irregularly spaced sites. The test statistic is based on discrepancy between parametric and nonparametric estimators of spectral density functions. One advantage of this test statistic is that its asymptotic expectation and* *variance depend on only the kernel function and the bandwidth of the nonparametric estimator.*

Keywords. *Statistical testing; Spatio-temporal stationary random fields.*

1.66

Detecting outlying behavior of curves

M. Hubert[1]

[1] *Mathematics Department and LStat, KU Leuven, Celestijnenlaan 200B, BE-3001 Leuven, Belgium; Mia.Hubert@wis.kuleuven.be*

Abstract.
Depth functions are statistical tools, used to attribute a sensible ordering to observations in a sample from the center outwards. Recently several depth functions have been proposed for functional data. These depth functions can for example be used for robust classification and for the detection of outlying curves. We present several diagnostic tools to study the outlying behavior of curves, based on the multivariate functional halfspace depth introduced in Claeskens (2014). It is illustrated how this depth function can be useful to detect globally outlying curves as well as curves that are only outlying on parts of their domain. Several graphical representations of the curves and their degree of outlyingness are presented.

Keywords. *Functional depth; Outlying curves.*

References

Claeskens, G., Hubert, M., Slaets, L., Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association* **109**, 411–423.

Hubert, M., Claeskens, G., De Ketelaere, B., Vakili, K. (2012). A new depth-based approach for detecting outlying curves. *Proceedings of COMPSTAT 2012, edited by A. Colubi, K. Fokianos, G. Gonzalez-Rodriguez, E.J. Kontoghiorghes,* 329–340.

1.67

Extending DD-classifier on the functional setting

J.A. Cuesta–Albertos[1], M. Febrero–Bande[2,*] and M. Oviedo de la Fuente[2]

[1] *Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain; juan.cuesta@unican.es*
[2] *Departamento de Estadística e Investigación Operativa, Universidade de Santiago de Compostela, Spain; manuel.febrero@usc.es, manuel.oviedo@usc.es*
[*] *Corresponding author*

Abstract. *The DD–plot is a tool based on data depth that was introduced in Liu* et al. *(1999) for the graphical comparison of two multivariate distributions or groups. Roughly speaking, a DD–plot is a two dimensional graph where the pair* $(D_1(x), D_2(x))$ *is plotted, being* $D_i(x)$*, the depth of element* x *respect to the i-th group. Using this graph and the data-depths, Li* et al. *(2012) develop a nonparametric classifier much better than the maximum depth principle but with several limitations. The objective of this paper is threefold: first to extend the DD–classifier to a number of groups greater than two, second to apply regular classification methods to DD–plots and third, to obtain useful insights about data–depths using the diagnostics of these classification methods. Along this paper different notions of functional data–depth are revised (and enlarged) at the same time that the new proposal, called* DD^h*–classifier, is developed. The paper is completed with a simulation study and the application to classical datasets in the functional context.*

Keywords. *Functional Data depth; DD–Classifier; DD–plot.*

References

Li, J., Cuesta–Albertos, J.A. and Liu, R. (2012). DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. *J. Amer. Statist. Assoc.* **107**, 737–753.

Liu, R., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27**, 783–840.

1.68

A penalized PLS approach in functional regression with ordinal response

Manuel Escabias[1,*], Ana M. Aguilera[1] and Carmen Aguilera-Morillo[2]

[1] *University of Granada; escabias@ugr.es, aaguiler@ugr.es*
[2] *University Carlos III; maguiler@est-econ.uc3m.es*
[*] *Corresponding author*

Abstract. *In this paper authors give a new step in their research line in functional regression models with discrete response, that began with the functional logit regression model (Escabias et al., 2004), recently grew up with one proposal for the estimation of the functional base-line logit models (Escabias et al., 2014) and now board the functional logit model for ordinal response. During these years different estimation proposals have been carried out for these models, all of them based on basis expansion methods. These proposals have tried to solve the multicollinearity problems that usually arise in these models by using functional principal component analysis as in Escabias et al. (2005) or Escabias et al. (2014), PLS regression as in Escabias et al. (2007) and more recently different forms of penalized methods and PLS-based methods (Aguilera-Morillo et al., 2013). Now we propose to combine penalized and PLS-based methods for the estimation of the functional regression model with ordinal response. In order to formulate the functional model for ordinal response, let us consider a set of curves and an ordinal response variable with S categories, then the most popular logit model for ordinal responses is the proportional odds model, where the logits are expressed in terms of predictor* *curves as a functional linear span of curves with the same slope functional parameter for all $S-1$ logits. These logits are defined as the logarithm of the quotient between the distribution and survival functions evaluated at each one of the first $S-1$ categories of the response. After considering that the predictor curves and parameter function belong to a finite space generated by a basis of functions the functional model turns to a multiple one. The ML estimation of this model is affected by high multicollinearity what makes the functional parameter to be unsmooth (Aguilera et al., 2006 and Escabias et al., 2004). In this work we use P-spline penalization of the curves to avoid this lack of smothness.*

Keywords. *Functional Regression; Penalized PLS; Ordinal Response.*

References

Aguilera A.M., Escabias M. and Valderrama M.J. (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics and Data Analysis 50(8):1905-1924

Escabias M., Aguilera A.M. and Valderrama M.J. (2004) Principal component estimation of functional logistic regression: discussion of two different approaches. Journal of Nonparametric Statistics 16(3-4): 365-384

1.69

Multiple-output functional quantile regression

Davy Paindaveine[1,*] and Germain Van Bever[2]

[1] *Université libre de Bruxelles; dpaindav@ulb.ac.be*
[2] *Université libre de Bruxelles; gvbever@ulb.ac.be*
[*] *Corresponding author*

Abstract. *We consider a multiple-output functional regression problem where the response is a random m-vector Y and the covariate $X(t)$ is of a functional nature. In that framework, we define a concept of directional regression quantile, which extends the finite-dimensional concept from Hallin, Paindaveine and Šiman (2010) and Hallin, Lu, Paindaveine and Šiman (2014). This requires considering a (single-output) quantile regression problem that is of a "partial" nature, in the sense that it involves both finite-dimensional and functional covariates. We show that, parallel to the finite-dimensional case, conditional depth regions of the response Y can be obtained from the proposed directional regression quantiles. The results are illustrated on simulated and real data sets.*

Keywords. *Functional regression, Quantile regression, Conditional Depth.*

References

Hallin, M., 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.

Hallin, M., Lu, Z., Paindaveine, D., and Šiman, M. (2014). Local constant and local bilinear multiple-output quantile regression. *Bernoulli*, to appear.

1.70

Adaptive one-bit matrix completion

O. Klopp[1], J. Lafon[2], E. Moulines [2,*], J. Salmon

[1] *Modélisation aléatoire de Paris X (MODAL'X); Olga.KLOPP@math.cnrs.fr*
[2] *Institut Mines-Telecom / Telecom ParisTech; surname.name@telecom-paristech.fr*
[*] *Corresponding author*

Abstract. *In the past few years, a large variety of works has shown the benefits of using matrix completion techniques to improve recommender system (e.g. for movie or music recommendation). Most works have considered cases where the coefficients to be determined are continuous scores. Here, we investigate the case where the observations are binary. More precisely, we deal with the problem of matrix completion when the matrix coefficients follow a logistic distribution with a known concave link function. We assume a general sampling scheme for the acquisition process of the coefficients. We study the performance of a nuclear-norm penalized estimator. More precisely we derive bounds for the Kullback-Leibler divergence between the true and estimated distribution. In practice we propose an algorithm based on coordinate gradient descent in order to tackle potentially high dimensional settings.*

Keywords. *nuclear norm regularization; matrix completion; high-dimensional inference*

1.71

Solution of linear inverse problems using flexible dictionaries

Marianna Pensky

Department of Mathematics, University of Central Florida, Orlando; Marianna.Pensky@ucf.edu

Abstract. *We consider solution of a general statistical linear inverse problem. Construction of an adaptive optimal solution for problems of this sort usually is either based on a singular value decomposition or relies on the knowledge of exact inverse images for a specific set of functions (like wavelets). The shortcoming of both approaches lies in the fact that, in many situations, neither the eigenbasis of the linear operator nor a standard dictionary (wavelets, trigonometric* *polynomials) constitutes an appropriate collection of functions for sparse representation of f.*

In the context of regression problems, there have been enormous amount of effort to recover an unknown function using a flexible, overcomplete dictionary. One of the most popular methods is Lasso and its versions (group Lasso, adaptive Lasso and others). Application of those methods is based on minimizing empirical likelihoof and, unfortunately, requires stringent assumptions on the dictionary, the, so called, compatibility conditions. While these conditions may be satisfied for the functions in the original dictionary, they usually do not hold for their images due to contraction imposed by the linear operator. In the talk, we show how one can go around compatibility conditions and apply Lasso for construction of adaptive solutions of general linear inverse problems using flexible, overcomplete dictionaries.

Methodology is applied to analysis of Dynamic Contrast Enhanced Imaging data.

Keywords. *Linear inverse problem; Lasso; adaptive estimation; oracle inequality*

1.72

Diagnostics of mammograms by wavelet-based scaling tools

Pepa Ramírez-Cobo[1,*] and Brani Vidakovic[2]

[1] *Departament of Statistics and Operations Research, Universidad de Cádiz; pepa.ramirez@uca.es*
[2] *Georgia Institute of Technology; brani@bme.gatech.edu*
[*] *Corresponding author*

Abstract. *Two wavelet-based spectra for the analysis of images which scale are described and compared. In the application part, classification of digital mammograms to benign and malignant are considered. Differences in image backgrounds between malignant and normal cases are found, in terms of different fractal descriptors. Also, robust versions of the spectra are delineated.*

Keywords. *Breast cancer diagnostic; Image mono- and multi- fractal analysis; wavelet transform.*

1.73

Learning with localized support vector machines

Mona Eberts[1], Ingo Steinwart[1,*]

[1] *University of Stuttgart, Department of Mathematics, Institute for Stochastics and Applications; ingo.steinwart@mathematik.uni-stuttgart.de, mona.eberts@mathematik.uni-stuttgart.de*
[*] *Corresponding author*

Abstract. *One of the limiting factors of using support vector machines (SVMs) in large scale applications are their super-linear computational requirements in terms of the number of training samples. To address this issue, several approaches that train SVMs on many small chunks of large data sets separately have been proposed in the literature. So far, however, almost all these approaches have only been empirically investigated. In addition, their motivation was always based on computational requirements. In this work, we consider a localized SVM approach based upon a partition of the input space. For this local SVM, we derive a general oracle inequality. Then we apply this oracle inequality to least squares regression using Gaussian kernels and deduce local learning rates that are essentially minimax optimal under some standard smoothness assumptions on the regression function. We further introduce a data-dependent parameter selection method for our local SVM approach and show that this method achieves the same learning rates as before. Finally, we present some larger scale experiments for our localized SVM showing that it achieves essentially the same test performance as a global SVM for a fraction of the computational requirements. In addition, it turns out that the computational requirements for the local SVMs are similar to those of a vanilla random chunk approach, while the achieved test errors are significantly better. This talk is based on Eberts and Steinwart (2013, 2014).*

Keywords. *Regression; Classification; Support Vector Machines; Learning Rates; Large Data Sets*

1.74

Regression estimation based on Bernstein density copulas

Taoufik Bouezmarni$^{(1,*)}$, Benedikt Funk^2 and Félix Camirand Lemyre^1

[1] *Département de Mathématiques, Université de Sherbrooke, Sherbrooke, Québec, Canada J1K 2R1; Taoufik.Bouezmarni@USherbrooke.ca, Felix.Camirand.Lemyre@USherbrooke.ca*
[2] *Technische Universität Dortmund. Fakultät für Mathematik. Lehrstuhl LSIV. Vogelpothsweg 87, 44227 Dortmund; Benedikt.Funke@mathematik.uni-dortmund.de*
[*] *Corresponding author*

Abstract. *The regression function can be expressed in term of copula density and marginal distributions, see Noh, Ghouch and Bouezmarni (2013). In this paper, we propose a new method of estimating a regression function using the Bernstein copula density estimator for the copula density and the empirical distributions for the marginal distributions. The method is fully non-parametric and easy to implement. We provide some asymptotic results related to this copula-based regression modeling by showing the convergence in probability and the asymptotic normality of the estimator. We also study the finite sample performance of the estimator. and illustrate its usefulness by analyzing real data.*

Keywords. *Bernstein copula density estimator; Regression estimation; Nonparametric estimation; Asymptotic properties; Asymptotic normality.*

References

Noh, H. and El Ghouch A. and Bouezmarni, T. (2013). Copula-Based Regression Estimation and Inference. *Journal of American Society of Statistics* **109**, 676–688.

1.75

Quantiles and inequality indices estimation from heavy-tailed distribution

Arthur Charpentier[1] and Emmanuel Flachaire[2,*]

[1] *Université du Québec à Montréal ; charpentier.arthur@uqam.ca*
[2] *Aix-Marseille Université; emmanuel.flachaire@univ-amu.fr*
[*] *Corresponding author*

Abstract. *In this paper, we estimate quantiles and inequality indices from a nonparametric density estimation based on transformed data. A parametric cumulative distribution function is initially used to transform the data into values over the unit interval, from which a nonparametric density estimation is obtained. Finally, an estimation of the density of the original sample is obtained by back-transformation. This approach may be particularly useful to estimate heavy-tailed distributions. We discuss its implementation and its finite sample properties for density estimation, and for estimation and inference with quantiles and inequality indices.*

Keywords. *Quantiles; Inequality indices; Density estimation; Beta kernel; Beta mixture.*

1.76

Root-T consistent density estimation in GARCH models

Jeroen Rombouts

Abstract. *We consider nonparametric estimation of the stationary density of the logarithm of the volatility of the GARCH$(1, 1)$ model. This problem is particularly challenging since this density is still unknown, even in cases where the model parameters are given. Although the volatility variables are only observed with multiplicative independent innovation errors, we manage to construct a nonparametric procedure which estimates the log volatility density consistently. By carefully exploiting the specific GARCH dependence structure of the data, our iterative procedure even attains the striking parametric convergence rate $T^{\{} - 1/2\}$. We derive asymptotic theoretical properties of our estimator, and establish some new smoothness properties of the stationary density. Using numerical simulations, we illustrate the excellent performance of our estimator, and we provide an application to financial data.*

Keywords. *Keyword1; Keyword2. Include up to five keywords separated by semi–colons starting with capital letters.*

1.77

Cross-validated mixed datatype bandwidth selection for nonparametric cumulative distribution/survivor functions

Hongjun Li[1], Jeffrey S. Racine[2]

[1] *Department of Economics, Texas A&M University*
College Station, TX 77843-4228, USA
[2,*] *Department of Economics, McMaster University*
Hamilton, Ontario, Canada, L8S 4M4 racinej@mcmaster.ca
[*] *Jeffrey S. Racine*

Abstract. *We propose a data-driven least squares cross-validation method to optimally select smoothing parameters for the nonparametric estimation of cumulative distribution/survivor functions. We allow for general multivariate covariates that can be continuous, discrete/ordered categorical or a mix of either. We provide asymptotic analysis, examine finite-sample properties via Monte Carlo simulation, and consider an illustration involving nonparametric copula modeling.*

Keywords. *Kernel, Smoothing Parameter Selection, Quantile, Copula*

1.78

Semiparametric approach for regression with covariate subject to limit of detection

S. Kong[1] and B. Nan[1,*]

[1] *Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.; kongsc@umich.edu, bnan@umich.edu*
[*] *Corresponding author*

Abstract. *We consider generalized linear regression with left-censored covariate due to the lower limit of detection. The complete case analysis by eliminating observations with values below limit of detection yields valid estimates for regression coefficients, but loses efficiency. Substitution methods are biased; maximum likelihood method relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach for the regression parameters using an accelerated failure time model for the covariate subject to limit of detection. A two-stage estimation procedure is considered, where the conditional distribution* *of the covariate with limit of detection given other variables is estimated prior to maximizing the likelihood function for the regression parameters. The proposed method outperforms the complete case analysis and the substitution methods as well in simulation studies. Asymptotic properties are provided.*

Keywords. *Accelerate failure time model; Censored covariate; Empirical process; Generalized linear models; Pseudo-likelihood estimation.*

1.79

Bias correction in semiparametric models

A. Gamst*

* *University of California, San Diego; acgamst@math.ucsd.edu*

Abstract. *Consider the problem of estimating the nonlinear functional $\theta(p) = \int p^2 dx$ of the density p on the unit interval. Standard calculations imply that the efficient influence function for θ is $u(x,p) = 2\,(p(x) - \theta(p))$ and the semiparametric information bound $4\,var\,[p(X)] > 0$ suggests that θ should be estimable at the $n^{-1/2}$-rate for arbitrary p. However, the usual one-step adjustment to the plug-in estimator*

$$\widetilde{\theta}_n = \int \widehat{p}_n^2\, dx + n^{-1} \sum_{i=1}^{n} u(X_i, \widehat{p}_n),$$

where $\widehat{p}_n$ is a rate-optimal estimate of the density p, obtained from an auxiliary sample, has the property that

$$E\left(\widetilde{\theta}_n \,|\, \widehat{p}_n\right) = -\int \left(\widehat{p}_n - p\right)^2\, dx = O_P\left(n^{-2\alpha/(2\alpha+1)}\right),$$

assuming that p is α-smooth. This suggests that $\alpha > 1/2$ is required for estimation at the $n^{-1/2}$-rate. Suppose $p(x) = \sum_{k=0}^{\infty} \beta_k \psi_k(x)$ for some orthonormal basis $\{\psi_k\}$. We say that p is α-smooth if $\sum_k k^{2\alpha}\beta_k^2 < \infty$. In fact, expanding the one-step estimator in terms of $\widehat{p}_n(x) = \sum_{k=1}^{r} \widehat{\beta}_k \psi_k(x)$, where r is a "bandwidth" parameter and $\widehat{\beta}_k = n^{-1}\sum_{i=1}^{n} \psi_k(X_i)$, reveals that $\tilde{\theta}_n$ is a biased version of the statistic

$$\widehat{\theta}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{r} \psi_k(X_i)\psi_k(X_j),$$

which is a U-statistic with a truncated kernel, has mean squared error $MSE = O\big(n^{-2}(r \wedge n)\big) + O\big(r^{-4\alpha}\big)$ and is $n^{-1/2}$-consistent and semiparametrically efficient for θ, with $r = n$, so long as $\alpha > 1/4$. In this case, it is possible to demonstrate that $\alpha > 1/4$ is required for estimation at the $n^{-1/2}$-rate. Similar remarks can be made concerning the problem of estimating the residual variance in nonparametric regression with homoskedastic errors and many other problems. We discuss the role of bias correction in achieving information bounds in semiparametric models.

Keywords. *Bias Correction; Semiparametric Models.*

1.80

Analysis of the proportional hazard model with sparse longitudinal covariates

Hongyuan Cao[1]*, Matthew M. Churpek[2], Donglin Zeng[3] and Jason P. Fine[3]

[1] *Department of Health Studies, University of Chicago; hycao@uchicago.edu*
[2] *Department of Medicine, University of Chicago; Matthew.Churpek@uchospitals.edu*
[3] *Department of Biostatistics, UNC-Chapel Hill; dzeng@email.unc.edu, jfine@email.unc.edu*
* *Corresponding author*

Abstract. *Regression analysis of censored failure observations via the proportional hazards model permits time-varying covariates which are observed at death times. In practice, such longitudinal covariates are typically sparse and only measured at infrequent and irregularly spaced follow-up times. Full likelihood analyses of joint models for longitudinal and survival data impose stringent modelling assumptions which are difficult to verify in practice and which are complicated both inferentially and computationally. In this article, a simple kernel weighted score function is proposed with minimal assumptions. Two scenarios are considered: half kernel estimation in which observation ceases at the time of the event and full kernel estimation for data where observation may continue after the event, as with recurrent events data. It is established that these estimators are consistent and asymptotically normal. However, they converge at rates which are slower than the parametric rates which may be achieved with fully observed covariates, with the full kernel method achieving an optimal convergence rate which is superior to that of the half kernel method. Simulation results demonstrate that the large sample approx- imations are adequate for practical use and may yield improved performance relative to last value carried forward approach and joint modelling method. The analysis of the data from a cardiac arrest study demonstrates the utility of the proposed methods.*

Keywords. *Convergence rates; Cox model; Kernel weighted estimation; Sparse longitudinal covariates.*

2.01

The projection pursuit multiple index model

Michael G. Akritas[1]

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

Abstract. *The class of Projection Pursuit Multiple Index (PPMI) Models will be introduced. A subclass of PPMI models is almost equivalent to the class of multiple index (MI) models, but uses a parametrization related to that of the single index (SI) model. A different subclass of PPMI models generalizes the class of projection pursuit (PP) models. A method for estimation, and a unified asymptotic theory are presented.*

Keywords. *High-Dimensional Data Modeling: Index models; Projection pursuit models; Estimating equations; Empirical processes*

2.02

On the performance of the Lasso as a method of nonparametric estimation

A. Dalalyan[1,*], M. Hebiri[2] and J. Lederer[3]

[1] *ENSAE-CREST-GENES; arnak.dalalyan@ensae.fr,*
[*] *Corresponding author*
[2] *Université Paris-Est – Marne-la-Vallée; Mohamed.Hebiri@univ-mlv.fr*
[3] *Cornell University; johanneslederer@cornell.edu*

Abstract. *Although the Lasso has been extensively studied, the relationship between its prediction performance and the correlations of the covariates is not yet fully understood. The investigation of this relationship is particularly important for clarifying the accuracy of the Lasso as a method of nonparametric estimation with a large, overcomplete dictionary. In this talk, we give new insights into this relationship in the context of regression with deterministic design. We show, in particular, that the incorporation of a simple correlation measure into the tuning parameter leads to a nearly optimal prediction performance of the Lasso even when the elements of the dictionary are highly correlated. However, we also reveal that for moderately correlated dictionary, the performance of the Lasso can be mediocre irrespective of the choice of the tuning parameter. For the illustration of our approach with an important application, we deduce nearly optimal rates for the least-squares estimator with total variation penalty.*

Keywords. *Prediction risk; Correlated design; Total-variation; ℓ_1-penalty; fast rates.*

2.03

Robust rank correlation based screening

Gaorong Li[1], Heng Peng[2,*], Jun Zhang[3], and Lixing Zhu[2]

[1] *Beijing University of Technology; ligaorong@gmail.com*
[2] *Hong Kong Baptist University; hpeng@math.hkbu.edu.hk, lzhu@math.hkbu.edu.hk*
Shenzhn University; zhangjunstat@gmail.com
[*] *Corresponding author*

Abstract. *Independence screening is a variable selection method that uses a ranking criterion to select significant variables, particularly for statistical models with nonpolynomial dimensionality or "large p, small n" paradigms when p can be as large as an exponential of the sample size n. In this paper we propose a robust rank correlation screening (RRCS) method to deal with ultra-high dimensional data. The new procedure is based on the Kendall τ correlation coefficient between response and predictor variables rather than the Pearson correlation of existing methods. The new method has four desirable features compared with existing independence screening methods. First, the sure independence screening property can hold only under the existence of a second order moment of predictor variables, rather than exponential tails or alikeness, even when the number of predictor variables grows as fast as exponentially of the sample size. Second, it can be used to deal with semiparametric models such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation even when there are nonparametric functions in the models. Third, the procedure can be largely used against outliers and influence points in the observations. Last, the use of indicator functions in rank correlation screening greatly simplifies the theoretical derivation due to the boundedness of the resulting statistics, compared with pre vious studies on variable screening. Simulations are carried out for comparisons with existing methods and a real data example is analyzed.*

Keywords. *variable selection; rank correlation screening; dimensionality reduction; semiparametric models; large p small n.*

2.04

An improved quantitative structure-activity relationship (QSAR) analysis of peptides

Zhijun Dai[1,2], Lifeng Wang[1,2], Yuan Chen[1,2], Haiyan Wang[3], Lianyang Bai[2,4] and Zheming Yuan[1,2]

[1] *Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha, China;*
[2] *Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha, China;*
[3] *Department of Statistics, Kansas State University, Manhattan, Kansas, USA*
[4] *Hunan Academy of Agricultural Sciences, Changsha, China*

Abstract. *In this talk, I will present a method to perform improved QSAR analysis of peptides. The method contains feature selection, near-neighbor sample selection , and weighted regression and prediction. The modeling involves a double selection procedure that first performs feature selection and then conducts sample selection before the final regression analysis. Five hundred and thirty-one physicochemical property parameters of amino acids were used as descriptors to characterize the structure of peptides. These high-dimensional descriptors then go through a feature selection process given by the Binary Matrix Shuffling Filter (BMSF) to obtain a set of important low dimensional features. Each descriptor that passed the BMSF filtering also receives a weight defined through its contribution to reduce the estimation error. These selected features were served as the predictors for subsequent sample selection and modeling. Based on the weighted Euclidean distances between samples, a common range was determined with high-dimensional semivariogram and then used as a threshold to select the near-neighbor samples from the training set. For each sample to be predicted, the QSAR model was established using SVR with the weighted, selected features based on the exclusive set of near-neighbor training samples. Prediction was conducted for each test sample accordingly. The performances of this method are tested with the QSAR analysis of angiotensin-converting enzyme (ACE) inhibitors and HLA-A*0201 data sets. Improved prediction accuracy was obtained in both applications. This method can optimize the QSAR modeling from both the feature selection and sample selection perspectives, which leads to improved accuracy over single selection methods. We expect this method to have extensive application prospect in the field of regression prediction.*

Keywords. *Variable selection; Sample weighting; Quantitative structure-activity relationship (QSAR) modeling; Cross-validation*

2.05

Distributional inference: reproducibility and limitations

D A S Fraser[11]

[1] *Statistical Sciences, University of Toronto; dfraser@utstat.toronto.edu * Corresponding author*

Abstract. *The January 7, 2014 issue of Science has a lead Editorial on Reproduciblity in science, with emphasis on openness on methodology and possible pitfalls that can invalidate conclusions. And the June 30, 2013 issue had a Perspective by Bradley Efron recommending that Bayesian priors should be restricted to those describing just "Genuine prior information". This latter would invalidate the invariance approach emphasized by Laplace, as pointed out in the Science Letters of September 27, 2013. This all reflects a rising concern in science that stated results should mean what they say and not be opinion pieces, in whole or part. We examine reproducibility and potential pitfalls in the use of distributional inference for both frequentist and Bayesian procedures.*

Keywords. *Inference distribution: Confidence; Bayes; Fiducial.*

2.06

A general predictive distribution function and confidence distributions

Min-ge Xie[1]

[1] *Department of Statistics, Rutgers University; mxie@stat.rutgers.edu*

Abstract. *In this talk, we develop a new and general framework for prediction, in which a prediction is presented in the form of a distribution function, called predictive distribution function. This predictive distribution function, developed based on confidence distributions, has a clear frequentist probability interpretation and can provide meaningful answers for all sorts of questions related to prediction. It can also serve as a unifying point for existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. A simple yet broadly applicable algorithm by Monte-Carlo or bootstrapping is proposed. Numerical examples are provided to demonstrate the methodology.*

Keywords. *Frequentist prediction; Confidence interpretation; Unified approach*

2.07

Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness

Regina Liu1[1], Dungang Liu[2,*] and Minge Xie[1]

[1] *Rutgers University; rliu@stat.rutgers.edu, mxie@stat.rutgers.edu*
[2] *Yale University; dungang.liu@yale.edu*
[*] *Corresponding author*

Abstract. *Meta-analysis has been widely used to synthesize evidence from multiple studies for common hypotheses or parameters of interest. However, it has not yet been fully developed for incorporating heterogeneous studies, which arise often in applications due to different study designs, populations or outcomes. For heterogeneous studies, the parameter of interest may not be estimable for certain studies, and in such a case, these studies are typically excluded from conventional meta-analysis. The exclusion of part of the studies can lead to a non-negligible loss of information. We introduce a meta-analysis for heterogeneous studies by combining the* confidence density functions *derived from the summary statistics of individual studies, hence referred to as the CD approach. It includes all the studies in the analysis and makes use of all information, direct as well as indirect. Under a general likelihood inference framework, this new approach is shown to have desirable properties, i) it is asymptotically as efficient as the maximum likelihood approach using individual participant data (IPD) from all studies; ii) unlike the IPD analysis, it suffices to use summary statistics to carry out the CD approach. Individual-level data are not required; and iii) it is robust against misspecification of the working covariance* *structure of the parameter estimates. Besides its own theoretical significance, the last property also substantially broadens the applicability of the CD approach. All these properties of the CD approach are further confirmed by data simulated from a randomized clinical trials setting as well as by real data on aircraft landing performance. Overall, one obtains an unifying approach for combining summary statistics, subsuming many of the existing meta-analysis methods as special cases.*

The need and suitability of combining heterogeneous studies is also discussed briefly in the context of regulation or policy making.

Keywords. *Combining information; Confidence distribution, efficiency; Heterogeneous studies; Individual participant data; Multivariate meta-analysis.*

2.08

Testing for breaks in regression models with dependent data

J. Hidalgo[1], V. Dalla[2]

[1] *London School of Economics. Houghton Street, London WC2A 2AE, U.K.*
[2] *Athens University. Greece.*

Abstract. *The paper examines a test for smoothness/breaks in a nonparametric regression model with dependent data. In particular we examine tests for k breaks against the alternative of $k+k_0$ breaks fro some k_0. The test is based on the supremum of the difference between the one-sided kernel regression estimates. When the errors of the model are strong dependent, we have that surprisingly, the behaviour of the test depends on whether the regressors are deterministic or stochastic. In the latter case, the normalization constants to obtain the asymptotic Gumbel distribution are standard, whereas in the former case, those constants are data dependent and the critical values are difficult to obtain, if possible. This motivates, together with the fact that the rate of convergence to the Gumbel distribution is only logarithmic, the use of a bootstrap analogue of the test. We show that it is asymptotic validity. One interesting finding is that neither subsampling nor the sieve bootstrap will lead to asymptotic valid inferences, they do not provide valid estimates of the critical values of the test, in our scenario. Finally, we present a Monte-Carlo experiment to shed some light on the finite sample behaviour of the test(s).*[a]

Keywords. *Nonparametric regression; Breaks/smoothness; Strong dependence; Extreme-values distribution; Frequency domain; Bootstrap algorithms*

[a] We like to thank Marie Huskova for their comments on a previous version of the paper. Of course, any remaining errors are our sole responsibility

2.09

Nonparametric estimation of fixed effects panel data varying coefficient models

Juan M. Rodríguez-Poo[1,*] and Alexandra Sobcrón[2]

[1] *Departamento de Economía. Universidad de Cantabria. Avda. de los Castros s/n. 39005. Santander. Spain ; rodrigjm@unican.es*
[2] *Departamento de Economía. Universidad de Cantabria ; soberonap@unican.es*
[*] *Corresponding author*

Abstract. *In this paper, we consider the estimation of a panel data model where the heterogeneity term is arbitrarily correlated with the covariates and the coefficients are unknown functions of some explanatory variables. The estimator is based in a deviation from the mean transformation of the regression model and then a local linear regression is applied to estimate the unknown varying coefficient functions. It turns out that the standard use of this technique rends a non-negligible asymptotic bias. In order to avoid it, in the estimation procedure, we introduce a high dimensional kernel weight. As a consequence, the resulting estimator shows a bias that asymptotically tends to zero at usual nonparametric rates. However, the variance is enlarged, and therefore the estimator shows a very slow rate of convergence. In order to achieve the optimal rate, we propose a one-step backfitting algorithm. The resulting two step estimator is shown to be asymptotically normal and its rate of convergence is optimal within its class of smoothness functions. Furthermore, the estimator is oracle efficient. Finally, we show some Monte Carlo results that confirm the theoretical findings.*

Keywords. *Varying coefficients model; Fixed effects; Panel data; Local linear regression; Oracle efficient estimator.*

2.10

Distribution-free tests of inequality constraints on conditional moments

Delgado, Miguel A.[1] and Escanciano, Juan Carlos[2]

[1] *Departamento de Economía, Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain. E-mail: miguelangel.delgado@uc3m.es*
[2] *Economics Department, Indiana University, Bloomington, Indiana, USA; E-mail: jescanci@indiana.edu, Corresponding author*

Abstract. *We present a methodological approach for testing inequality constraints on conditional moments. The null hypothesis of an inequality restriction is equivalently expressed as an equality using the least concave majorant operator applied to the integrated conditional moments. A suitable time transformation of the basic process renders an asymptotic distribution-free test,* *with critical values that can be easily tabulated. Monte Carlo experiments provide evidence of the satisfactory finite sample performance of the proposed test.*

Keywords. *Nonparametric testing; Least concave majorant; Moment inequalities; Distribution-free.*

2.11

Small-b and fixed-b asymptotics for weighted covariance estimation in fractional cointegration

Javier Hualde[1], Fabrizio Iacone[2]

[1] *Universidad Pública de Navarra*
[2] *University of York*

Abstract. *In a standard cointegrating framework, Phillips (1991) introduced the weighted covariance (WC) estimator of cointegrating parameters. Later, Marinucci (2000) applied this estimator to various fractional circumstances and, like Phillips (1991), analyzed the so-called small-b asymptotic approximation to its sampling distribution. Recently, an alternative limiting theory has been successfully employed to approximate the sampling distribution of nonparametric estimators of spectral densities and long run covariance matrices more accurately than by traditional asymptotics. This has been named fixed-b asymptotics, and, while it has been hardly used in cointegration, the form of the WC estimator leads to a natural application of this type of limiting theory. Thus, we derive the fixed-b limit of WC estimators in a fractional setting, filling also some gaps in the traditional (small-b) theory. Additionally, we compare the small-b and fixed-b limiting approximations to the sampling distribution of a WC estimator by means of a Monte Carlo experiment, finding that the fixed-b limit is more accurate.*

2.12

Max-stable processes on river networks

S. Engelke[1,2,*], P. Asadi[2] and A.C. Davison[1]

[1] *École Polytechnique Fédérale de Lausanne; sebastian.engelke@epfl.ch, anthony.davison@epfl.ch*
[2] *Université de Lausanne; peiman.asadi@unil.ch*
[*] *Corresponding author*

Abstract. *Max-stable processes are suitable models for extreme events that exhibit spatial dependencies. The dependence measure is usually a function of Euclidean distance between two locations. In this talk, we model extreme river discharges on a river network in the upper Danube catchment, where flooding regularly causes huge damage. Dependence is more complex in this*

2.13

Prediction and estimation of random fields on quarter plane

Priya Kohli[1,*] and Mohsen Pourahmadi[2]

[1] *Department of Mathematics and Statistics, Connecticut College, USA, pkohli@conncoll.edu*
[2] *Department of Statistics, Texas A& M University,USA, pourahm@stat.tamu.edu.*

Abstract. *We study solutions of several prediction problems for random fields (2-D processes) by extending some nonstandard prediction problems for a stationary time series (1-D process) based on the modified pasts. The goal is to provide informative and explicit prediction error variance formulas in terms of either the AR or MA parameters of the random fields so that the effect of an additional observation is linked to its spatial location via the size of these parameters. Using the Wold decomposition of stationary random fields, their multi-step ahead prediction errors and variances, solutions are provided for various nonstandard prediction problems when a number of observations are either added to or deleted from the quarter-plane past. Unlike the time series situation, the prediction error variances for random fields seems to be expressible only in terms of the MA parameters, and attempts to express them formally in terms of the AR parameters lead to a new and mysterious projection operator which captures the nature of the "edge-effects" encountered in the estimation of the spectral density function of stationary random fields. The approach leads to a number of technical issues and open problems. We study* cepstral random field models*, which are formulated in the frequency domain by ensuring a positive spectral density and provide a general way of specifying a random field. We obtain recursive formulas expressing the AR and MA coefficients in terms of the spatial cepstral coefficients to facilitate easy and fast computation of the proposed predictor coefficients. The procedure is illustrated using a simulation study and application to real data.*

Keywords. *Cepstral Random Fields, Edge-effects; Projection Operator; Unilateral Representation*

2.14

Baxter´s inequality and sieve bootstrap for random fields

M. Meyer[1,*], C. Jentsch[2] and J.-P. Kreiss[1]

[1] *Technische Universität Braunschweig, Institut für Mathematische Stochastik, Pockelsstrasse 14, D–38106 Braunschweig, Germany; marco.meyer@tu-bs.de, j.kreiss@tu-bs.de*
[2] *Universität Mannheim, Abteilung Volkswirtschaftslehre, L 7, 3-5, D-68131 Mannheim, Germany, carstenjentsch@web.de*
[*] *Corresponding author*

Abstract. *The concept of the autoregressive sieve bootstrap for time series is expanded to the case of random fields. Given a finite data sample of rectangular shape, the procedure fits a finite-order autoregressive (AR) model to the sample using Yule-Walker-type methods. The residuals of this fit are resampled which allows for construction of a bootstrap sample. Typically, the order of the fitted AR model depends on the sample size and is assumed to increase as the sample size tends to infinity. We will explore the range of validity of this resampling procedure and provide for a large class of spatial processes a general check criterion, which allows to decide whether the sieve bootstrap asymptotically works for a specific statistic of interest or not. In the latter case we will also point out the exact reason which causes the bootstrap to fail.*

Crucial for the result is the fact that, under relatively mild conditions, rather general random fields inherit a certain AR structure. Furthermore, the validity of the bootstrap scheme depends on a generalization of Baxter's inequality, cf. Baxter (1962), to the case of random fields. This inequality connects the coefficients of a finite-order AR fit, i.e. the solutions of the Yule-Walker equations, to the (infinitely many) coefficients of the AR representation of the underlying spatial process.

Keywords. *Autoregression; prediction theory; bootstrap; random fields.*

References

Baxter, G. (1962). An Asymptotic Result for the Finite Predictor. *Math. Scand.* **10**, 137–144.

2.15

A test for stationarity for irregularly spaced spatial data

Suhasini Subba Rao[1]

[1] *Department of Statistics, Texas A&M University. College Station, U.S.A.; suhasini@stat.tamu.edu*

Abstract. *The analysis of spatial data is based on a set of assumptions, which in practice need to be checked. A commonly used assumption is that the spatial random field is second order stationary. In this paper, a test for spatial stationarity for irregular sampled data is proposed. The test is based on a transformation of the data (a type of Fourier transform), where the correlations between the transformed data is close to uncorrelated if the random field is second order stationary, whereas this property does not hold if the random field is second order nonstationary. Using this property a test for second order stationarity is constructed, which is based on measuring the degree of correlation in the transformed data. The asymptotic sampling properties of the test statistic is derived under both stationarity and nonstationary and the method is illustrated with simulations and a real data example.*

2.16

Principal components analysis without eigenanalysis

Jim Ramsay[1,*], Alois Kneip[2]

[1] Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, QC, Canada H3A 1B1; ramsay@psych.mcgill.ca
[2] Department of Statistics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany; akneip@uni-bonn.de
[*] Corresponding author

Abstract. *Principal components analysis is an invaluable tool in multivariate or functional data analysis, but it suffers from several defects. The fits to all variables are required to be by least squares, there is no distinction made between principal component vectors/functions and principal component scores as parameters, eigenvectors/functions as objects spanning the subspace are accorded an interpretive or substantive significance that they don't deserve, and it is difficult to devise estimation properties now considered essential with other methods, such as regularity and robustness. This new approach is based on a parameter cascade that defines factor scores and other case-specific parameters as smooth functions of principal component vectors/functions. The approach readily generalizes to the estimation of extended models, such as:*

- *GAM/PCA where measurement models can vary across variables*
- *partial least squares/PCA,*
- *functional PCA and*
- *functional PCA combined with registration*

In this approach, any suitable loss function may be employed for any variable in the multivariate case, or vary over time in the functional case. The method is also adaptable to non-flat manifold estimation. The talk will include an application of this extended PCA to NMR spectroscopy data.

Keywords. *Principal components analysis, generalized additive models, cure registration, functional data analysis*

2.17

Points impact in functional linear regression

Alois Kneip[1], Dominik Poss[1] and Pascal Sarda[2]*

[1] *Universität Bonn; akneip@uni-bonn.de, dposs@uni-bonn.de*
[2] *Université Paul Sabatier, Toulouse; sarda@math.univ-toulouse.fr*
* *Corresponding author*

Abstract. *In the setting where a scalar variable Y depends on the variability of a functional variable X, functional linear regression is one of the most popular model which has been widely studied in the literature (see for instance Cai and Hall (2007), Crambes et al. (2009)). The response Y is related to the whole curve X through a linear functional where a slope function is the parameter of the model. More recently, some authors as McKeague and Sen (2010) have pointed out that one or several sensitive points of the curve X may have also specific influence on the response. These points are denominated as points of impact.*

Following this idea, we propose a generalization of the functional linear regression where it is assumed that there exists an unknown number of points of impacts. The aim is then to estimate, from a sample drawn from (X, Y), the usual slope parameter and to determine number and locations of points of impact as well as corresponding regression coefficients. It is shown at first that the model is identifiable provided that the variables X possesses specific local variation. Roughly speaking it means that some parts of local variation of X in a small neighborhood of a point of impact is uncorrelated with the remainder of the curve.

We propose a method for estimating the number and locations of points of impact. We prove that this number can be estimated consistently and then we derive rates of convergence for location estimates and regression coefficients.

Keywords. *Functional linear regression; points of impact; stochastic processes.*

References

Cai, T. and Hall, P. (2007). Prediction in Functional Linear Regression. *Annals of Statistics* **34**, 2159–2179.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothnig Splines Estimators for Functional Linear Regression. *Annals of Statistics* **37**, 35–72.

McKeague, I.W. and Sen, B. (2010). Fractals with point impact in functional linear regression. *Annals of Statistics* **38**, 2559–2586.

2.18

Modelling electricity prices a functional data on random domains

Alois Kneip[1] and Dominik Liebl[2,*]

[1] *University Bonn; akneip@uni-bonn.de*
[2] *Université libre de Bruxelles; Dominik.Liebl@ulb.ac.be*
[*] *Corresponding author*

Abstract. *Motivated from the so-called merit-order model, a widely accepted theoretical model for electricity spot prices, we model electricity spot prices as noisy discretization points of a latent functional time series of daily price functions (see also Liebl (2013)). In order to consider also the seasonal influences, the price functions are defined as bi-variate functions of intra-daily electricity demand and daily air-temperature.*
This leads to a rarely studied phenomenon of densely sampled functional data observed only within random sub-domains, which harms the estimation of the covariance function over the total domain. We propose a procedure that allows to recover the price functions, and subsequently the covariance function, over the total domain. Furthermore, we present in-probability convergence rates for the local linear estimators of the mean and covariance function as well as for the functional principal components. In our application, we demonstrate that the re-covered price functions are useful for constructing forecast models with precise distribution forecasts.

Keywords. *Functional data analysis; random sub-domains; multivariate local linear estimation; electricity spot prices*

References

Liebl, D. (2013). Modeling and forecasting electricity spot prices: a functional data perspective. *The Annals of Applied Statistics* **7**, 1562–1592.

2.19

Registration and functional principal componens

Alois Kneip[1], Heiko Wagner[1]

[1] *University of Bonn*

Abstract. *Functional data analysis deals with modeling samples of smooth functions $x_1, \ldots, x_n$. A basic tool of statistical analysis consists in the use of functional principal components. But an inherent problem in many applications is the possible existence of two types* *of variations : a phase variation (horizontally) due to time lags, and an amplitude variation. Determining mean and principal components then often does not lead to interpretable results, unless a registration procedure is applied which eliminates phase variation. In this context usually registration is considered as a pre-processing step. In this talk we propose a new algorithm that combines registration and principal component analysis. The approach is illustrated by some real data examples.*

2.20

Risk hull method in linear ill-posed inverse problems

Yu. Golubev[1]

[1] *Aix Marseille University; golubev.yuri@gmail.com*

Abstract. *The talk is based on the paper Cavalier and Golubev (2006) devoted to spectral regularization methods of the linear inverse problem $Y = Af + \epsilon$, where ϵ is a white Gaussian noise and A is a known compact operator with singular values converging to zero with polynomial decay. The unknown function f is recovered with the help of a spectral regularization method with a regularization parameter governed by a data-driven procedure. This procedure is based on the method of the risk hull minimization. We provide non-asymptotic upper bounds for the mean square risk of this method and show, in particular, that in numerical simulations, this approach may substantially improve classical methods based on the unbiased risk estimation.*

Keywords. *Inverse problem; Regularization; Risk hull; Oracle inequality.*

References

Cavalier, L. and Golubev, Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *The Annals of Statist.* **34**, 1653–1677.

2.21

Bound-to-bound data collaboration

Michael Frenklach* and Andrew Packard

Department of Mechanical Engineering, University of California at Berkeley; frenklach@berkeley.edu, apackard@berkeley.edu
* *Corresponding author*

Abstract. *We present a methodology of uncertainty quantification developed in a series of studies and termed Bound-to-Bound Data Collaboration (abbreviated to B2B-DC and described in the references below). B2B-DC is framework for combining models and training data from multiple sources to explore their collective information content. It is built on an underlying physical process and associated model, a collection of experimental observations with specified uncertainties, algebraic surrogate models (response surfaces) representing parametric dependence of the physical-model predictions of the experimental observables on the uncertain parameters, and specialized constrained-optimization algorithms. The methodology makes predictions on the true feasible set, transfers the uncertainties of both model parameters and training-set experiments directly into prediction, tests and quantifies consistency among data and models, explores sources of inconsistency, discriminates among differing models, and enables analysis of global sensitivities of uncertainty in prediction to the uncertainties in data and model. Applications of the approach include combustion science and engineering, atmospheric chemistry, and system biology.*

Keywords. *Uncertainty quantification; Predictive modeling; Response surfaces, Global sensitivities, Optimization.*

References

Frenklach, M., Packard, A., Seiler, P., and Feeley, R. (2004). Collaborative data processing in developing predictive models of complex reaction systems. *Int. J. Chem. Kinet.* **36**, 57–66.

Feeley, R., Seiler, P., Packard, A., and Frenklach, M. (2004) Consistency of a reaction dataset. *J. Phys. Chem. A* **108**, 9573–9583.

Frenklach, M. (2007). Transforming data into knowledge—process informatics for combustion chemistry. *Proc. Combust. Inst.* **31**, 125–140.

Russi, T., Packard, T., Feeley, R., and Frenklach, M. (2008). Sensitivity analysis of uncertainty in model prediction. *J. Phys. Chem. A* **112**, 2579–2588.

Russi, T., Packard, A., and Frenklach, M. (2010) Uncertainty quantification: Making predictions of complex reaction systems reliable. *Chem. Phys. Lett.* **499**, 1–8.

2.22

What do I make of your latinorum? Sensitivity auditing of mathematical modelling

N1. Andrea Saltelli[1,*], and N2. Silvio Funtowicz[2]

[1] *Econometrics and Applied Statistics, Joint Research Centre, European Commission; andrea.saltelli@jrc.ec.europa.eu,*
[2] *University of Bergen, Centre for the Study of the Sciences and Humanities; Silvio.Funtowicz@svt.uib.no*
[*] *Corresponding author*

Abstract. *More stringent quality criteria are needed for models used at the science/policy interface, e.g. in the context of policy ex ante impact assessment studies. While modelers strive to improve the predictive capacity of models by use of emulators, history matching, data* *assimilation, ensemble modelling, model averaging, and other tools of the sort, the same modelers may at times claim to have tamed model structural uncertainty. We believe that risk is different from uncertainty and suggest that when the result of a modelling activity feeds into the policy process an additional level of verification is needed, which calls into question normative dimensions, the treatment of uncertainty, the identity of the story-teller, the relevance and the salience of the narrative.*

We call this approach sensitivity auditing of the modelling process, and encode it into a set of seven simple rules.

Keywords. *Sensitivity analysis; Sensitivity auditing; NUSAP; Post normal science (PNS); Science for policy.*

References

Saltelli, A. and Guimarães Pereira, Â. and Van der Sluijs, J.P. and Funtowicz, S. (2013). What do I make of your latinorum? Sensitivity auditing of mathematical modelling. *Int. J. Foresight and Innovation Policy*, **9**, 213–234.

Saltelli, A., Funtowicz, S. (2014). When all models are wrong: More stringent quality criteria are needed for models used at the science-policy interface. *Issues in Science and Technology*, Winter 2014, 79–85.

2.23

Bayesian model calibration in the presence of model discrepancy

Ralph Smith*[1] and Jerry McMahan, Jr.[1]

[1] *Department of Mathematics, North Carolina State University, Raleigh, NC 27695; rsmith@ncsu.edu, jamcmaha@ncsu.edu*

* *Corresponding author*

Abstract. *Measurement and model errors produce uncertainty in model parameters estimated through least squares fits to data or Bayesian model calibration techniques. In many cases, model errors or discrepancies are neglected during model calibration. However, this can yield nonphysical parameter values for applications in which the effects of unmodeled dynamics are significant. It can also produce prediction intervals that are inaccurate in the sense that they do not include the correct percentage of future observations. In this presentation, we discuss techniques to quantify model discrepancy terms in a manner that yields physical parameters and correct prediction intervals. We illustrate aspects of the framework in the context of distributed structural models with highly nonlinear parameter dependencies.*

Keywords. *Bayesian model calibration; Model discrepancy terms; Prediction intervals.*

2.24

Mini-Minimax uncertainty quantification for emulators

Jeffrey C. Regier[1] and Philip B. Stark[1,*]

[1] *Department of Statistics, University of California, Berkeley {jeff,stark}@stat.berkeley.edu*
[*] *Corresponding author*

Abstract. *Consider approximating a function f by an* emulator $\hat{f}$ *based on n noiseless observations of f. Let w be a point in the domain of f. How big might the error $|\hat{f}(w) - f(w)|$ be? If f could be arbitrarily rough, this error could be arbitrarily large. Suppose f is smooth: Lipschitz with a known constant. We find a lower bound on the smallest n for which, for the best emulator $\hat{f}$, $|\hat{f}(w) - f(w)| \leq \epsilon$. But in general, we will not know whether f is Lipschitz, much less know its Lipschitz constant. Assume optimistically that f is Lipschitz-continuous with the smallest constant consistent with the n data. We find the maximum (over such regular f) of $|\hat{f}(w) - f(w)|$ for the best possible emulator $\hat{f}$; this the "mini-minimax uncertainty" at w. In reality, f might not be Lipschitz or—if it is—it might not attain its Lipschitz constant on the data. Hence, the mini-minimax uncertainty at w could be much smaller than $|\hat{f}(w) - f(w)|$. But if the mini-minimax uncertainty is large, then—even if f satisfies the optimistic smoothness assumption—$|\hat{f}(w) - f(w)|$ could be large, even for the best $\hat{f}$. For the Community Atmosphere Model (CAM), the maximum (over w) of the mini-minimax uncertainty based on a set of 1154 observations of f is no smaller than it would be for a single observation of f at the centroid of the 21-dimensional parameter space. We also find lower confidence bounds for quantiles of the mini-minimax uncertainty and its mean over the domain of f. For the CAM, these lower confidence bounds are an appreciable fraction of the maximum. In contrast, for simpler functions—even in high-dimensional spaces—the mini-minimax uncertainty can be small.*

Keywords. *Uncertainty quantification; Emulator; Information-based complexity; Lipschitz condition; Empirical smoothness.*

2.25

Fast DD-classification of functional data

K. Mosler[1,*], P. Mozharovskyi[1]

[1] *Statistics and Econometrics, Universität zu Köln, 50923 Köln, Germany; mosler@statistik.uni-koeln.de, mozharovskyi@statistik.uni-koeln.de*
[*] *Corresponding author*

Abstract. *A fast nonparametric procedure for classifying functional data is introduced. It consists of a two-step transformation of the original data plus a classifier operating on the unit* *cube. The functional data are first mapped into a finite-dimensional location-slope space and then transformed by a multivariate depth function into the DD-plot, which is a subset of the unit cube. This transformation yields also a new notion of depth for functional data. Three alternative depth functions are employed for this, as well as two rules for the final classification on* $[0,1]^2$*. The entire methodology does not involve smoothing techniques and is completely nonparametric. It is robust, efficiently computable, and has been implemented in an R environment. The new procedure is compared with known ones, including the componentwise approach of Delaigle, Hall and Bathia (2012), and its applicability is demonstrated by simulations as well as a benchmark study.*

Keywords. *Functional depth; Supervised learning; Central regions; Location-slope depth; DD-plot.*

2.26

Geostatistics for Hilbert data

Piercesare Secchi[1]

[1] *MOX, Dipartimento di Matematica, Politecnico di Milano; piercesare.secchi@polimi.it*

Abstract. *When dealing with high-dimensional georeferenced data, the need of spatial prediction results in both theoretical and practical issues. The talk addresses these problems by proposing an extension of some geostatistical techniques to non-stationary functional random fields. A new theoretical framework is established to perform Universal Kriging of spatially dependent functional data belonging to a Hilbert space; moreover, estimators of the spatial mean and the spatial covariance structure are derived. The generality of the proposed approach allows to include pointwise and differential information brought by data by embedding the analysis in a proper Hilbert space, possibly other than* L^2*. Three environmental case studies will serve as illustration. The first deals with temperature curves embedded in* L^2 *while the second considers functional compositional data handled through the Aitchison geometry. In the third case study a Hilbert space approximation is exploited to make spatial predictions for data belonging to a Riemannian manifold, the space of positive definite symmetric matrices.*

The talk is based on work developed in collaboration with Alessandra Menafoglio (Mox, Dipartimento di Matematica, Politecnico di Milano), Davide Pigoli (CRiSM, Department of Statistics, University of Warwick), Alberto Guadagnini (Dipartimento di Ingegneria Civile e Industriale, Politecnico di Milano) and Matilde Dalla Rosa (ENI S.p.A., Exploration & Production Division).

Keywords. *Spatial statistics; Kriging; Functional Data Analysis; Hilbert Data*

References

Menafoglio A. , Secchi P., Dalla Rosa M. (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electron. J. Stat.*, **7**, 2209–2240.

Menafoglio A., Guadagnini A., Secchi P. (2014). A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers. *Stoch. Environ. Res. Risk Assess.*, in press.

Pigoli D., Secchi P. (2012). Estimation of the mean for spatially dependent data belonging to a Riemannian manifold. *Electron. J. Stat.*, **6**, 1926–1942.

Pigoli D., Menafoglio A., Secchi P. (2013). Kriging prediction for manifold-valued random field. *MOX Technical Report*, 63/2013. Politecnico di Milano.

2.27

Visualization of shape outliers in functional data samples

Ana Arribas Gil[1]* and Juan Romo[1]

[1] *Departamento de Estadística, Universidad Carlos III de Madrid, Spain; ana.arribas@uc3m.es, juan.romo@uc3m.es*
* *Corresponding author*

Abstract. *We propose a new method to visualize and detect shape outliers in samples of curves. In functional data analysis, we observe curves defined over a given real interval and shape outliers may be defined as those curves that exhibit a different shape from the rest of the sample. Whereas magnitude outliers, that is, curves that lie outside the range of the majority of the data, are in general easy to identify, shape outliers are often masked among the rest of the curves and thus difficult to detect. In this article, we exploit the relationship between two measures of depth for functional data to help to visualize curves in terms of shape and to develop an algorithm for shape outlier detection. We illustrate the use of the visualization tool, the outliergram, through several examples and analyze the performance of the algorithm on a simulation study.*

Keywords. *Depth measure; Functional data; Shape outliers.*

2.28

An alternative to FPCA for functional data in two arguments

K. Chen[1], P. Delicado[2,*] and H.G. Müller[3]

[1] *Department of Statistics and Department of Psychiatry, University of Pittsburgh, Pittsburgh, USA; khchen@pitt.edu*
[2] *Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain; pedro.delicado@upc.edu*
[3] *Department of Statistics, University of California, Davis, USA; hgmueller@ucdavis.edu*
* *Corresponding author*

Abstract. *We consider a square integrable random function that depends on two arguments, possibly with asymmetric roles (for instance, one of the arguments can be time). We propose an asymmetric representation of this random function, that is an alternative to the Karhunen-Loève (KL) representation, leading to an alternative to the standard FPCA. Interesting properties of our proposal are the following: it is the optimal solution of an approximation problem (similar to that leading to KL representation); it allows to treat asymmetrically the two arguments of the functional data; in the new approach the functional data are represented as a sum of terms that are the product of two functions, each depending on only one argument, this way to look for an intuitive interpretation of each term is easier than in the standard FPCA. We also study the convergence of the sampling (or empirical) version of our proposal to the theoretical counterpart.*

We illustrate our proposal analyzing functional data coming from the Human Fertility Database. Our interest is focused on the Age-Specific Fertility Rate (ASFR), that for a given age s (expressed in years) and a given calendar year t is defined as a quotient: the number of births during the year t corresponding to women aged s, divided by the number of women aged s during the year t. Our proposal leads to much more straightforward analysis of the dynamics of the fertility process compared to standard FPCA

Keywords. *Asymptotics; Demography; Functional data analysis; Functional principal component analysis; Karhunen-Loève expansion.*

2.29

Saddle-point model selection

K. Efimov[1], V. Spokoiny[2]

[1] *Humboldt University of Berlin, Moscow Institute of Physics and Technology, kirill.efimovs@gmail.com*
[2] *Weierstrass-Institute, Humboldt University of Berlin, Moscow Institute of Physics and Technology, spokoiny@wias-berlin.de*

Abstract. *Within the penalized model selection approach the selected model is defined by minimization of penalized empirical risk. Such procedures enjoy nice algorithmic properties especially if both the empirical risk and the penalty function are convex functions of the parameter. A number of "oracle" risk bounds for such methods are available. However, the choice of penalty is critical and there is no unified approach for fixing this penalty. This paper presents another method of model selection based on a saddle point optimization. The basic observation behind the saddle point method is that the empirical risk minimizer is a saddle point of the bivariate function built as the difference between empirical risks of two models. An extension of this idea is to define the model selector via a saddle point of a penalized difference. The penalty is also a bivariate function and it has to be selected by the condition that the "oracle" model can be rejected against a larger model only with a very small probability.*

Keywords. *Model selection; Unbiased risk estimation; Penalty calibration.*

2.30

Conditional moment restrictions estimation: finite sample theory

V. Patilea[1,*], V. Spokoiny[2,3] and N. Zhivotovskiy[3]

[1] *CREST & Ensai, Campus de Ker-Lann, 35172 Bruz, France; valentin.patilea@ensai.fr*
[2] *Weierstrass Institute for Applied Analysis and Stochastics & Humboldt University, Berlin; spokoiny@wias-berlin.de*
[3] *Moscow Insitute of Physics and Technology, Institute for Information Transmission Problems of RAS, Moscow; nikita.zhivotovskiy@phystech.edu*
[*] *Corresponding author*

Abstract. *In this contribution, we are interested by statistical models where parameters are identified by a set of conditional estimating equations, also called moment restrictions. Such statistical models are semiparametric in the sense that one aims at estimating a finite dimensional vector of parameter without specifying entirely the distribution of the variables of interest. Nonlinear mean or quantile regressions, econometric instrumental variables models, are only few examples of statistical models that fit into this framework. We consider the smooth minimum distance (SMD) estimation method introduced by Lavergne & Patilea (2013) for inference on the parameters of the model. We investigate the statistical properties of the SMD estimates using modern tools as proposed by Spokoiny (2012). The main feature of the new approach is the non-asymptotic nature of the results. Moreover, the model could be misspecified. We deduce concentration and confidence sets, risk bounds and local expansions of the minimum of the empirical criterion and the corresponding estimate. The previous asymptotic results can be easily derived as corollaries of the new non-asymptotic statements. At the same time, the new approach works well in the situations with large or growing parameter dimension in which the previous parametric theory fails. The results apply for any dimension of the parameter space and provide a quantitative lower bound on the sample size yielding the root-n accuracy.*

Keywords. *Semiparametric models; Conditional moment equations; U−statistics; Non-asymptotic results.*

References

Lavergne, P. & Patilea, V. (2013). Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory. *Journal of Econometrics* **177**, 47–59.

Spokoiny, V. (2012). Parametric Estimation. Finite sample Theory. *Annals of Statistics* **40**, 2877–2909.

2.31

Critical dimension in parametric and semiparametric Bernstein - von Mises Theorem

M. Panov[1,*] and V. Spokoiny[2]

[1] *Moscow Institute of Physics and Technology, Institute for Information Transmission Problems of RAS, Datadvance Company, Pokrovsky blvd. 3 building 1B, 109028 Moscow, Russia; panov.maxim@gmail.com*
[2] *Moscow Institute of Physics and Technology, Weierstrass Institute and Humboldt University Berlin, Mohrenstr. 39, 10117 Berlin, Germany; spokoiny@wias-berlin.de*
[*] *Corresponding author*

Abstract. *The classical semiparametric Bernstein - von Mises (BvM) result is reconsidered in a non-classical setup allowing finite samples and model misspecification. The bracketing approach of Spokoiny (2012) is used for obtaining a finite sample version of BvM theorem even if the full parameter dimension grows with the sample size. We establish an upper bound on the error of Gaussian approximation of the posterior distribution of the target parameter which is explicit in the dimension of the full and target parameters. This helps to identify the so called* critical dimension p_n *of the full parameter for which the BvM result is applicable. In many important statistical models like i.i.d., linear regression or Generalized Linear models we show that the condition "p_n^3/n is small" is sufficient for BvM result to be valid under general assumptions on the model, which improves previous results by Ghosal (1999) and Ghosal (1997). We also provide an example of a model with the phase transition effect: the statement of the BvM theorem fails when the dimension* p_n *approaches* $n^{1/3}$*. The results are extended to the case of semiparametric estimation with certain restrictions on the smoothness of the nonparametric component.*

Keywords. *Prior, Posterior, Bayesian inference, Semiparametric, Critical dimension.*

References

Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *The Annals of Statisics*, **40(6)**: 2877-2909. ArXiv:1111.3029.

Ghosal, S. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, **5(2)**: 315-331.

Ghosal, S. Normal approximations to the posterior distributions for Generalized Linear Models with many covariates. *Mathematical methods of statistics*, **6(3)**: 332-348.

2.32

Functional linear instrumental regression

Jan Johannes

Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve; jan.johannes@uclouvain.be

Abstract. *The estimation of a slope function β is considered in functional linear instrumental regression, where in the presence of a functional instrument W the dependence of a scalar response Y on the variation of an endogenous explanatory random function X is modelled by $Y = \int_0^1 \beta(t)X(t)dt + \sigma U$, $\sigma > 0$, for some error term U. Taking into account that the functional regressor X and the error term U are correlated in many economical applications, the random function W and the error term U are assumed to be uncorrelated. Given an iid. n-sample of (Y, X, W) a lower bound of the maximal mean integrated squared error is derived for any estimator of β over certain ellipsoids of slope functions. This bound is essentially determined by the mapping properties of the cross-covariance operator associated to the functional regressor X and the best linear predictor W_o of X given the instrument W. Assuming first that W_o is known in advance a least squares estimator of β is introduced based on a dimension reduction technique and additional thresholding. It is shown that this estimator can attain the lower bound up to a constant under mild additional moment conditions. The best linear predictor of X given the instrument W is generally, however, not known. Therefore, in a second step W_o is replaced by an estimator and sufficient conditions are provided to ensure the minimax-optimality of the resulting two stage least squares estimator. The results are illustrated by considering Sobolev ellipsoids and finitely or infinitely smoothing cross-covariance operators.*

Keywords. *Functional linear model; Instrumental variable; Linear Galerkin approach; Minimax theory.*

References

Cardot H. and J. Johannes. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis* **101(2)**, 395–408.

Comte F. and J. Johannes (2012). Adaptive functional linear regression. *The Annals of Statistics* **40(6)**, 2765–2797.

Hall P. and J. L. Horowitz. (2005) Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics* **33(6)**, 2904–2929.

Johannes J. (2014) Functional linear instrumental regression. *Technical report, Université catholique de Louvain.*

2.33

Universal asymptotics for high-dimensional sign tests

Davy Paindaveine[1,*] and Thomas Verdebout[2]

[1] *Université libre de Bruxelles; dpaindav@ulb.ac.be*
[2] *Université Lille 3; thomas.verdebout@univ-lille3.fr*
[*] *Corresponding author*

Abstract. *In a small-n large-p hypothesis testing framework, most procedures in the literature require quite stringent distributional assumptions, and restrict to a specific scheme of* (n, p)*-asymptotics. More precisely, multinormality is almost always assumed, and it is imposed, typically, that* $p/n \to c$*, for some* c *in some given convex set* $C \subset (0, \infty)$*. Such restrictions clearly jeopardize practical relevance of these procedures. In this paper, we consider several classical testing problems in multivariate analysis, directional statistics, and multivariate time series : the problem of testing uniformity on the unit sphere, the spherical location problem, the problem of testing that a process is white noise versus serial dependence, the problem of testing for multivariate independence, and the problem of testing for sphericity. In each case, we show that the natural sign tests enjoy nonparametric validity and are distribution-free in a "universal"* (n, p)*-asymptotics framework, where* p *may go to infinity in an arbitrary way as* n *does. Simulations confirm our asymptotic results.*

Keywords. *Double asymptotics; High-dimensional data; Robust statistics; Sign tests.*

2.34

A Depth notion for random networks and some nonparametric results

D. Fraiman[1], R. Fraiman[2,*]

[1] *Departamento de Matemática, Universidad de San Andrés, Argentina; dfraiman@udesa.edu.ar*
[2] *Centro de Matemática, Facultad de Ciencias, UdeLAR, Uruguay; rfraiman@cmat.edu.uy*

Abstract. *We herein consider random networks in a nonparametric setup, when the number of nodes of the network is fixed. We introduce some notion of center, scale, covariance and ordering. A new notion of depth is introduced and its asymptotic behavior is analyzed for a sample of dependent random graphs, which typically corresponds to the evolution of the network in time. Some examples and applications are considered. References should be included as Rosemblatt (1956) or as (Peligrad, 1986).*

Keywords. *Random networks; Erdos–Renyi; Depth notions; Mixing processes.*

References

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *arXiv:1102.2650v5.*

Peligrad M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. *Prog. Probab. Statist* **11**, 193–224.

Rosenblatt M. S. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42**, 43–47.

2.35

Some recent results on kernel methods for independent or for dependent data with special emphasis on additive models

Andreas Christmann[1,*] and Ding-Xuan Zhou[2]

[1] *Department of Mathematics, University of Bayreuth, Germany; andreas.christmann@uni-bayreuth.de*
[2] *Department of Mathematics, City University of Hong Kong, China; mazhou@cityu.edu.hk*
[*] *Corresponding author*

Abstract. *Regularized kernel based methods including support vector machines play an important role in modern nonparametric statistics and in machine learning theory, see e.g. Vapnik (1998), Cucker and Zhou (2007), and Steinwart and Christmann (2008). Although often treated in a purely nonparametric case, they can also be used for additive models which are popular in semiparametric statistics, see e.g. Christmann and Hable (2012). Some recent results on consistency and robustness of regularized kernel based methods for independent or for dependent data will be given as well as learning rates for regularized kernel based methods for additive models. These learning rates compare favourably in particular in high dimensions to recent results on optimal learning rates for purely nonparametric regularized kernel based methods using the Gaussian radial basis function kernel, provided the assumption of an additive model is valid.*

Keywords. *Kernel; Additive Model; Learning rate; Robustness; Nonparametric; Semiparametric.*

References

Christmann, A. and Zhou, D.X. (2014). Learning rates for the risk of support vector machines in additive models. *Preprint.*

Christmann, A. and Hable, R. (2012). Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis* **56**, 854 – 873.

Cucker, F. and Zhou D. X. (2007). *Learning Theory. An Approximation Theory Viewpoint.* Cambridge University Press. Cambridge.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Springer. New York.

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley & Sons. New York.

2.36

Robust estimators in additive models with missing responses

N1. Graciela Boente[1,*], N2. Alejandra Martínez[1] and N3. Matías Salibian–Barrera[2]

[1] *Universidad de Buenos Aires and CONICET; gboente@dm.uba.ar, ale_m_martinez@hotmail.com*
[2] *University of British Columbia; matias@stat.ubc.ca*
[*] *Corresponding author*

Abstract. *As is well known, kernel estimators of the regression function in nonparametric multivariate regression models suffer from the so-called curse of dimensionality, which occurs because the number of observations lying in neighbourhoods of fixed radii decreases exponentially with the dimension. Additive models are widely used to avoid the difficulty of estimating regression functions of several covariates without using a parametric model. They generalize linear models, are easily interpretable, and are not affected by the curse of the dimensionality. Different estimation procedures for these models have been proposed in the literature, and some of them have also been extended to the situation when the data may contain missing responses. It is easy to see that most of these estimators can be unduly affected by a small proportion of atypical observations, since they are based on local averages or local polynomials. For that reason, robust procedures to estimate the components of an additive model are needed. We consider robust estimators for additive models based on local polynomials that can also be used on data sets with missing responses. These estimators simultaneously avoid the curse of dimensionality and the sensitivity to atypical observations. Our proposal is based on the method of marginal integration, and adapted to the missing responses situation. If time permits, we will also introduce a robust kernel estimator for additive models via the back-fitting algorithm.*

Keywords. *Additive Models; Backfitting; Marginal Integration; Missing Responses; Robustness*

2.37

Change-point inference for time-varying erdos-renyi graphs

Moulinath Banerjee[1]

[1] *moulib@umich.edu*

Abstract. *We investigate an Erdos-Renyi graph, where the edges can be in a set of finite states (e.g. present/absent). The state of each edge evolves as a Markov chain independently of the other edges, and whose parameters exhibit a common change-point in time. We derive the maximum likelihood estimator for the change-point and characterize its distribution. Depending on the signal-to-noise ratio present in the data, different limiting regimes emerge. Nevertheless, a* *unifying adaptive scheme can be used in practice that covers all cases. Our treatment incorporates the situation where the number of edges is allowed to increase with n: the number of time measurements. Illustrations of the model and our approach are demonstrated on both synthetic and real data.*

This is joint work with Elena Yudovina and George Michailidis.

2.38

Graph estimation with joint additive models

Arend Voorman[1], Ali Shojaie[1,*] and Daniela Witten[1]

[1] *Department of Biostatistics, University of Washington; voorma@uw.edu, ashojaie@uw.edu, dwitten@uw.edu*

[*] *Corresponding author*

Abstract. *In recent years, there has been considerable interest in estimating conditional independence graphs in the high-dimensional setting. Most prior work has assumed that the variables are multivariate Gaussian, or that the conditional means of the variables are linear. Unfortunately, if these assumptions are violated, the resulting conditional independence estimates can be inaccurate. I will present a semi-parametric method, SpaCE JAM, which allows the conditional means of the features to take on an arbitrary additive form. I will discuss computational and theoretical aspects of the problem, as well as extensions of the method to estimation of directed acyclic graphs with known causal ordering. Using simulated and real data examples, I will demonstrate that SpaCE JAM enjoys superior performance to existing methods when there are non-linear relationships among the features, and is comparable to methods that assume multivariate normality when the conditional means are linear.*

Keywords. *Conditional independence; Graphical model; Nonlinearity; Non-Gaussianity; Sparse additive model.*

2.39

Multiview and dynamic network analysis with matrix factorization

S. Mankad[1] and G. Michailidis[2]

[1] *University of Maryland; smankad@rhsmith.umd.edu*

[2] *University of Michigan; gmichail@umich.edu*

Abstract. *Social media platforms, such as Twitter and Facebook, have significantly affected the exchange of ideas, and even assisted in social revolution. In response, there is a growing literature that attempts to understand and exploit social networking platforms for resource optimization and marketing, as it is a major interest for private enterprises and political cam-*

paigns attempting to propagate particular opinions or products. However, most studies of social networking data focus on static relationships, such as following-follower networks in Twitter, which are not necessarily well-correlated with actual patterns of conversation. A popular hypothesis is that it is, therefore, necessary to move beyond standard (static) network analysis in order to comprehensively answer most interesting questions. Along these lines, we discuss a modeling approach based on matrix factorization that captures influence from multiple networks that are sequential (time-series) or multiview (static, but featuring different link relations between the same set of nodes). We find that the approach with multiview data can accurately identify influential users, a critical problem in marketing and intelligence gathering.

Keywords. *Network analysis; Matrix factorization.*

2.40

Estimation in high-dimensional vector autoregresive models

George Michailidis

Department of Statistics, The University of Michigan

Abstract. *Vector Autoregression (VAR) is a widely used method for learning complex interrelationship among the components of multiple time series. Over the years it has gained popularity in the fields of control theory, statistics, economics, finance, genetics and neuroscience. We consider the problem of estimating stable VAR models in a high-dimensional setting, where both the number of time series and the VAR order are allowed to grow with sample size. In addition to the "curse of dimensionality" introduced by a quadratically growing dimension of the parameter space, VAR estimation poses considerable challenges due to the temporal and cross-sectional dependence in the data. Under a sparsity assumption on the model transition matrices, we establish estimation and prediction consistency of ℓ_1-penalized least squares and likelihood based methods. Exploiting spectral properties of stationary VAR processes, we develop novel theoretical techniques that provide deeper insight into the effect of dependence on the convergence rates of the estimates. We study the impact of error correlations on the estimation problem and develop fast, parallelizable algorithms for penalized likelihood based VAR estimates.*

Keywords. *Sparsity; VARs; Spectral Density*

2.41

Quadratic-type classifications for non-Gaussian, high-dimensional data

M. Aoshima[1,*] and K. Yata[1]

[1] *University of Tsukuba, Japan; aoshima@math.tsukuba.ac.jp, yata@math.tsukuba.ac.jp*
[*] *Corresponding author*

Abstract. *In this talk, we consider multiclass classification when the sample size is much lower than the dimension. The inverse matrix of sample covariance matrix does not exist, and hence typical discriminant analysis such as Fisher discriminant analysis does not work straightforwardly. When the population covariance matrices are common, Bickel and Levina (2004, Bernoulli) considered the inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. When the population covariance matrices are not common, Dudoit et al. (2002, JASA) considered using the inverse matrix defined by only diagonal elements of the sample covariance matrices. Aoshima and Yata (2011) created a new quadratic-type discriminant rule by using the difference of a geometric representation between the two classes. Aoshima and Yata (2013) developed distance-based classifiers for multiclass, non-Gaussian, high-dimensional data.*

In this talk, we discuss a quadratic-type classifier for multiclass, non-Gaussian, high-dimensional data. We do not assume the equality of population covariance matrices. The classifier considered in this talk includes preceding classifiers as a special case. We first show that the classifier holds the consistency property in misclassification rates for multiclass, non-Gaussian data when the dimension goes to infinity. We verify that the classifier is asymptotically distributed as a normal distribution when the dimension goes to infinity. With the help of the asymptotic normality, we evaluate misclassification rates of the classifier. We emphasize that the classifier based on Mahalanobis' distance does not always give a preferable performance. We also consider quadratic-type classifiers by feature selection or by regularized estimation of covariance matrices.

Keywords. *Asymptotic normality; Geometric representations; HDLSS; Heteroscedasticity; Large p, small n.*

References

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis* (*Editor's special invited paper*) **30**, 356–399.

Aoshima, M. and Yata, K. (2013). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, in press.

2.42

Two-sample thresholding tests for high dimensional means

Song Xi Cheni[1], Jun Li[2] and Ping-Shou Zhongi[3]

[1] *Peking University and Iowa State University,csx@gsm.pku.edu.cn*

Abstract. *We consider testing for two-sample means of high dimensional populations by thresholding. Two tests are investigated which are designed for better power performance when the two sample mean vectors differ only in sparsely populated coordinates. The first test is constructed by carrying out thresholding to remove those non-signal bearing dimensions. The second test combines data transformation with the precision matrix and thresholding. The benefits of the thresholding and the data transformations are demonstrated by reduced variance of the test* *statistics, and the improved power and wider detection region of the tests. Numerical analyses and empirical study are performed to confirm the theoretical findings and to demonstrate the practical implementations*

2.43

Testing equality of a large number of densities

J. D. Hart[1,*] and D. Zhan[2]

[1] *Department of Statistics, Texas A&M University; hart@stat.tamu.edu*
[2] *Data and Research Services, Texas A&M University; dzhan@tamu.edu*
[*] *Corresponding author*

Abstract. *The problem of testing equality of a large number of densities is considered. The classical k-sample problem compares a small, fixed number of distributions and allows the sample size from each distribution to increase without bound. In our asymptotic analysis the number of distributions tends to infinity but the size of individual samples remains fixed. The proposed test statistic is motivated by the simple idea of comparing kernel density estimators from the various samples to the average of all density estimators. However, a novel interpretation of this familiar type of statistic arises upon centering it. The asymptotic distribution of the statistic under the null hypothesis of equal densities is derived, and power against local alternatives is considered. It is shown that a consistent test is attainable in many situations where all but a vanishingly small proportion of densities are equal to each other. The test also has applications to settings where the data come from distributions generated by a hidden Markov chain.*

Keywords. *Kernel density estimation; k-sample problem; Local alternatives; Omnibus test; U-statistics.*

2.44

LP nonparametric network modeling

Subhadeep Mukhopadhyay[1,*] and Emanuel Parzen[2]

[1] *Temple University, Philadelphia, PA, USA; deep@temple.edu*
[2] *Texas A&M University, College Station, TX, USA; eparzen@stat.tamu.edu*
[*] *Corresponding author*

Abstract. *Network (or Graph) modeling is one of the growing interdisciplinary field of research. In this paper we introduce a new nonparametric framework to understand the the 'structure' of the network. Our innovation is fundamentally distinct from all the previous research in this direction: Traditional approaches either use adjacency matrix calculus or parametric* *model, whereas our approach is based on functional nonparametric statistical algorithm. The fundamental conditional comparison density function plays a vital role. We provide representation of conditional comparison density as infinite linear combination of especially designed piece-wise constant orthonormal LP score functions, which provides a low-rank parametrization of network connectivity pattern.*

We provide LP nonparametric algorithm for (i) community detection; (ii) smooth estimation of Graphon-like object $[0,1]^2 \to [0,1]$ with checkerboard shape (bivariate step-function) that completely characterizes the 'structure' of the network. Moreover, we provide novel statistical interpretation of our algorithm from the perspective of 'shape' of conditional comparison density captured by LP score coefficients.

Application to real-world network will be demonstrated. .

Keywords. *LP orthonormal score function; Comparison density; LP score coefficients; Nonparametric graphon estimation; Community detection.*

References

Parzen, E. and Mukhopadhyay, S. (2013a). LP Mixed Data Science: Outline of Theory. *arXiv:1311.0562.*

Parzen, E. and Mukhopadhyay, S. (2013b). United Statistical Algorithms, LP comoment, Copula Density, Nonparametric Modeling. *59th ISI World Statistics Congress (WSC), Hong Kong.*

Parzen, E. and Mukhopadhyay, S. (2013c). United Statistical Algorithms, Small and Big Data, Future of Statisticians. *arXiv:1308.0641.*

2.45

Convex function estimation

Dragi Anevski[1]

[1] *Mathematical Sciences, Lund University, Sweden, dragi@maths.lth.se*

Abstract. *We discuss limit distribution results for a nonparametric regression estimator of an unknown convex function, e.g. a regression function or a density function. We discuss previously obtained results for independent data as well as state new results for dependent data, for both weakly and long range dependent data. An attempt to give a unified approach, with respect to type of dependence and type of estimand, for the estimation problem is made.*

Keywords. *Limit distribution; Convexity; Dependent data*

2.46

Semiparametric Bernstein-von Mises theorem: second order studies

Yun Yang[1], Guang Cheng[2] and David Dunson[1]

[1] *Department of Statistical Science, Duke University; yy84@stat.duke.edu, dunson@stat.duke.edu*
[2] *Department of Statistics, Purdue University; chengg@purdue.edu*

Abstract. *Semiparametric Bernstein von-Mises Theorem has been successfully developed by Bickel and Kleijn (2012) in a general setup among others, e.g., Cheng and Kosorok (2008a, b). This talk mainly focuses on its second order extension with an attempt to figure out the influence of nonparametric Bayesian prior on the semiparametric inference, i.e., parametric component. Such results can provide us new theoretical insight in guiding the choice of objective prior in a general semiparametric setup.*

2.47

Confidence bands for distribution functions: A new look at the law of the iterated logarithm

L. Dümbgen[1,*] and J.A. Wellner[2]

[1] *University of Bern; duembgen@stat.unibe.ch*
[2] *University of Washington; jaw@stat.washington.edu*
[*] *Corresponding author*

Abstract. *We present a general law of the iterated logarithm for stochastic processes on the open unit interval having subexponential tails in a locally uniform fashion. It applies to standard Brownian bridge but also to suitably standardized empirical distribution functions. This leads to new goodness-of-fit tests and confidence bands which refine the procedures of Berk and Jones (1979) and Owen (1995). Roughly speaking, the high power and accuracy of the latter procedures in the tail regions of distributions are esentially preserved while gaining considerably in the central region.*

Keywords. *Confidence band; Limit distribution; Sub-exponential tails; Submartingale, Tail regions*

References

Berk, R. H., and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47**, 47–59.

2.48

Two-stage plans for estimating the inverse of a monotone function

George Michailidis

Department of Statistics, The University of Michigan

Abstract. *This study investigates two-stage plans based on nonparametric procedures for estimating an inverse regression function at a given point. Specifically, isotonic regression is used at stage one to obtain an initial estimate followed by another round of isotonic regression in the vicinity of this estimate at stage two. It is shown that such two stage plans accelerate the convergence rate of one-stage procedures and are superior to existing two-stage procedures that use local parametric approximations at stage two when the available budget is moderate and/or the regression function is Ôill-behavedÕ. Both Wald and Likelihood Ratio type confidence intervals for the threshold value of interest are investigated and the latter are recommended in applications due to their simplicity and robustness. The developed plans are illustrated through a comprehensive simulation study and an application to car fuel efficiency data.*

Joint work with Mouli Banerjee, Runlong Tang and Shawn Mankad

Keywords. *Adaptive sampling; nonparametric estimation; convergence rate acceleration*

2.49

Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions

A. Jokiel-Rokita*, M. Pulit

Institute of Mathematics and Computer Science
Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27,
50-370 Wroclaw, Poland;
alicja.jokiel-rokita@pwr.wroc.pl, michal.pulit@pwr.wroc.pl*
** Corresponding author*

Abstract. *The receiver operating characteristic (ROC) curve is a graphical representation of the relationship between false positive and true positive rates. It is a widely used statistical tool for describing the accuracy of a diagnostic test. In this paper we propose a new nonparametric ROC curve estimator based on the smoothed empirical distribution functions. We prove its strong consistency and perform a simulation study to compare it with some other popular nonparametric estimators of the ROC curve. We also apply the proposed method to a real data set.*

Keywords. Receiver operating characteristic (ROC) curve; Empirical distribution function; Nonparametric estimation.

2.50

Nonparametric inference for covariate-specific summary indices of ROC curves

Juan Carlos Pardo-Fernández[1,*], Elisa M. Molanes-López[2] and Emilio Letón[3]

[1] Universidade de Vigo, Spain; juancp@uvigo.es
[2] UC3M, Madrid, Spain; emolanes@est-econ.uc3m.es
[3] UNED, Madrid, Spain; emilio.leton@dia.uned.es
[*] Corresponding author

Abstract.
In medical studies, the diagnostic of a patient is very often based on some characteristic of interest, which may lead to classification errors. These classification errors are calibrated on the basis of two indicators: sensitivity (probability of diagnosing an ill person as ill) and specificity (probability of diagnosing a healthy person as healthy). When the diagnostic variable is continuous, the classification will necessarily be based on a cut-off value: if the diagnostic variable exceeds the cut-off then the patient is classified as ill, otherwise the patient is classified as healthy. The representation of the sensitivity versus one minus the specificity for each cut-off value leads to the operating characteristic (ROC) curve, which is a statistical tool extensively used to analyze the discriminative power of the diagnostic variable. Some summary indicators, such as the area under the curve or the Youden index, are employed to describe the main features of the ROC curve.

In many studies, a covariate is available along with the diagnostic variable. The information contained in the covariate may modify the discriminatory capability of the ROC curve, and therefore it is interesting to study the impact of the covariate on the conditional or covariate-specific ROC curve. This work will be devoted to the study of a nonparametric estimation procedure of the covariate-specific ROC curve and its associated summary indices, specifically, the covariate-specific AUC and the covariate-specific Youden index.

Keywords. *Area under the curve; Nonparametric regression; ROC curve; Youden index.*

2.51

Functional partial area under the curve regression: a metabolic syndrome case study

Vanda Inácio de Carvalho[1,*], Miguel de Carvalho[2], Todd A. Alonzo[3] and Wenceslao González-Manteiga[4]

[1] Department of Statistics, Pontificia Universidad Católica de Chile; icalhau@mat.puc.cl
[2] Department of Statistics, Pontificia Universidad Católica de Chile; mdecarvalho@mat.puc.cl

[3]Division of Biostatistics, University of Southern California; talonzo@childrensoncologygroup.org
[4]Department of Statistics and Operations Research, University of Santiago de Compostela; wenceslao.gonzalez@usc.es
[*]Corresponding author

Abstract. *The statistical evaluation of diagnostic tests is of great importance in public health and medical research. New diagnostic tests must be rigorously evaluated to determine their abilities to discriminate between diseased and nondiseased states and it is of crucial importance to understand the covariate influence to determine the optimal and suboptimal populations to perform such tests on. Due to advances in technology, medical diagnostic data have become increasingly complex and, nowadays, applications where measurements are functions, curves, or images are becoming more and more common. We develop nonparametric regression methods for the partial area under the receiver operating characteristic curve, a well-accepted measure of diagnostic test accuracy when only a region of the curve is of interest, for the case where the covariate influencing the test's performance is functional. The simulation study shows that our method produces estimates with small bias and small mean squared error. The application of our method to assess the ability of the gamma-glutamyl-transferase to detect women with metabolic syndrome reveals that the nocturnal levels of arterial oxygen saturation of hemoglobin are key to the test's performance.*

Keywords. *Diagnostic testing; Functional data; Kernel regression; Metabolic syndrome; Partial area under the curve.*

2.52

Bayesian nonparametric Youden index modeling

V. Inácio de Carvalho[1,*], M. de Carvalho[1] and A. Branscum[2]

[1] *Pontificia Universidad Católica de Chile; icalhau@mat.puc.cl, mdecarvalho@mat.puc.cl*
[2] *Oregon State University; adam.branscum@oregonstate.edu*
[*]*Corresponding author*

Abstract. *Accurate diagnosis of disease is of crucial importance in health care and medical research. The major goal of a diagnostic test is to distinguish diseased patients from nondiseased patients and, before a test is routinely used in practice, its ability to distinguish between these two states must be rigorously evaluated. The accuracy of the test at any given threshold can be measured by the probability of a true positive (sensitivity) and a true negative (specificity). The receiver operating characteristic (ROC) curve, a popular graphical tool for evaluating the discriminatory ability of a continuous scale diagnostic test, plots the sensitivity,* $\mathrm{Se}(c)$*, against 1−specificity,* $1-\mathrm{Sp}(c)$*, as the threshold* c *varies through the range of possible test results. To evaluate the discriminatory ability of a test it is common to summarize the information of the ROC curve into a single global value or index. The two most popular summary indices of diagnostic accuracy are the area under the ROC and the Youden index. In this work we focus on the Youden index (YI), which can be defined as* $\mathrm{YI} = \max_c\{\mathrm{Se}(c) + \mathrm{Sp}(c) - 1\}$ *and has the attractive feature of providing a criterion for choosing the optimal threshold value* c^* *to screen* *subjects in practice. With the aim of having a flexible model that can handle skewness, multimodality and other nonstandard features of the data, without the need of knowing in advance their existence, we propose to estimate the Youden index and its associated optimal threshold c^*, using Bayesian nonparametric techniques, namely, Dirichlet process mixtures. The performance of the estimator is evaluated through a simulation study and a real data application is provided.*

Keywords. *Bayesian nonparametrics; diagnostic testing; Dirichlet process mixtures; Markov chain Monte Carlo; Youden index.*

2.53

Some tools for the assessment of contamination models in general dimension

Eustasio del Barrio[1]

[1] *IMUVA, Universidad de Valladolid; tasio@eio.uva.es*

Abstract. *Starting from the classical distinction between statistical and practical significance in a goodness-of-fit test, we consider several proposals of a tolerance zone around a parametric model. We show the connection of one type of these tolerance zones, expressable in terms of contamination, and the concept of trimmings of a probability. We present some basic properties of trimmings. Different contaminated models can be formulated in terms of different parametrizations of sets of probability trimmings. Deviation from these models can be expressed in terms of a metric. We study the case of transportation cost and Kolmogorov metrics and discuss the associated empirical partial mass problems in general dimension. Finally, we show how to use trimming techniques for testing fit to several types of contaminated models.*

Keywords. *Essential Model Validation; Contaminated model; Trimming; Partial mass problems*

2.54

Limit theorems for strictly stationary random fields satisfying strong mixing conditions

Cristina Tone

Abstract. *In this talk we will introduce a common technique, the Bernstein "blocking argument", in proving CLT's for dependent random fields. Under just ρ'-mixing and finite seconds moments, we will obtain a CLT for strictly stationary random fields, which will help us construct an invariance principle for empirical processes endowed from that strictly stationary random field. Using the same blocking argument, for a sequence of strictly stationary random fields that* *are uniformly rho'-mixing and satisfy a Lindeberg condition, a CLT is obtained for sequences of "rectangular" sums from the given random fields. This result is then used to prove a CLT for some kernel estimators of probability density for ρ'-mixing random fields with probability density and joint densities which are absolutely continuous.*

2.55

Can honest Bayes methods apply to complex models?

Y. Ritov[1], P. Bickel[2], A. Gamst[3], B. Kleijn[4]

[1,*] *The Hebrew University of Jerusalem, yaacov.ritov@gmail.com*
[2] *University of California at Berkeley, bickel@stat.berkeley.edu*
[3] *University of California at San Diego, acgamst@math.ucsd.edu*
[4] *University of Amsterdam, B.Kleijn@uva.nl* * *Corresponding author*

Abstract. *We consider the Bayesian analysis of a few complex, high-dimensional models and show that intuitive priors, which are not tailored to the fine details of the model and the estimated parameters, produce estimators which perform poorly in situations in which good, simple frequentist estimators exist. The talk would concentrate with the partial linear model. We present a strong version of Doob's consistency theorem which demonstrates that the existence of a uniformly square root of n-consistent estimator ensures that the Bayes posterior is square root of n-consistent for values of the parameter in subsets of prior probability 1. We also demonstrate that it is, at least, in principle, possible to construct Bayes priors giving both global and local minimax rates, using a suitable combination of loss functions. We argue that there is no contradiction in these apparently conflicting findings.*

Keywords. *Doob's theorem, Bayesian inference; Partial linear models; Complex models*

2.56

Non-parametric cross-covariance functions for multivariate spatial data

H. Zhang[1,*], Y. Wang[2]

[1] *Department of Statistics, Purdue University, USA; zhanghao@purdue.edu*
[2] *Department of Mathematics, East Kentucky University, USA; wang.yong@eky.edu*
* *Corresponding author*

Abstract. *Multivariate spatial data such as temperature and precipitation are observed in many studies. Cokriging is the technique for best linear unbiased prediction using multivariate spatial data, which utilizes both marginal and cross covariance functions. Recent years see a significant advance in parametric models of multivariate covariance functions. Statistical inferences for* *these parameters can be problematic due to the increased number of parameters that have to satisfy some constraints in order for the fitted model to be a valid multivariate covariance function. We propose a new and different approach. We first fit the marginal models by some parametric family (e.g., Matern family) which is relatively easier to do. We then estimate the cross-covariance function non-parametrically and the resulting multivariate covariance function is guaranteed to be positive definite. We evaluate the performance of our approach through some real data sets and compare it with some existing parametric models.*

Keywords. *Cross Covariance Function; Cokring; Multivariate Spatial Model;*

2.57

Asymptotic efficiency of tests and asymptotic equivalence of locally stationary Gaussian processes

Inder Tecuapetla-Gómez[1]

[1] *University of Göttingen, itecuap@mathematik.uni-goettingen.de*

Abstract. *In the first part of this talk we show how a large deviation result for the log-likelihood ratio between the laws of two locally stationary Gaussian processes (LSGP) can be used to obtain the analogue of classical results on the asymptotic efficiency of tests such as Stein's lemma, the Chernoff bound and the general Hoeffding bound.*

We dedicate the second part of this talk to show that asymptotically, and with respect to Le Cam's pseudodistance, the problem of estimating parametrically the log-time-varying spectral density function of a LSGP can be thought of as a Gaussian white noise problem plus drift.

Keywords. *local stationarity, Le Cam's pseudodistance, large deviations, asymptotic efficiency*

2.58

On the estimation of long-memory locally stationary processes

Wilfredo Palma[1]

[1] *Pontificia Universidad Católica de Chile, wilfredo@mat.puc.cl*

Abstract. *This talk addresses some estimation problems in the context of locally stationary time series exhibiting long range dependence. Different aspects of the estimation process are discussed, including the estimation of trends and model parameters. The theoretical results established are illustrated with several Monte Carlo experiments and applications to real life time series data.*

Keywords. *Long memory, local stationarity*

2.59

Constructing adaptive interference-reduced Wigner-Ville spectral estimators of non-stationary time series

Jean-Marc Freyermuth[1]

[1] *Katholique Université of Leuven:, Jean-Marc.Freyermuth@kuleuven.be*

Abstract. *In this talk we propose estimators of the time-frequency spectrum of a (zero mean) non-stationary time series with second order structure which varies across time. It is obtained by smoothing the empirical Wigner-Ville (WV) spectrum (Matin and Flandrin, 1985) which is a highly localized time-frequency spectrum. Using the empirical WV avoids prior time-frequency segmentation (such as for the segmented periodogram (Schneider and von Sachs, 1996)) nevertheless it suffers from low and heterogeneous signal-to-noise ratios and from severe interferences. In addition, the associated time-frequency spectrum is best modeled as an anisotropic object with locally varying smoothness in both time and frequency directions (Neumann and von Sachs, 1997). All this make smoothing very challenging. Our approach is to project the empirical WV data onto a specifically designed hyperbolic wavelet basis (Autin et al., 2013b) and to use a tree-structured thresholding (Autin et al., 2011, 2013) under constraints inspired notably by the Heisenberg's uncertainty principle. Such approach is expected to ensure an adaptive time-frequency representation and to reduce the cross-interferences of the WV spectrum.*

Keywords. *Wigner-Ville spectra, non-stationarity, tree-structured thresholding, Heisenberg's uncertainty principle*

References

Autin, F., Freyermuth, J-M., von Sachs, R. (2011). Ideal denoising within a family of tree-structured wavelet estimators . *Electronic Journal of Statistics*, 5, 829-855.

Autin, F., Claeskens, G., Freyermuth, J-M. (2013). Hyperbolic wavelet thresholding rules: the curse of dimensionality through the maxiset approach. *Applied Computational Harmonic Analysis*, e-pub.

Autin, F., Claeskens, G., Freyermuth, J-M. (2013). On the performances of isotropic and hyperbolic wavelet estimators. *Tech report, KBI.*

Martin, W., Flandrin, P. (1985). Wigner-Ville spectral analysis of non-stationary processes. *IEEE Trans. on Acoust., Speech and Signal Proc.*, ASSP-33(6), 1461-1470.

Neumann, M.H., von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Annals of statistics*, 25(1), 38-76.

Schneider, K., von Sachs, R. (1996). Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Applied and Computational Harmonic Analysis*, 3, 268-282.

2.60

Simultaneous quantile inference for locally stationary long memory time series

Zhou Zhou [1]

[1] *University of Toronto, zhou@utstat.toronto.edu*

Abstract. *We consider simultaneous or functional inference of time-varying quantile curves for a class of non-stationary and long memory time series. New uniform Bahadur representations and Gaussian approximation schemes are established for a wide class of non-stationary and long memory linear processes. Furthermore, an asymptotic distributional theory is developed for the maxima of a class of non-stationary long memory Gaussian processes. With the latter theoretical results, we construct simultaneous confidence bands for the above mentioned quantile curves with asymptotically correct coverage probabilities.*

Keywords. *local stationarity, long memory processes quantile regression*

2.61

Analysis of aggregated functional data from mixed populations applied to transformer electricity load data

R. Dias[1], A. Lenzi[2,*], C. P. E. de Souza[3], N. L. Garcia[4] and N. Heckman[5]

[1] *University of Campinas and University of British Columbia; dias@ime.unicamp.br, camila.souza@stat.ubc.ca*
[2] *Technical University of Denmark; amle@dtu.dk, nancy@ime.unicamp.br*
[3] *University of British Columbia, nancy@stat.ubc.ca*
[*] *Amanda Lenzi*

Abstract. *Understanding the energy consumption patterns of different types of consumers is essential in any planning of energy distribution. However, obtaining consumption information for single individuals is often either not possible or too expensive. Therefore, we consider data from aggregations of energy use, that is, from sums of individuals' energy use, where each individual falls into one of C consumer classes. Unfortunately, the exact number of individuals of each class may be unknown: consumers do not always report the appropriate class, due to various factors including differential energy rates for different consumer classes. We develop a methodology to estimate the expected energy use of each class as a function of time and the true number of consumers in each class. We also provide some measure of uncertainty of the resulting estimates. To accomplish this, we assume that the expected consumption is a function of time that can be well approximated by a linear combination of B-splines. Individual consumer perturbations from this baseline are modeled as B-splines with random coefficients. We treat the reported numbers of consumers in each category as random variables with distribution depending* *on the true number of consumers in each class and on the probabilities of a consumer in one class reporting as another class. We obtain maximum likelihood estimates of all parameters via a maximization algorithm. We introduce a special numerical trick for calculating the maximum likelihood estimates of the true number of consumers in each class. We apply our method to a data set and study our method via simulation.*

Keywords. *aggregated functional data; nonparametric regression, energy consumption*

2.62

Switching nonparametric regression models with an application to building power usage data

Camila P. E. de Souza [1,*] and Nancy E. Heckman[2]

[1,2] *University of British Columbia; camila.souza@stat.ubc.ca, nancy@stat.ubc.ca*
[*] *Corresponding author*

Abstract. *We propose a methodology to analyze data arising from a curve that, over its domain, switches among J states. We consider a sequence of response variables, where each response y depends on a covariate x according to an unobserved state z, also called a hidden or latent state. The states form a stochastic process and their possible values are $j = 1, \ldots, J$. If z equals j the expected response of y is one of J unknown smooth functions evaluated at x. We call this model a switching nonparametric regression model. We consider two different data structures: one with N replicates and the other with one single realization. For the hidden states, we consider those that are independent and identically distributed and those that follow a Markov structure. We develop an EM algorithm to estimate the parameters of the latent state process and the functions corresponding to the J states. Standard errors for the parameter estimates of the state process are also obtained. We investigate the frequentist properties of the proposed estimates via simulation studies. As an application we analyze daytime power usage on business days in a building treating each day as a replicate and modeling power usage as arising from two functions, one function giving power usage when the cooling system of the building is off, the other function giving power usage when the cooling system is on.*

Keywords. *Nonparametric regression; Machine learning; Latent variables; EM algorithm; Building power usage data.*

2.63

Modeling and forecasting daily electricity loads via functional clustering and curve linear regression

H. Cho[1,*], Y. Goude[2], X. Brossat[2] and Q. Yao[3]

[1] *School of Mathematics, University of Bristol, UK; haeran.cho@bristol.ac.uk*
[2] *Électricité de France, France.*

[3] *Department of Statistics, London School of Economics, UK; Guanghua School of Management, Peking University, China.*
[*] *Corresponding author*

Abstract. *We propose a methodology for modeling and forecasting daily electricity load. Two main ingredients of our approach are (i) clustering pairs of successive daily load curves into homogeneous sub-groups, and (ii) modeling the dependence between the successive curves within each of such sub-groups via curve linear regression. For the former task, we adopt the k-centers functional clustering (k-CFC) from Chiou and Li (2007) which simultaneously accounts for the dissimilarities between clusters in terms of both the mean and the covariance functions. Besides, we propose some model selection criteria which are applicable to the functional data together with the k-CFC, to identify the number of clusters as well as the optimal clustering of the data. For the latter part of the methodology, the curve linear regression technique from Cho et al. (2013) plays a significant part, which reduces the curve regression problem to a finite number of scalar linear regression problems via singular value decomposition in a Hilbert space. The combined methodology is applied to a range of simulated datasets as well as to the French electricity load data collected between 1996 and 2009, where various model selection methods are investigated and the SVD-based curve linear regression methods are compared to other competitors.*

2.64

Creating and using low voltage network templates to identify network stresses

G. Shaddick[1*]

[1] *Department of Mathematical Sciences, University of Bath, BA2 7AY, UK; g.shaddick@bath.ac.uk*

Abstract. *Distribution network operators need to maintain security and quality of supply as customers connect low carbon technologies (LCT) and distributed generation (DG) to the Low Voltage (LV) networks. LCT and DG penetration and traditional loads vary across the network and information on the potential effects of LCT and DG is currently limited. We seek to address this by developing a number of LV Network Templates. These Templates allow a network planner to estimate the load flow at a substation without the need for costly monitoring. They consist of load profiles (patterns of delivered power over time) for different times and locations and are created by clustering data collected in the UK from multiple locations (ca. 1000 substations and 4000 customer homes) measured at high temporal resolution (10 mins). An important factor is to be able to predict load profiles, together with associated measures of uncertainty, for locations where monitoring has not been performed. As a result of these analyses, load profiles were produced for different times of the day and show the variation between the working week and weekends and across different seasons. These profiles allow planners to estimate the capacity and voltage headroom available and so to determine whether installation of low carbon technologies are expected to cause voltage or thermal issues.*

Keywords. *Low voltage networks, clustering, classification, data reduction*

2.65

Data-dependent smoothing of nonparametric estimates by discrepancy method

N. Markovich[1]

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

Abstract. *The application of nonparametric estimates of the probability density function requires the evaluation of smoothing parameters like bandwidths of kernel estimates. We consider the so-called discrepancy method. The latter was proposed and investigated in Vapnik et al. (1992), (Markovich, N.M., 2007) as an alternative data-dependent smoothing tool to cross-validation. It is based on the usage of well-known nonparametric statistics like the von Mises-Smirnov's (M-S) and the Kolmogorov-Smirnov (K-S) as measures in the space of distribution functions. The unknown smoothing parameter is proposed to find as a solution of the discrepancy equation. On its left-hand side it stands a discrepancy between the empirical distribution function and the nonparametric estimate of the distribution function. The latter is obtained as a corresponding integral of the density nonparametric estimator. The right-hand side is equal to some quantile of the asymptotic distribution of the M-S or K-S statistic. The discrepancy method demonstrates better results than cross-validation for nonsmooth (e.g., triangular and uniform) distributions. The method can avoid the problem of cross-validation falling into local extremes. In Vapnik et al. (1992) it is derived that the rate of convergence in L_2 for a projection estimator with the smoothing parameter found by the M-S discrepancy method is close to the best in the class of densities with a bounded variation of the kth derivative. The properties of the discrepancy method in case of heavy-tailed densities is also discussed. The application of the discrepancy method to the nonparametric estimation of the extremal index is proposed.*

Keywords. *Data-dependent smoothing tool; Discrepancy method; Nonparametric density estimation; Cross-validation; Extremal index*

References

Vapnik, V. N., Markovich, N.M. and Stefanyuk, A.R. (1992). Rate of convergence in L_2 of the projection estimator of the distribution density. *Automation and Remote Control* **53**, 677–686.

Markovich, N.M. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data.* Wiley. Chichester.

2.66

A new class of the tail index estimators and the improvement of the Hill estimator

Vygantas Paulauskas

Vilnius University, Department of Mathematics and Informatics, Naugarduko 24, Vilnius 03225, Lithuania; vygantas.paulauskas@mif.vu.lt

Abstract. *In the talk we discuss a new type of the tail index estimators. By including the logarithmic function, which is essential in construction of most tail index estimators, into a family of power functions we obtain this new class of estimators. These estimators are natural generalizations of the classical Hill, moment, and moment ratio estimators. Moreover, the generalized Hill estimator has smaller asymptotic mean square error comparing with the Hill estimator over all range of the parameters present in the second order regular variation condition. The relation with a recently introduced the so-called dual divergence estimator of the tail index (see Bouzebda and Cherfi (2012)) will be mentioned.*
The talk is based on the paper Paulauskas and Vaičiulis (2013) and some new results.

Keywords. *Tail index estimation; Hill-type estimators; heavy tails.*

References

Bouzebda, S. and Cherfi, M. (2012). Dual Divergence Estimators of the Tail Index. *ISNR Probability and Statistics* , doi:10.5402/2012/746203.

Paulauskas, V. and Vaičiulis, M. (2013). On the improvement of Hill and some others estimators. *Lithuanian Math. J.* **53**, 336–355.

2.67

Prediction of stable random fields and time series

Elena Shmileva[1,*]

[1] *St.Petersburg State University, Chebyshev Laboratory, St.Petersburg 199178, Russia; elena.shmileva@gmail.com*

Abstract. *The purpose of the research is to elaborate methods to predict values of random fields with heavy tailed marginal distributions.*
For stable non-Gaussian random fields we propose a methodology based on a nice theoretical structure of these fields. In our construction we use the integral representation of these fields via stable random measures as well as deterministic spectral measures to describe multivariate

distribution of the field values.
Three methods for $\alpha \in (1,2)$ and several methods for $\alpha \in (0,1)$ (the most heavy tailed case) stable fields were found. Our methods are similar, in some sense, to the kriging techniques but instead of the covariance function (that does not exists for the stable non-Gaussian fields) different dependence measures are used. Thus, covariation and codifference could be considered as quantities reflecting dependence between two stable random variables. Numerical comparison of the methods is proposed.
Papers containing these results are Karcher, Shmileva, Spodarev (2013) or as (Spodarev, Shmileva, Roth, 2013).

Keywords. *Stable distribution; Dependence; Random fields; Geostatistics; Kriging.*

References

Karcher, W. Shmileva, E., Spodarev, E.(2013). Extrapolation of stable random fields. *Journal of Multivariate Analysis* **115**, 516–536.

Spodarev, E., Shmileva, E., Roth S. (2013). *LaTeX Extrapolation of stationary random fields.* A book chapter, preprint, http://arxiv.org/abs/1306.6205.

2.68

On favorable extremes modeling

M. Stehlík[1]

[1] *Department of Applied Statistics,*
Johannes Kepler University Linz, Austria
Department of Statistics
University of Valparaiso, Chile
Email: milan.stehlik@jku.at

Abstract. *Within this talk we will concentrate on several methodological issues of parametric models for Extreme Values. Hill Hill (1975) derived a procedure of Pareto tail estimation by the MLE. Later on, many authors tried to robustify the Hill estimator, but they still rely on maximum likelihood, e.g. Fraga Alves (2001) has introduced a new lower bound. However, the influence function of Hill estimator is slowly increasing, but unbounded. Hill procedure is thus no robust and many authors tried to make the original Hill robust (see Beran and Schell (2012) or Vandewalle (2007)). In Fabián (2001) a new method of score moment estimators has been proposed. It appeared that these score moment estimators are robust for a heavy tailed distributions (see Stehlík et al. (2010)) We will compare t-estimation with other favorable heavy-tails estimation introduced in Brazauskas and Serfling (2001). For the case of Pareto distribution, the t-Hill estimator (see Fabián and Stehlík (2009)) procedure based on the score moment estimator has been investigated in Stehlík et al. (2013) for optimal testing for normality against Pareto tail. We will illustrate that t-Hill estimator is a "naturally" robust, distribution sensitive heavy tail estimator and prove its weak consistency together with its good small sample properties and some further structural properties (see Beran et al. (2014); Jordanova et al. (2013,a)).*

Theoretical aspects of the talks will be highlighted on two applications, namely modelling of methane flux (see Jordanova et al. (2013a)) and snow extremes modelling (see Stehlík et al. (2013)).

Keywords. *t- Hill estimator, small sample, robustness, methane flux.*

2.69

Nonparametric regression with one-sided irregular errors

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*
[*] *Corresponding author*

Abstract. *We consider nonparametric regression models with one-sided errors and regression functions in Hölder classes with different smoothness orders. Our aim is, on the one hand, the estimation of the regression function and, on the other hand, to develop hypotheses tests for the error distribution.*

We discuss suitable nonparametric regression estimators in models with one-sided errors and prove uniform rates of convergence. We see that the regression function can be estimated at a faster rate than in conventional nonparametric settings when the error distribution is irregular. Here 'irregular' means that sufficient mass is concentrated in the neighbourhood of zero.

Further we consider the empirical process of estimated residuals. It is shown that, under appropriate conditions, the asymptotic distribution is not influenced by the nonparametric estimation of the regression function. This is remarkably different from corresponding results in mean regression models with regular error distributions. The results are applied to derive goodness-of-fit tests for the error distribution.

Keywords. *Local maxima; Local polynomial approximation; Uniform rate of convergence; Residual empirical process; Goodness-of-fit testing*

2.70

Variable and shape selection for the generalized additive model

M. Meyer[1*]

[1] *Colorado State University; meyer@stat.colostate.edu*
[*] *Corresponding author*

Abstract. *The partial linear generalized additive model is considered, where the goal is to choose a subset of predictor variables and describe the component relationships with the response, in the case where there is very little* a priori *information. For each predictor, the user need only specify a set of possible shape or order restrictions. For example, the systematic component associated with a continuous predictor might be assumed to be increasing, decreasing, convex, or concave. The effect of a treatment variable might have a tree ordering or be unordered. A model selection method chooses the nature of the relationships as well as the variables. Given a set of predictors and shape or order restrictions, the maximum likelihood estimator for the constrained generalized additive model is found using iteratively re-weighted cone projections. The cone information criterion is used to select the best combination of variables and shapes. Because the shapes and orderings impose some degree of smoothness, no tuning parameters are required. The methods are introduced using two classical data sets, and a case study involving injuries in head-on collisions is presented. Simulations show that the model selection criterion performs well in comparison to the AIC with parametric assumptions. The* R *package* `cgam` *contains routines and data sets for the methods and examples.*

Keywords. *Isotonic; Convex; Information Criterion; Cone Projection*

2.71

A unified framework for spline estimators

Tatyana Krivobokova[1,*] and Katja Schwarz[1]

[1] *University of Göttingen; tkrivob@gwdg.de, k.sayevich@stud.uni-goettingen.de*
[*] *Corresponding author*

Abstract. *We discuss a unified framework to study the (asymptotic) properties of all (periodic) spline based estimators, that is of regression, penalized and smoothing splines. The explicit form of the periodic Demmler-Reinsch basis of general degree in terms of exponential splines allows to derive the exact expression for the equivalent kernel of all spline estimators simultaneously. The corresponding bandwidth, which drives the asymptotic behavior of spline estimators, is shown to be a function of both – the number of knots and the smoothing parameter. A strategy for the optimal bandwidth selection is discussed.*

Keywords. *B-splines; Demmler-Reinsch basis; Equivalent kernels*

2.72

Nonparametric estimation and forecasting of edge probabilities in a time variable random graph model

M. Birke[1,*]

[1] *University of Bayreuth; melanie.birke@uni-bayreuth.de*
[*] *Corresponding author*

Abstract. *In a random graph model where the position of the vertices is time variable we want to estimate or forecast the probability of an edge between two vertices at the present time given the distance of these vertices at an earlier time. To this end we use a nonparametric density estimator similar to that proposed in Bontemps et al. (2009). Consider the movement of the knots as independent identically distributed stochastic processes in some metric space. For a fixed time point the random graph model is based on the vertex set of independent identically distributed random vectors in that metric space. Furthermore, edges can appear or disappear when time moves forward, so the random egde sets vary over time. Motivated by a physical experiment with magneticed particles which are attracted if they come close to each other it makes sence to consider a random graph model where the position of the vertices at an earlier time influences the probability of the presence of an edge at a later time. The distribution of the distances of knots is continuous while the distribution of the random edges is Bernoulli conditionally on the distances. Following the proposal of Bontemps et al. (2009) for a conditional density estimator with both continuous and discrete distributions we get an estimator which looks like a randomly weighted U-statistic with the difference that the weights and the random knots are dependent. We use a central limit theorem for martingal differences and methods for U-statistics with a symmetric kernel depending on the sample size to establish the asymptotic behavior of the estimator. The practical performance of the estimator is demonstrated in a simulation study*

Keywords. *Edge Probability; Nonparametric Conditional Density Estimation; Random Graph*

References

Bontemps,C., Racine, J.S. and Simioni, M. (2009). Nonparametric vs parametric binary choice models: An empirical investigation. *Agricultural & Applied Economics Association Selected Paper* N° **611468**

2.73

Quantifying and reducing uncertainties on excursion sets under a gaussian random field prior

D. Ginsbourger[1,*], C. Chevalier[1], J. Bect[2], D. Azzimonti[1], and I. Molchanov[1]

[1] *Department of Mathematics and Statistics, Institute of Mathematical Statistics and Actuarial Sciences, University of Bern, Switzerland; ginsbourger@stat.unibe.ch, clément.chevalier@stat.unibe.ch, dario.azzimonti@stat.unibe.ch, Ilya@stat.unibe.ch*

[2] *Département Signaux & Systèmes Électroniques, Supélec, France; julien.bect@supelec.fr*

[*] *Corresponding author*

Abstract. *We focus on the problem of estimating the excursion set of a function above a given threshold under a limited evaluation budget. We adopt a Bayesian approach where the objective function is assumed to be a realization of some random field, typically assumed Gaussian, with mean and covariance functions known up to some parameters. In this setting, the posterior distribution on the objective function gives rise to a posterior distribution of excursion sets. In practice, estimates of the set of interest but also of the associated uncertainty can be derived from such posterior distribution. As notions of expectation and variability are not straightforward to* *define for random sets, several approaches have been proposed (Molchanov, 2005). In Chevalier et al. (2013), we considered the Vorob'ev expectation and deviation of an excursion set under a Gaussian random field prior, and found out that these quantities can be easily computed relying on the analytically tractable posterior mean and covariance functions of the field. However, further notions of expectation and variability appear intractable, which motivates Monte Carlo estimators relying on Gaussian field conditional simulations. In the present work we mainly deal with the optimal choice of simulation points, and we propose Monte Carlo estimates of non-linear functionals of the field (including various kinds of expectations and variability indices) obtained through enhanced approximations of field and set realizations.*

Keywords. *Set estimation, Conditional simulations, Vorob'ev deviation, Optimal design*

References

C. Chevalier, D. Ginsbourger, J. Bect, and I. Molchanov (2013). Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with gaussian process models. *mODa 10, Advances in Model-Oriented Design and Analysis, Contributions to Statistics*, pp 35–43.

I. Molchanov (2005). *Theory of Random Sets.* Springer, London, 2005.

2.74

On the comparison of the shape variability

Miguel López Díaz[1*]

[1] *Dpto. Estadística e I.O. y D.M. Universidad de Oviedo, Spain; mld@uniovi.es*
[*] *Corresponding author*

Abstract. *Shape analysis plays a very important role in many fields of research. A key matter in shape analysis is shape variability. For instance, the analysis of shape variability of anatomical structures is of great importance in many clinical disciplines. Usually, abnormality in shape is often related to disorders. In this communication we propose some criteria to compare shape variability of image-valued mappings. Such critera involve some concepts of classical geometry like the curvature of a special parameterization of bidimensional closed curves or the radial mapping of star-shaped sets. Applications of the methods are illustrated.*

Keywords. *Bidimensional closed curve; Radial function; Shape analysis; Stochastic order.*

2.75

Some geometric aspects of manifold estimation

A. Cuevas[1]

[1] *Departamento de Matemáticas, Universidad Autónoma de Madrid; antonio.cuevas@uam.es*

Abstract. *In the setting of this talk, the term "manifold estimation" will refer to those techniques aimed at reconstructing a manifold (or estimating some of its relevant features) from noisy observations taken around that manifold. We will summarize some recent results in this line, regarding the estimation of the surface area and the estimation of the "medial axis" of a set (which roughly represents the "central core" of the "median" of that set). Some relevant connections with statistical methodology and image analysis will be briefly commented.*
The contents of this talk are a summary of recent joint work with diferent co-authors: **J.R. Berrendero, A. Cholaquidis, R. Fraiman, P. Llop** *and* **B. Pateiro-López**.

Keywords. *Manifold estimation; Medial axis; Minkowski content.*

2.76

Comparing population clusterings

J.E. Chacón[1,*]

[1] *Departamento de Matemáticas, Universidad de Extremadura; jechacon@unex.es*
[*] *Corresponding author*

Abstract. *A population clustering can be understood as an essential partition of the support of a probability distribution. Different notions of cluster lead to different concepts of ideal population clusters, but no matter what approach is taken, eventually the researcher needs to evaluate the performance of a clustering methodology by measuring the distance between a data-driven clustering and the ideal population goal. In this talk, two new distances are proposed for this aim, by extending well-known distances between sets to distances between clusterings.*

Keywords. *Distance between partitions; Clustering consistency; Modal clustering; Population clustering.*

2.77

Estimation of varying coefficient models with randomly censored data

I. Van Keilegom[1,*], S. J. Yang[1], A. El Ghouch[1] and C. Heuchenne[1,2]

[1] *Université catholique de Louvain; ingrid.vankeilegom@uclouvain.be, seong.j.yang@gmail.com, anouar.elghouch@uclouvain.be*
[2] *Université de Liège; cedric.heuchenne@uclouvain.be*
[*] *Corresponding author*

Abstract. *The varying coefficient model is a useful alternative to the classical linear model, since the former model is much richer and more flexible than the latter. We propose estimators of the coefficient functions for the varying coefficient model in the case where different coefficient functions depend on different covariates and the response is subject to random right censoring. Since our model has an additive structure and requires multivariate smoothing we employ a smooth backfitting technique, that is known to be an effective way to avoid 'the curse of dimensionality' in structured nonparametric models. The estimators are based on synthetic data obtained by an unbiased transformation. The asymptotic normality of the estimators is established and a simulation study illustrates the reliability of our estimators.*

Keywords. *Smooth backfitting; Unbiased transformation; Random right censoring; Local polynomial smoothing, Curse of dimensionality.*

2.78

Characterizing the association between disease onset and survival under cross-sectional sampling

M. Carone[1,*], L. Pozzi[2], M.J. van der Laan[2], D.O. Scharfstein[3] and M. Asgharian[4]

[1] *Department of Biostatistics, University of Washington; mcarone@uw.edu*
[2] *Div. of Biostatistics, University of California, Berkeley; luca.pozzi@berkeley.edu, laan@berkeley.edu*
[3] *Department of Biostatistics, Johns Hopkins University; dscharf@jhsph.edu*
[4] *Department of Mathematics and Statistics, McGill University; masoud@math.mcgill.ca*

[*] *Corresponding author*

Abstract. *Common measures of association between random variables usually quantify departure from independence. Such measures often fail to provide a useful interpretation when the variables are successive durations with a constrained sum, such as age at disease onset and residual lifetime from onset, whose relationship we study in this work. To provide a more natural description of the association between disease onset and survival, we construct and study an accelerated residual lifetime model. We then consider estimation of parameters based on this* *model using survival data obtained through a cross-sectional survey with follow-up. This design is often employed to study the natural history of a disease, and yields systematically biased data with loss to follow-up. A first consistent estimator is presented, though because of its reliance on smoothing techniques it fails to have regular asymptotic behaviour. As an alternative, using the framework of targeted maximum likelihood estimation, we construct a modification of our estimator that not only retains consistency but also follows standard distributional limit theory. Confidence intervals can thus be readily constructed. Using data from the Canadian Study of Health and Aging, our methodology is used to infer about aspects of the natural history of dementia in the Canadian elderly population.*

Keywords. *Disease onset; Residual lifetime; Cross-sectional sampling; Targeted maximum likelihood estimation.*

2.79

Generalized copula-graphic estimation with left-truncated and right-censored data

Jacobo de Uña-Álvarez[1,*] and Noël Veraverbeke[2]

[1] *University of Vigo, Spain; jacobo@uvigo.es*
[2] *Hasselt University, Belgium, and North-West University, Potchefstroom, South Africa; noel.veraverbeke@uhasselt.be*
[*] *Corresponding author*

Abstract. *In this paper a copula-graphic estimator is proposed for left-truncated and right-censored survival data. It is assumed that there is some dependent censoring acting on the variable of interest, which may come from an existing competing risk. Furthermore, the full process is independently right-censored by some administrative censoring time, while there is an independent left-truncation variable which complicates the sampling procedure. The dependent censoring is modeled through an Archimedean copula function, which is supposed to be known. An asymptotic representation of the estimator as a sum of independent and identically distributed random variables is obtained and, consequently, a central limit theorem is established. These results extend to the truncated setting those in de Uña-Álvarez and Veraverbeke (2013). We investigate the finite sample performance of the estimator through simulations. A real data illustration is included.*

Keywords. *Competing risks; Cross-sectional sampling; Dependent censoring; Nonparametric estimation; Survival analysis.*

References

de Uña-Álvarez and Veraverbeke (2013). Generalized copula-graphic estimator. *Test* **22**, 343–360.

2.80

A covariate adjusted mann-whitney test for comparing two sojourn times under right censoring

Somnath. Datta[1,*]

[1] *somnath.datta@louisville.edu*
[*] *Corresponding author*

Abstract. *We develop a Mann-Whitney type test statistic based on the residuals from an accelerated failure time model fitted to two groups of sojourn times with a common set of covariates. This covariate adjusted test statistics handles right censoring via the inverse probability of censoring weights. This weights were devised to improve efficiency in the sense that certain pairs where at least one state entry time is uncensored could be compared. Extensive simulation studies were undertaken to evaluate the performance of this test. A real data illustration of our methodology is also provided.*

Keywords. *AFT Model; Waiting Times; Gap Times; Two Sample Test; Wilcoxon Test*

2.81

Robust testing of CAPM portfolio betas

Z. Prášková [1,*], M. Hušková[1], J. Steinebach[2] and O. Chochola[1]

[1] *Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 18675 Prague, Czech Republic; praskova@karlin.mff.cuni.cz, huskova@karlin.mff.cuni.cz, chochola@karlin.mff.cuni.cz*
[2] *University of Cologne, Mathematical Institute, Weyertal 86-90, D-50931 Cologne, Germany; jost@math.uni-koeln.de*
[*] *Corresponding author*

Abstract. *Capital assets pricing model (CAPM) represents a multivariate regression model with time varying parameters that measure risk of assets with respect to the market (portfolio betas). For testing stability (constancy) of portfolio betas we propose a sequential robust multivariate procedure that is based on M-estimators and partial weighted sums of M-residuals. Assuming weak dependency of both regressors and errors we study asymptotic properties of the test statistic under the null hypothesis of no change in the portfolio betas as well as under local alternatives. The asymptotic results are further extended to the CAPM model with high-frequency observations (functional CAPM). The theoretical results are accompanied by a simulation study that compares the proposed procedures with those based on the ordinary least-squares estimates. An application to a real data set is also presented.*

Keywords. *Robust monitoring; M-estimator; Portfolio beta; Change-point; Invariance principle*

2.82

Tests for structural changes in time series of counts

Š. Hudecová [1], M. Hušková[1,] and S. Meintanis[2]*

[1] *Charles University in Prague ; hudecova@karlin.mff.cuni.cz, huskova@karlin.mff.cuni.cz,*
[2] *National and Kapodistrian University of Athens; simosmei@econ.uoa.gr*
* *Corresponding author*

Abstract. *We propose methods for detecting structural changes in time series with discrete observations. The detector statistics come in familiar L2-type formulations incorporating the empirical probability generating function. Special emphasis is given to the popular models of integer autoregression (INAR) and Poisson autoregression (PAR). For both models we study structural changes due to a change in distribution, as well as the classical problem of parameter change. The asymptotic properties of the proposed test statistics are studied under the null hypothesis as well as under alternatives. A Monte Carlo power study on bootstrap versions of the new methods is also included along with real data examples.*

Keywords. *INAR model; Poisson autoregression; Change-point test; Empirical probability generating function; Binomial thinning.*

2.83

Dependent wild bootstrap for the empirical process

P. Doukhan[1], G. Lang[2], A. Leucht[3,*] and M. H. Neumann[4]

[1] *Université Cergy-Pontoise; doukhan@u-cergy.fr*
[2] *AgroParisTech ENGREF; gabriel.lang@engref.agroparistech.fr*
[3] *Universität Mannheim; leucht@uni-mannheim.de*
[4] *Friedrich-Schiller-Universität Jena; michael.neumann@uni-jena.de*
* *Corresponding author*

Abstract. *Many statistics can be rewritten as or approximated by functionals of the empirical process. When knowledge of the distribution of these statistics is required, e.g. for the construction of confidence sets or the determination of critical values for tests, knowledge of the distributional properties of the empirical process is useful. If the data generating process exhibits a (possibly unspecified) dependence structure, the distribution and also the asymptotics of the empirical process will depend on these (unknown) features of the underlying data generating process. It is well known that blockwise bootstrap methods provide a valid approximation. In this talk, we propose an alternative model-free bootstrap method for the empirical process under absolute regularity. More precisely, consistency of an adapted version of the so-called* *dependent wild bootstrap, that was introduced by Shao (2010) and is very easy to implement, is proved under minimal conditions on the tuning parameter of the procedure. We apply our results to construct confidence intervals for quantiles and to approximate critical values for statistical tests, e.g. the Kolmogorov-Smirnov test. A simulation study shows that our method is competitive to standard block bootstrap methods in finite samples.*

Keywords. *Bootstrap; Empirical process; Time series, Quantiles, Kolmogorov-Smirnov test.*

References

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association* **105**, 218–235.

2.84

Change-point tests for martingale difference hypothesis

Z.Hlávka [1], M. Hušková[1,*] C. Kirch[2] and S. Meintanis[3]

[1] *Charles University in Prague ; hlavka@karlin.mff.cuni.cz, huskova@karlin.mff.cuni.cz,*
[2] *Karlsruhe Institute of Technology; claudia.kirch@kit.edu*
[3] *National and Kapodistrian University of Athens; simosmei@econ.uoa.gr*
[*] *Corresponding author*

Abstract. *We will present testing procedures which detect if the observed time series is a martingale difference sequence. Furthermore, tests detection of change-points in the conditional expectation of the series given its past. The test statistics are formulated following the approach of Fourier-type conditional expectations first proposed by Bierens (1982) and have the advantage of computational simplicity. The limit behavior of the test statistics is investigated under the null hypothesis as well as under alternatives. A bootstrap procedure is proposed in order to get approximations for critical values. The performance of the bootstrap version of the test is compared in finite samples with other methods for the same problem. A real-data application is also included.*

Keywords. *Martingale difference hypothesis; Change-point test; Bootstrap test; Empirical characteristic function.*

2.85

Semi- and non-parametric methods for genomics big data

Jun Xie

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA; junxie@purdue.edu

Abstract. *Semi- and non-parametric methods have an advantage in genomics data analysis, as they can accommodate complicated data structures and do not assume specific model forms. In this talk, large scale genomics data will be examined, including examples from genome wide association studies and pharmacogenomics research. These examples are often referred to as big data problems, where tens or hundreds of thousands of genetic variables are measured. To tackle the challenge of modeling and analyzing the genomics big data, a semi-parametric method is first applied to scan the large number of variables for significance, then non-parametric variable selection and dimension reduction techniques are developed to further examine genetic information and how it may affect a clinical or disease outcome. Applications on predicting drug response based on patients' genetic and clinical data will be used to demonstrate the semi- and non-parametric methods. This research offers an example of utilizing semi- and non-parametric methods in the modern era of genomics big data.*

Keywords. *Big data; Dimension reduction; Genomics; Nonparametric*

2.86

Estimating false inclusion rates in penalized regression models

P. Breheny[1]

[1] *Department of Biostatistics, University of Iowa; patrick-breheny@uiowa.edu*

Abstract. *Penalized regression methods are an attractive tool for feature selection with many appealing properties, although their widespread adoption has been hampered by the difficulty of applying inferential tools. In particular, the question "How reliable is the selection of those features?" has proved difficult to address, partially due to the complexity of defining a false discovery in the penalized regression setting. Here we define a false inclusion as a variable that is independent of the outcome regardless of whether other variables are conditioned on. We show that this definition permits straightforward estimation of the number of false inclusions in near-independence conditions. We also develop a permutation-based approach and show that it yields more accurate estimation of the false inclusion rate in highly correlated settings. Extensions to generalized linear models and the group lasso are also discussed, including the problem of variable selection in high-dimensional additive modeling.*

Keywords. *Variable selection; False discovery rates; Penalized regression; Lasso; Group lasso.*

2.87

High dimensional ODEs coupled with mixed-effects modeling techniques for gene regulatory network identification

Hua Liang

Department of Statistics, George Washington University, 801 22nd St. NW, Washington, D.C. 20052, hliang@gwu.edu

Abstract. *Gene regulation network (GRN) is a complicated biological system. The understanding of the GRN system is essential to uncover the mystery of biological life. It is challenging to construct and identify GRN based on high-throughput time course microarray data. In this paper, we employ high-dimensional ordinary differential equation models (ODE) to describe the dynamic GRN. A complete procedure with five steps is proposed to construct the GRN and identify the significant gene-gene interactions from time course microarray data: 1) Use a mixed-effects nonparametric model with random effects following a mixture normal distribution to cluster genes into functional modules. 2) Apply a standard mixed-effects nonparametric s-moothing approach to each of the functional modules (clusters) to estimate the population mean curves and their derivatives for each functional module. 3) Construct an ODE system to connect all functional modules and formulate a regression model by plugging the estimates of the mean curves and their derivatives for each functional module from the previous step into the ODE models. Then apply the smoothly clipped absolute deviation (SCAD) technique to identify significant connections in the ODE system. 4) Apply the stochastic approximation EM (SAEM) technique to update the parameter estimates for the selected ODE system from the previous step. 5) Employ the Gene Ontology (GO) for function enrichment analysis to understand and interpret the established GRN. We apply this five-step procedure to time course microarray data from a yeast cell cycle study and construct the GRN for yeast cell cycle related modules.*

Keywords. *Differential equations; SAEM; time course microarray data; Two-stage method; Yeast cell cycles.*

2.88

A flexible semiparametric forecasting model for time series

Degui Li[1,*], Oliver Linton[2] and Zudi Lu[3]

[1] *Department of Mathematics, University of York, UK; degui.li@york.ac.uk.*
[2] *Faculty of Economics, University of Cambridge, UK; obl20@cam.ac.uk.*
[3] *School of Mathematical Sciences, University of Southampton, UK; Z.Lu@soton.ac.uk.*
[*] *The presenter*

Abstract. *We consider approximating a multivariate regression function by an affine combination of one-dimensional conditional component regression functions. The weight parameters involved in the approximation are estimated by least squares on the first-stage nonparametric kernel estimates. We establish asymptotic normality for the estimated weights and the regression function in two cases: the number of the covariates is finite, and the number of the covariates is diverging. As the observations are assumed to be stationary and near epoch dependent, the approach in this paper is applicable to both the estimation and forecasting issues in time series analysis. Furthermore, the methods and results are augmented by a simulation study and illustrated by applications in the analysis of the Australian annual mean temperature anomaly series and forecasting the high frequency volatility of the FTSE100 index.*

Keywords. *Forecasting, Marginal regression, Model averaging, Nadaraya-Watson Kernel estimation, Near epoch dependence, Semiparametric estimation.*

2.89

Semiparametric GEE analysis in parially linear single-index models for longitudinal data

Jia Chen[1,*], Degui Li[2], Hua Liang[3] and Suojin Wang[4]

[1] *Department of Economics and Related Studies, University of York, York, UK; jia.chen@york.ac.uk*
[2] *Department of Mathematics, University of York, York, UK; degui.li@york.ac.uk*
[3] *Department of Statistics, George Washington University, Washington, D.C., US; hliang@gwu.edu*
[3] *Department of Statistics, Texas A&M University, College Station, TX, US; sjwang@stat.tamu.edu*
[*] *Presenter*

Abstract. *In this article, we study a partially linear single-index model for longitudinal data under a general framework which includes both the sparse and dense longitudinal data cases. A semiparametric estimation method based on the combination of the local linear smoothing and generalized estimation equations (GEE) is introduced to estimate the two parameter vectors as well as the unknown link function. Under some mild conditions, we derive the asymptotic properties of the proposed parametric and nonparametric estimators in different scenarios, from which we find that the convergence rates and asymptotic variances of the proposed estimators for sparse longitudinal data would be substantially different from those for dense longitudinal data. We also discuss the estimation of the covariance (or weight) matrices involved in the semiparametric GEE method. Furthermore, we provide some numerical studies to illustrate our methodology and theory.*

Keywords. *GEE; local linear smoothing; longitudinal data; semiparametric estimation; single-index models.*

2.90

Separation of amplitude and phase variation in point processes

Victor M. Panaretos[1,*]

[1] *Department of Mathematics, EPFL; victor.panaretos@epfl.ch*
[*] *Corresponding author*

Abstract. *The amplitude variation of a real random field $\{X(t)\}$ consists in its random oscillations in the y-axis, typically encapsulated by its (co)variation around a mean level. In contrast, phase variation refers to fluctuations in the x-axis, often caused by random time changes or spatial deformations. We consider the problem of identifiably formalising similar notions for (potentially spatial) point processes, and of nonparametrically separating them based on realisations of iid copies of the phase-varying point process. The key element of our approach is the use of the theory of optimal transportation of measure, which is proven to allow the consistent separation of the two types of variation for point processes over Euclidean domains. (Based on joint work with Y. Zemel, EPFL).*

Keywords. *Fréchet mean; Monge-Kantorovich Problem; Warping.*

References

[1] Panaretos, V. M., & Zemel, Y. (2014). Separation of Amplitude and Phase Variation in Point Processes. *Technical Report, Chair of Mathematical Statistics, EPFL.*

2.91

A deformation model for empirical distributions with Wasserstein's distance

Jean-Michel Loubes[1]

[1] *Institut de Mathématiques de Toulouse, Université de Toulouse 3, loubes@math.univ-toulouse.fr*

Abstract. *We tackle the problem of comparing distributions of random variables and defining a mean pattern between a sample of random events. Using barycenters of measures in the Wasserstein space, we propose an iterative version as an estimation of the mean distribution. Moreover, when the distributions are a common measure warped by a centered random operator, then the barycenter enables to recover this distribution template. We provide also a goodness of fit procedure to check the deformation model.*

Keywords. *Deformation model, Wasserstein's distance*

2.92

Convergence of the Wasserstein distance of empirical Gaussian distributions

A. Munk[1], T. Rippl[2,*] and A. Sturm[1]

[1] *University of Göttingen; munk@math.uni-goettingen, asturm@math.uni-goettingen.de*
[2] *University of Erlangen; rippl@math.fau.de*
[*] *Corresponding author*

Abstract. *We deduce a central limit theorem for the l^2-Wasserstein distance of two empirical distributions when the true distributions are Gaussian. The cases are distinguished whether these two distributions are the same or different. In the case of different distributions we give an explicit formula for the asymptotic variance. We also show that the Wasserstein distance in the Gaussian case is Fréchet differentiable.*

Keywords. *Wasserstein distance; Gaussian distribution; Empirical distribution; Central limit theorem; Fréchet differentiability.*

2.93

Confidence sets based on thresholding estimators in high-dimensional gaussian regression

Ulrike Schneider[1]

[1] *Vienna University of Technology; ulrike.schneider@tuwien.ac.at*

Abstract. *We study confidence intervals based on hard-thresholding, soft-thresholding, and adaptive soft-thresholding in a regression model where the number of regressors k may depend on and diverge with sample size n. In addition to the case of known error variance, we define and study versions of the estimators when the error variance is unknown. In the known variance case, we provide an exact analysis of the coverage properties of such intervals in finite samples. We show that these intervals are always larger than the standard interval based on the least-squares estimator. Asymptotically, the intervals based on the thresholding estimators are larger even by an order of magnitude when the estimators are tuned to perform consistent variable selection. For the unknown-variance case, we provide non-trivial lower bounds for the coverage probabilities in finite samples and conduct an asymptotic analysis where the results from the known-variance case can be shown to carry over asymptotically if the number of degrees of freedom $n - k$ tends to infinity fast enough in relation to the thresholding parameter.*

Keywords. *Thresholding; Lasso; High-dimensional regression model; Confidence set*

References

Schneider, U. (2013). Confidence Sets Based on Thresholding Estimators in High-Dimensional Gaussian Regression. **arxiv:1308.3201**.

3.01

OODA, visualization, backwards PCA & HDLSS asymptotics

J. S. Marron[1]

[1] *Department of Statistics and Operations Research. University of North Carolina*

Abstract. *Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. In non-Euclidean analysis, the approach of Backwards PCA is seen to be quite useful. An overview of insightful mathematical statistics for object data is given, based on High Dimension Low Sample Size asymptotics, where the dimension grows, but the sample size is fixed.*

3.02

Nonstationary and nonparametric modelling of multivariate correlated data via kernel processes mixing

Montse Fuentes[1]

[1] *North Carolina State University. Statistics Department; fuentes@ncsu.edu*

Abstract. *We introduce a nonparametric multivariate spatial model that avoids specifying a Gaussian distribution for spatial random effects. Our nonparametric model extends the stick-breaking (SB), which is frequently used in Bayesian modelling to capture uncertainty in the parametric form of an outcome. The stick-breaking prior is extended here to the spatial setting by assigning each location a different, unknown distribution, and smoothing the distributions in space with a series of space-dependent kernel functions that have a space-varying bandwidth parameter. This results in a flexible non stationary model for correlated data, as different kernel functions lead to different relationships between the distributions at nearby locations. This approach is the first to allow both the probabilities and the point mass values of the SB prior to depend on space. Thus, there is no need for replications and we obtain a continuous process in the limit. We extend the model to the multivariate setting by having for each process a different kernel function, but sharing the location of the kernel knots across the different processes. The resulting covariance for the multivariate process is in general nonstationary and nonseparable. The modelling framework proposed here is also computationally efficient because it avoids inverting large matrices and calculating determinants, which often hinders the spatial analysis of large data sets. We study the theoretical properties of the proposed multivariate spatial process. The methods are illustrated using simulated examples and an air pollution application to model components of fine particulate matter.*

3.03

A semiparametric framework for rank tests

Jan De Neve[1,*], Olivier Thas[1,2] and Jean–Pierre Ottoy[1]

[1] *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium; JanR.DeNeve@UGent.be, JeanPierre.Ottoy@UGent.be*
[2] *Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Australia; Olivier.Thas@UGent.be*
[*] *Corresponding author*

Abstract. *We demonstrate how classical rank tests, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis, and Friedman tests can be embedded in a statistical modelling methodology and how our approach can be used for constructing new rank tests for more complicated designs.*

In particular, rank tests for unbalanced and multi-factor designs, and rank tests that allow for correcting for continuous covariates are included. The method also allows for the estimation of meaningful effect sizes. Our method results from two particular parametrizations of probabilistic index models (Thas et al., 2012).

Keywords. *Factorial design; Kruskal–Wallis rank test; Wilcoxon–Mann–Whitney test; Probabilistic index model*

References

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.

3.04

Diagnostic tests for the location-shift assumption

O. Thas*[1,2,*] and J.C.W. Rayner[2,3]

[1] *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium; Olivier.Thas@UGent.be*
[2] *National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia*
[3] *School of Mathematical and Physical Sciences, University of Newcastle, Australia; John.Rayner@newcastle.edu.au*
[*] *Corresponding author*

Abstract. *The Wilcoxon rank-sum test is often considered as the nonparametric version of the t-test for comparing means. However, the hypotheses of the Wilcoxon test can only be expressed in terms of means under restrictive distributional assumptions. The most common assumption demands that the distributions of the two populations belong to a location-shift family, i.e. the distributions agree in shape except for a location shift. We present tests for testing the location-shift assumption. A first test is based on the Wasserstein distance between the two empirical quantile functions. A second class of tests is based on the set of semiparametric efficient score statistics. In an empirical power study we compare the tests. Finally, we present some meaningful graphical diagnostic tools for assessing the location-shift assumption.*

Keywords. *Rank tests, goodness-of-fit*

3.05

On robust mutivariate Kaplan-Meier estimator

Purba Mondal[1], Anannya Nath[1] and Subhra Sankar Dhar[2]

[1] *Indian Statistical Institute, Chennai, India; purba@isichennai.res.in, anannya@isichennai.res.in*
[2] *Indian Institute of Technology, Kanpur, India; subhra@iitk.ac.in*

Abstract. *For censored data, Kaplan Meier Estimator has widely been used to estimate survival function. In this article, we extend the univariate Kaplan Meier Estimator based on the product integral representation into multivariate setup. Some properties and the performance of the estimator have been extensively investigated on different simulated and real data related to medical sciences. In addition, we also propose a modified version of the estimator, which is robust against the outliers.*

Keywords. *Kaplan Meier Estimator; Multivariate Censored Data; Product Integral Representation; Robustness*

References

Dabrowska, D. M. (1988) Kaplan-Meier Estimate on the Plane. *The Annals of Statistics*, **16**, 1475–1489.

3.06

Choosing smoothing parameters in a kernel estimation of a multivariate regression function

I. Horová[1,*] and J. Koláček[1]

[1] *Masaryk University, Brno, Czech Republic; horova@math.muni.cz, kolacek@math.muni.cz*
[*] *Corresponding author*

Abstract. *Kernel estimation of a regression function is a useful technique in exploratory data analysis. The crucial factor which influences the quality of the kernel estimation is a smoothing parameter. The present paper is devoted to the extension of the univariate kernel regression estimation to the multivariate setting. However, this extension is not without its problems. The kernel estimation of the d-variate regression function requires the specification of entries of a $d \times d$ positive definite matrix of smoothing parameters. We aim to develop a method for choosing smoothing parameters which is based on a property of the asymptotic integrated square error of this estimation and its suitable approximation. The analysis of statistical properties of the proposed method confirms its rationality. A simulation study compares the least squares* *cross-validation method and the proposed method.*

Keywords. *Regression Function; Multivariate Kernel; Smoothing Parameters Selection.*

3.07

Kernel selection in nonparametric density estimation

M.I. Borrajo[1,*], J.E. Chacón[2] and A. Rodrígucz-Casal[1]

[1] *Departamento de Estadística e Investigación Operativa, Universidade de Santiago de Compostela, Spain; mariaisabel.borrajo@usc.es, alberto.rodriguez.casal@usc.es*
[2] *Departamento de Matemáticas, Universidad de Extremadura, Spain; jechacon@unex.es*
[*] *Corresponding author*

Abstract. *Density estimation has been a widely studied field in nonparametric statistics. Since the introduction of the histogram to the kernel density estimator, including bandwidth selection methods, many advances has been made. We propose a new density estimator based on the theory developped in Watson and Leadbetter (1963), which only assumes the kernel to be a L^2 function. In this way we let the kernel vary in shape and scale, and at the same time, we avoid the problem of bandwidth selection.*

Our purpose is based on the Fourier transform of the optimal kernel, which can be empirically estimated since it depends on the characteristic function of the density that we aim to estimate. In order to remove the noise of the empirical characteristic function for large frequencies, we propose a new data-based method for selecting the truncation point in the frequency domain.

Finally, the behaviour of our method is illustrated in a simulation study, in which we have included the self-consistent density estimator of Bernacchia and Pigolotti (2011), as well as the main bandwidth selectors for the kernel density estimator.

Keywords. *Density estimation; Kernel selection.*

References

Bernacchia, A. and Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society: Series B* **73(3)**, 407–422.

Watson, G. S. and Leadbetter, M. R. (1963). On the estimation of the probability density. *Annals of Mathematical Statistics* **34**, 480–491.

3.08

Nonparametric kernel density estimation for grouped data

M. Reyes[1,*], M. Francisco-Fernández[1] and R. Cao[1]

[1] *Universidad de A Coruña, Facultad de Informática, Campus de Elviña, s/n, 15071 A Coruña, Spain; m.reyes@udc.es, mariofr@udc.es, rcao@udc.es*
[*] *Corresponding author*

Abstract. *Grouped data appear when the event of interest cannot be directly observed and it is only known to have occurred within an interval. In this framework, a nonparametric kernel density estimator, based on the Parzen-Rosenblatt approach (Parzen, 1962; Rosenblatt, 1956), is proposed and studied. Following up on the papers of Hall (1982) and Scott and Sheather (1985), we analyze the effect of using non equally-spaced grouped data in our estimator. Under usual assumptions about the density function, bandwidth and kernel function, as well as additional assumptions on the interval partition, asymptotic bias and variance are derived and a plug-in bandwidth selector is proposed. Through a comprehensive simulation study we show the behavior of both the estimator and its bandwidth selector under different scenarios of data grouping. An application to real data confirms the simulation results, showing the good performance of the estimator whenever data are not heavily grouped.*

Keywords. *Kernel density estimation; Non equally-spaced grouped data; Plug-in bandwidth selector.*

References

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM Journal on Applied Mathematics* **42**, 390–399.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, 832–837.

Scott, D.W. and Sheather, S.J. (1985). Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods* **14**, 1353–1359.

3.09

Multiple adaptive tests and the false discovery rate control (FDR) under dependent p-values

A. Janssen[1,*] and P. Heesen[1]

[1] *Heinrich-Heine University, Düsseldorf, Germany; janssena@math.uni-duesseldorf.de; heesen@math.uni-duesseldorf.de*
[*] *Corresponding author*

Abstract. *High dimensional testing problems are frequently attacked by multiple tests. A useful quantity for the judgement of multiple tests is the false discovery rate (FDR), the expected ratio of the number of false rejected null and the total number of rejections. We refer to Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Storey* et al. *(2004), and Finner* et al. *(2009) among others. In this talk we introduce new adaptive multiple tests with data dependent critical values. In particular, the given p-values may be dependent. The core of the work are inequalities for the FDR. These results may be applied to dependent genome data.*

Keywords. *PRDS; Positive dependence; Adaptive Benjamini Hochberg methods, Finite sample FDR control.*

References

Benjami, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach in multiple testing. *J. Roy. Stat. Soc. B* **57**, 289-300.

Benjami, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **10**, 1165-1188.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Stat. Soc. B* **66**, 187-205.

Finner, H., Dickhaus, T. and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Stat.* **37**, 596-618.

3.10

Family-wise separation rates for multiple testing

M. Fromont[1], M. Lerasle[2], and P. Reynaud-Bouret[2]

[1] *Univ. Européenne de Bretagne, IRMAR, CNRS UMR 6625, Rennes, France; magalie.fromont@univ-rennes2.fr*
[2] *Univ. Nice Sophia Antipolis, LJAD, CNRS UMR 7351, Nice, France; mlerasle@unice.fr; Patricia.Reynaud-Bouret@unice.fr*

Abstract. *Noting that aggregation-based adaptive tests of a single null hypothesis can be viewed as tests of multiple hypotheses, we here investigate the parallel that can be drawn between these tests. From this parallel, we aim at proposing a new second kind error-related tool to evaluate multiple testing procedures. Aggregation-based adaptive tests were proposed for instance by Baraud (2002), Baraud, Huet, Laurent (2003), or Fromont, Laurent (2006). They were evaluated on the one hand through their first kind error rate, which achieves the exact desired level, and on the other hand through their separation rates over various classes of alternatives, which were proved to be of the same order as the corresponding adaptive minimax separation rates. We show that these tests amount to particular step-down procedures, so they can also be evaluated from the multiple testing point of view, through a control of the Family-Wise Error Rate. Conversely, several multiple tests, such as Holm (1979)'s or Romano, Wolf (2005)'s procedures, may be studied from the aggregation-based testing point of view. More precisely, we propose to extend the notion of separation rate to multiple tests. We thus define the* weak Family-Wise Separation Rate *and its stronger counterpart, the* Family-Wise Separation Rate *(FWSR). We derive general properties, which in particular allow us to control the FWSRs of some multiple testing procedures in classical cases, and to prove that they are optimal in a minimax sense.*

Keywords. *Multiple testing; Family-Wise Error Rate; Step-down procedure; Adaptive test; Minimax separation rate*

References

Baraud, Y. (2002). Non asymptotic minimax rates of testing in signal detection. *Bernoulli* **8**, 5, 561–696.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31**, 1, 225–251.

Fromont, M., and Laurent, B. (2006). Adaptive goodness-of-fit tests in a density model. *Ann. Statist.* **34**, 2, 680–720.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Romano, J. P., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100**, 469, 94–108.

3.11

Data integration in early drug development with nonparametric FDR control for feature extraction

F. Mattiello[1,*], O. Thas[1,2] and B. Verbist[1]

[1] *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Belgium; Federico.Mattiello@UGent.be, Olivier.Thas@UGent.be, Bie.Verbist@UGent.be*

[2] *National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia*

[*] *Corresponding author*

Abstract. *In this work we propose a method for data integration of three data sources with applications in early drug development. The method was developed for QSTAR (Quantitative Structure Transcriptional Activity Relationship) analysis, which extents classical QSAR methods by including transcriptomics data (gene expressions). This new methodology was proposed by the QSTAR consortium (www.qstar-consortium.org).*

The objective is to identify genes and chemical substructures of molecules that are associated with a bioassay outcome, which is assumed to be related to the clinical outcome of the drug (molecule). In particular, we look for substructures that affect the bioassay through gene regulation. The core of the method is a sparse singular value decomposition (Witten and Tibshirani, 2009a and 2009b), which assures the identification of only the most important genes.

We have developed a nonparametric method for controlling the false discovery rate (FDR) for the detection of genes and chemical substructures within this data integration method. The method inherits several properties of principal component analysis and the nonparametric SAM method (Tusher et al., 2001).

The method is empirically evaluated in a simulation study and it is illustrated on a real dataset.

Keywords. *False discovery rate (FDR), Genomics, Multivariate method, Sparse singular value decomposition*

References

Daniela M. Witten and Robert J. Tibshirani (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Daniela M. Witten and Robert J. Tibshirani (2009). Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology* **8**(1), Article 28.

Virginia Goss Tusher, Robert J. Tibshirani and Gilbert Chu (2001). Significance Analysis of Microarrays applied to the ionizing radiation response. *PNAS* **98**(9), pp. 5116–5121.

3.12

Robust multivariate regression using data depth with applications to high-dimensional sparse data

Subhajit Dutta[1,*] and Marc G. Genton[1]

[1] *CEMSE, King Abdullah University of Science and Technology, Thuwal 23955-6900, KSA; subhajit.dutta@kaust.edu.sa, marc.genton@kaust.edu.sa*
[*] *Corresponding author*

Abstract. *A robust method for multivariate regression is developed based on robust estimators of the joint location and scatter matrix of the explanatory and response variables using the notion of data depth. We show that the multivariate regression estimator possesses desirable affine equivariance properties, achieves the best breakdown point of any affine equivariant estimator, and has a bounded influence function. To increase the efficiency of this estimator, we further construct a re-weighted estimator based on robust clustering of the residual vectors. When the* *data dimension is quite high compared to the sample size, we can still use some notions of data depth meaningfully, and use the corresponding depth values to construct a robust and sparse estimator. The resulting multivariate regression technique is still computationally quite feasible, and turns out to perform better than several existing and popular methods of robust multivariate regression when applied to various simulated as well as real benchmark datasets.*

Keywords. *Depth-weighted estimators; LASSO; Outliers; Projection depth; Spatial depth*

3.13

Consistency of infimal depth for functional data

S. Nagy

KU Leuven, Dept. of Mathematics, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium; stanislav.nagy@wis.kuleuven.be
Charles University in Prague, Dept. of Probability and Mathematical Statistics, Czech Republic

Abstract. *Recently, Mosler and Polyakova (2012) proposed a simple depth of infimal type for functional data. Kuelbs and Zinn (2013) immediately raised concerns about the consistency of the sample version of this depth, and gave examples where the consistency fails to hold. In the contribution we find sufficient conditions for the uniform consistency of this depth, and show that these cannot be weakened substantially. Connections to other depth functionals used in practice are outlined.*

Keywords. *Data Depth; Functional Data, Consistency.*

References

Mosler, K. and Polyakova, Y. (2012). General notions of depth for functional data. *arXiv preprint arXiv:1208.1981*

Kuelbs, J. and Zinn, J. (2013). Concerns with Functional Depth. *ALEA. Latin American Journal of Probability and Mathematical Statistics*, **10**(2), 831–855.

3.14

A comparative study of statistical functional depths

Alicia Nieto-Reyes[1]

[1] *Departamento de Matemáticas, Estadística y Computación. Universidad de Cantabria; alicia.nieto@unican.es*

Abstract. *This talk presents a comparative study among the existing functional depths, taking into account that different definitions of depth have different aims. This comparison pays special attention to the properties satisfied by those depths, theoretical properties as well as empirical. Functional depths have appeared as an extension to the continuum of multidimensional depth functions. As a consequence, the aim has been to satisfy the established properties of multidimensional depth, which do not take into account the special nature of functional data. Here, it will be described until what extend the existing functional depths do take into account the functional characteristic of the data.*

Keywords. *Depth properties; Functional data analysis; Functional depth; Multivariate data analysis; Statistical data depth.*

3.15

Expectile trimming

I. Cascos[1]

[1] *Department of Statistics, Universidad Carlos III de Madrid; ignacio.cascos@uc3m.es*

Abstract. *Expectiles were defined by Newey and Powell (1987) as the solution to a minimization problem arising in the context of linear regression. They were so named because they resemble some similarities with the quantiles of a random variable. Inspired by Tukey (1975) who used the quantiles of the univariate projections of a bivariate data cloud to produce a family of central (depth-trimmed) regions, Eliers (2010) suggested to use the expectiles to build the expectile regions of a multivariate data set. Tukey's central regions inspired the definition of the most popular depth function, the so-called halfspace depth, while the expectile regions can be used to introduce the expectile depth.*
We show how to compute efficiently the expectile contours of a bivariate data cloud and also the expectile depth of a given point with respect to a bivariate sample.

Keywords. *Depth function; Depth-trimmed region; Expectile.*

References

Eilers, P. H. C. (2010). Expectile contours and data depth. *Proceedings of the 25th International Workshop on Statistical Modelling.* A.W. Bowman (ed.), 167–172.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrika* **55**, 819–847.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematitians*, vol. 2, 523–531.

3.16

Kernel density estimation with berkson error

J. Long[1,*], N. El Karoui[2] and J. Rice[2]

[1] *Department of Statistics, Texas A&M University; jlong@stat.tamu.edu*
[2] *Department of Statistics, University of California, Berkeley; nkaroui@stat.berkeley.edu, rice@stat.berkeley.edu*
[*] *Corrupted author*

Abstract. *Given a sample $\{X_i\}_{i=1}^n$ from f_X, we construct kernel density estimators for f_Y, the convolution of f_X with a known error density f_ϵ. This problem is known as density estimation with Berkson error and has applications in epidemiology and astronomy. We derive an asymptotic approximation to the mean integrated squared error (MISE) of a kernel density estimator of f_Y as a function of the bandwidth parameter. Using these asymptotic rates we determine the optimal rate of convergence for the bandwidth. We compare MISE rates for several bandwidth selection strategies both asymptotically and at finite samples. The finite sample results demonstrate the importance of smoothing when the error term f_ϵ is concentrated near 0. We apply our methodology to NO_2 exposure data.*

Keywords. *Berkson error; Measurement error; Bandwidth selection; Kernel density estimation; Multivariate density estimation.*

3.17

Robust nonnegative garrote variable selection in linear regression

I. Gijbels[1] and I. Vrinssen[1,*]

[1] *KU Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat); irene.gijbels@wis.kuleuven.be, inge.vrinssen@wis.kuleuven.be*
[*] *Corresponding author*

Abstract. *Variable selection in a linear regression model is an important topic in applied data analysis, since often many variables are measured. Many variable selection methods are available in the literature. One of them is the nonnegative garrote method, originally developed for linear regression, but successfully extended to additive regression models and varying coefficient models, among others.*

In this talk we aim at robustifying the nonnegative garrote method to make it less sensitive to vertical outliers and leverage points. Our method is based on the S-estimator and an extra improvement step is proposed to increase the efficiency. Theoretical properties of the estimator, such as the convergence in estimation and variable selection and breakdown point, are estab-

lished. The performance of the method is illustrated by a simulation study and by a real data example.

Keywords. *Multiple linear regression; Nonnegative garrote; S-estimation; Variable selection.*

3.18

Nonparametric lack-of-fit testing and consistent variable selection

Adriano Z. Zambom[1,*], Michael G. Akritas[2]

[1] *UNICAMP - State University of Campinas- Brazil; zambom@ime.unicamp.br,*
[2] *Pennsylvania State University; akritas@stat.psu.edu*
[*] *Corresponding author*

Abstract. *Let* $\mathbf{X}$ *be a* d *dimensional vector of covariates and* Y *be the response variable. Under the nonparametric model* $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ *we develop an ANOVA-type test for the null hypothesis that a particular coordinate of* $\mathbf{X}$ *has no influence on the regression function. The asymptotic distribution of the test statistic, using residuals based on local polynomial regression, is established under the null hypothesis and local alternatives. Simulations suggest that the test outperforms existing procedures in heteroscedastic settings. Using p-values from this test, a variable selection method based on False Discovery Rate corrections is proposed, and proved to be consistent in estimating the set of indices corresponding to the significant covariates. Simulations suggest that, under a sparse model, dimension reduction techniques can help avoid the curse of dimensionality. A backward elimination version of this procedure, called BEAMS (Backward Elimination ANOVA-type Model Selection), performs competitively against well established procedures in linear regression settings, and outperforms them in nonparametric settings. A real data set is analyzed.*

Keywords. *Backward elimination; dimension reduction; model checking; multiple testing; nonparametric regression.*

3.19

On consistency of the least squares estimators in linear errors-in-variables models with infinite variance errors

Yu.V. Martsynyuk[1]

[1] *Department of Statistics, University of Manitoba, 338 Machray Hall, Winnipeg, MB R3T 2N2, Canada; Yuliya.Martsynyuk@UManitoba.CA*

Abstract. *We deal simultaneously with linear structural and functional errors-in-variables models (SEIVM and FEIVM), revisiting in this context the ordinary least squares estimators*

(LSE) for the slope and intercept of the corresponding simple linear regression. It has been known that, subject to some model conditions, these estimators become weakly and strongly consistent in the linear SEIVM and FEIVM with the measurement errors having finite variances when the explanatory variables have an infinite variance in the SEIVM, and a similar infinite spread in the FEIVM, while otherwise, the LSE's require an adjustment for consistency with the so-called reliability ratio. We prove weak and strong consistency, with and without the possible rates of convergence being determined, for the LSE's of the slope and intecept, assuming that the measurement errors are in the domain of attraction of the normal law (DAN) and thus are, for the first time, allowed to have infinite variances. Moreover, these results are obtained under the conditions that the explanatory variables are in DAN, have an infinite variance, and dominate the measurement errors in terms of variation in the SEIVM, and under appropriately matching versions of these conditions in the FEIVM. This duality extends a previously known interplay between SEIVM's and FEIVM's.

Keywords. *Linear structural and functional errors-in-variables models; Least squares estimators; Reliability ratio; Domain of attraction of the normal law; Weak and strong consistency.*

3.20

Semiparametric M estimation with missing covariates

F. Bravo

[1] *University of York, York, YO10 5DD, United Kingdom; francesco.bravo@york.ac.uk*

Abstract. *This paper considers M estimation in semiparametric models when some of the covariates are missing at random. The paper proposes an iterative M estimator based on inverse probability weighting and local linear estimation of the nonparametric component. The resulting estimator is very general and can be used in the context of semiparametric maximum likelihood, quasi likelihood and robust estimation. The paper establishes the asymptotic normality of the M estimator using both nonparametric and parametric estimation of the unknown probability weights. Monte Carlo simulations show that the proposed estimator has good finite sample properties.*

Keywords. *Inverse probability weighting; Local linear approximation; Missing at random*

3.21 Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units

Sahar Z Zangeneh[1,*], Robert W Keener [2] and Roderick J.A. Little[3]

[1] *Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center; saharzz@fhcrc.org*
[2] *Department of Statistics, University of Michigan; keener@umich.edu*
[3] *Department of Biotatistics, University of Michigan; rlittle@umich.edu*
[*] *Corresponding author*

Abstract. *In Probability proportional to size (PPS) sampling, the sizes for nonsampled units are not required for the usual Horvitz-Thompson or Hajek estimates, and this information is rarely included in public use data files. Previous studies have shown that incorporating information on the sizes of the nonsampled units through semiparamteric models can result in improved estimates. When the design variables that govern the selection mechanism, are missing, the sample design becomes informative and predictions need to be adjusted for the effect of selection. We present a general framework using Bayesian nonparametric mixture modeling with Dirichlet process priors for imputing the nonsampled size variables, when such information is not available to the statistician analyzing the data. The method is applied on a data sets from the Swedish municipalities.*

Keywords. *Bayesian nonparametric modeling; Mixture models; Dirichlet process priors; Probability proportional to size sampling.*

3.22 Particle learning for Bayesian non-parametric stochastic volatility models

A. Virbickaite[1], H.F. Lopes[2], P. Galeano[1] and M.C. Ausin[1,*]

[1] *Department of Statistics, Universidad Carlos III de Madrid; audrone.virbickaite@uc3m.es, pedro.galeano@uc3m.es, concepcion.ausin@uc3m.es*
[2] *Booth School of Business; hlopes@chicagobooth.edu*
[*] *Corresponding author*

Abstract. *A Bayesian non parametric Stochastic Volatility model with Dirichlet Process Mixture errors is considered for financial time series. Inference and prediction are carried out using a class of Sequential Monte Carlo Methods called Particle Learning algorithms which, compared to other alternative particle filtering methods, are known to be more efficient. The performance of the proposed particle method is compared with the standard Bayesian estimation approach for Stochastic Volatility models based on Markov Chain Monte Carlo methods. It is shown that the Particle Learning approach performs similarly to the conventional Bayesian estimation meth-*

ods being, at the same time, much faster and allowing for on-line type inference. The proposed method is illustrated using simulated and real data.

Keywords. *Dirichlet Process Mixture; MCMC; Particle Learning; Stochastic Volatility; Sequential Monte Carlo.*

References

Ausín, M.C., Galeano, P., Ghosh, P., (2014). A semiparametric Bayesian approach to the analysis of financial time series with applications to value at risk estimation. *European Journal of Operational Research*, 232 , 350-358.

Carvalho, C.M., Johannes, M.S., Lopes, H.F., Polson, N.G. (2010). Particle Learning and Smoothing. *Statistical Science*, 25 , 88-106.

Carvalho, C.M., Lopes, H.F., Polson, N.G., Taddy, M.A. (2010). Particle Learning for General Mixtures. *Bayesian Analysis*, 5 , 709-740.

3.23

Leplace-Monge-Kantorovith depth

Marc Hallin[1]

[1] *ECARES, Université libre de Bruxelles and ORFE, Princeton University*

Abstract. *Existing concepts of statistical depth at best are affine-invariant, the corresponding depth contours being affine-equivariant. Affine-invariance, and monotonicity along straight lines running through a deepest point are part of the four basic properties required from any depth concept in the axiomatic approach of Zuo and Serfling (2000) (see also Liu 1990). Those two properties have a somewhat regrettable linear flavor, though. And, in dimension one, the invariance/equivariance properties of depth hold under the much larger group of continuous transformations, that is, homeomorphisms—a group that, contrary to the affine group, generates the whole family of absolutely continuous distributions. Based on the concept of optimal measure transportation, we propose a new definition of depth, the Monge-Kantorovitch depth, that similarly enjoys invariance/equivariance properties under the group of homeomorphisms of the K-dimensional real space. The price to be paid is a weakening of the monotonicity-along-straight-lines property into a weaker property of nested depth contours.*
Based on joint work with A. Galichon (Sciences-Po, Paris) and M. Henry (University of Pennsylvania).

3.24

Depth functions in multivariate and other data settings: concepts, perspectives, tools, applications

R. Serfling[1]

[1] *Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080 USA; serfling@utdallas.edu*

Abstract. *Depth functions were developed, notably beginning with Tukey (1975) and Liu (1988), to extend the univariate notions of median and order statistics to the setting of multivariate data. Whereas a density function measures local probability weight, a depth function measures centrality. Its contours induce closely associated multivariate outlyingness, quantile, sign, and rank functions. Together, these functions comprise a powerful methodology for nonparametric multivariate data description, data analysis, and inference, including for example location and scatter estimation, tests of symmetry, generalized signed-rank tests, and multivariate boxplots. Depth-based tools for multivariate analysis were provided in Liu et al. (1999), approaches to formulation of depth functions were studied in Zuo and Serfling (2000), and an overview of depth and related functions was provided in Serfling (2006). In recent years, however, considerable further advances in the formulation and applications of depth functions have been made, accommodating additional goals and data settings. For example, variants sensitive to modality, and versions for functional data and shape-fitting problems, have been developed. Connections with level sets have been examined. We may ask some questions. What new possibilities are of interest? What fundamental features of depth and related functions have extensive generality of application? This talk will seek to provide an updated perspective on depth, outlyingness, quantile, and rank functions, through an overview coherently treating concepts, roles, properties, interrelations, data settings, applications, open issues, and new potentials.*

Keywords. *Depth functions; Rank functions; Multivariate; Functional data; Shape-fitting problems.*

References

Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Science USA* **85**, 1732-1734.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27**, 783-858.

Serfling, R. (2006). Depth functions in nonparametric multivariate analysis. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications* (R. Liu, R. Serfling, D. Souvaine, eds.), pp. 1-16. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 72, American Mathematical Society.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver 1974 (R. D. James, ed.)* **2**, 523-531.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28**, 461-482.

3.25

A goodness–of–fit test for parametric models with directional predictors

E. García–Portugués[1,*], I. Van Keilegom[2], R. M. Crujeiras[1] and W. González–Manteiga[1]

[1] *University of Santiago de Compostela; eduardo.garcia@usc.es, rosa.crujeiras@usc.es, wenceslao.gonzalez@usc.es*
[2] *Université catholique de Louvain; ingrid.vankeilegom@uclouvain.be*
[*] *Corresponding author*

Abstract. *A new test for assessing if the regression function of a linear variable on a directional predictor (circular, spherical or q–spherical, $q \geq 3$) belongs to a certain parametric family is proposed in this work. The test is based on kernel smoothing and confronts the parametric hypothesis against a nonparametric alternative. The squared distance between the smoothed parametric estimate and a new proposal of the local linear estimator for the regression function, adapted to deal with a directional predictor, is used as test statistic. The asymptotic distribution of the test is obtained, under simple and composite null hypotheses, and also for a family of local alternatives. A consistent resampling procedure for the practical calibration of the test is also provided. The performance of the method is illustrated in a simulation study. Finally, the test is applied to analyse several real datasets.*

Keywords. *Directional data; Goodness–of–fit; Local linear regression; Nonparametric testing.*

3.26

New goodness-of-fit diagnostics for conditional discrete response models

I. Kheifets[1,*], C. Velasco[2]

[1] *New Economic School, Moscow; ikheifets@nes.ru,*
[2] *Universidad Carlos III de Madrid; carlos.velasco@uc3m.es*
[*] *Corresponding author*

Abstract. *This paper proposes new specification tests for conditional models with discrete responses. In particular, we can test the static and dynamic ordered choice model specifications, which is key to apply efficient maximum likelihood methods, to obtain consistent estimates of partial effects and to get appropriate predictions of the probability of future events. The traditional approach is based on probability integral transforms of a jittered discrete data which leads to continuous uniform iid series under the true conditional distribution. Then, standard specification testing techniques could be applied to the transformed series, but the extra randomness from jitters affects the power properties of these methods. We investigate in this paper an alternative transformation based only on original discrete data. We analyze the asymptotic* *properties of goodness-of-fit tests based on this new transformation and explore the properties in finite samples of a bootstrap algorithm to approximate the critical values of test statistics which are model and parameter dependent. We show analytically and in simulations that our approach dominates the traditional approach in terms of power. We apply the new tests to models of the monetary policy conducted by the Federal Reserve.*

Keywords. *Specification tests; Dynamic discrete choice models; Conditional probability integral transform*

3.27

Goodness-of-fit tests in semiparametric transformation models

B. Colling[1,*], I. Van Keilegom[1]

[1] *Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium; benjamin.colling@uclouvain.be, ingrid.vankeilegom@uclouvain.be*
[*] *Corresponding author*

Abstract. *Consider a semiparametric transformation model of the form $\Lambda_\theta(Y) = m(X) + \epsilon$, where Y is a univariate dependent variable, X is a d-dimensional covariate, and ε is independent of X and has mean zero. We assume that $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions, while m is the unknown regression function. We use a profile likelihood estimator for the parameter θ and a semiparametric kernel estimator for m, and develop a test for the parametric form of the regression function m. The two test statistics that we propose are a Kolmogorov-Smirnov and a Cramér-von Mises type statistics, where the basic idea is to compare the distribution function of ε estimated in a semiparametric way to the distribution function of ε estimated under the null hypothesis. The limiting distributions of these two test statistics are established under the null hypothesis and under a local alternative. We use a bootstrap procedure to approximate the critical values of the test statistics under the null hypothesis. Finally, a simulation study is carried out to illustrate the performance of our testing procedure and an analysis of a real data set is presented.*

Keywords. *Goodness-of-fit; Kernel smoothing; Profile likelihood; Semiparametric regression; Transformation model.*

References

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society - Series B* **26**, 211–252.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21**, 1926–1947.

Linton, O., Sperlich, S., Van Keilegom, I. (2008). Estimation of a Semiparametric Transformation Model. *Annals of Statistics* **36**, 686–718.

Neumeyer, N., Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis* **101**, 1067–1078.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* **25**, 613–641.

Van Keilegom, I., González-Manteiga, W., Sánchez Sellero, C. (2008). Goodness of fit tests in parametric regression based on the estimation of the error distribution. *TEST* **17**, 401–415.

3.28

Quantile estimation in varying coefficient models and non-crossingness of conditional quantile curves

Y. Andriyana[1,*], I. Gijbels[1] and A. Verhasselt[2]

[1] *KU Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Leuven, Belgium; yudhie.andriyana@wis.kuleuven.be , irene.gijbels@wis.kuleuven.be*
[2] *Universiteit Hasselt, Interuniversity Institute for Biostatistics and statistical Bioinformatics, CenStat, Hasselt, Belgium; anneleen.verhasselt@uhasselt.be*
[*] *Corresponding author*

Abstract. *A (unconditional) quantile function is an increasing function in its argument, say p. In real applications, the impact of explanatory variables on a variable of interest, leads to the study of conditional quantile functions. For a given value of p, a conditional quantile function is thus a function of a covariate (or several covariates). In practice, the conditional quantile functions are estimated, from data, for various fixed values p, i.e. we estimate the conditional median function, but also the conditional first and third quartile function. These conditional quantile functions are helpful to further describe/predict the impact of covariates on the response variable. See e.g. Koenker and Basset (1978). In Andriyana et al. (2014) we developed procedures for estimating conditional quantile functions in varying coefficient models (see e.g. Hastie and Tibshirani (1993)). Such flexible models can adequately describe complex data structures. We employ P-spline estimation techniques (see e.g. Eilers and Marx (1996)). Since the conditional quantile functions are estimated for a given set of values of p, the resulting estimated conditional quantile curves may cross each other (in particular for small data sets), whereas this is not possible conceptually. In this talk, we explain how to obtain estimated quantile curves that do not cross, in a varying coefficient modeling setting. We illustrate the practical use of estimating conditional quantile curves in real data applications, involving longitudinal data.*

Keywords. *Longitudinal data; Non-crossing quantile curves; P-splines; Varying coefficient models.*

References

Andriyana, Y., Gijbels, I., Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficient models. *Test.* DOI: 10.1007/s11749–013–0346–2.

Eilers, P., Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–102.

Hastie, T., Tibshirani, R. (1993). Varying-coefficient models. *Journal of The Royal Statistical Society, Series B* **55**, 757–796.

Koenker, R., Bassett, G., Jr. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

3.29

Testing for constancy in varying coefficient models

M. Ahkim[1,*] and A. Verhasselt[2]

[1] *Applied Mathematics, Universiteit Antwerpen, Belgium; Mohamed.Ahkim@uantwerpen.be,*
[2] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium; Anneleen.Verhasselt@uhasselt.be*
[*] *Corresponding author*

Abstract. *Linear regression models are often too rigid for regression analysis. We consider varying coefficient models, which are an extension of the classical linear regression models in the sense that the regression coefficients are replaced by functions in certain variables (often time t). Varying-coefficient models have been popular in longitudinal data and panel data studies, and have been applied in fields such as finance and health sciences. We estimate the coefficient functions by the flexible P-spline technique. An important question in a varying coefficient model is whether an estimated coefficient function is statistically significantly different from a constant (or zero). We develop testing procedures based on the P-spline coefficients by making use of properties of B-spline basis expansions.*
The performances of the proposed testing methods are illustrated on simulated data and on epidemiology data.

Keywords. *varying-coefficient models, longitudinal data, hypothesis testing*

3.30

Precautionary savings over the life cycle: a simple two step locally constant least squares estimator

Juan M. Rodríguez-Poo[1] and Alexandra Soberón[2,*]

[1] *Departamento de Economía. Universidad de Cantabria. Avda. de los Castros s/n. 39005. Santander. Spain ; rodrigjm@unican.es*
[2] *Departamento de Economía. Universidad de Cantabria ; soberonap@unican.es*
[*] *Corresponding author*

Abstract. *This paper considers the nonparametric estimation of a structural model of optimal life-cycle savings, controlling for both uncertainty about health-care expenditures and household risk aversion. The main attraction of this estimator is that, compared to those already proposed* *in the literature, it allows to deal simultaneously with different problems such us unobserved cross-sectional heterogeneity, varying parameters of unknown form in the Euler equation and endogenous covariates. The estimator of the function of interest turns out to have the simple form of a two-step weighted locally constant least-squares estimator. Additionally, some marginal integration technique is needed to compute a subset of the functionals of interest. In the paper, we establish the asymptotic properties of these estimators. To illustrate the feasibility and possible gains of this method, the paper concludes with an application about household's precautionary savings over the life-cycle. From this empirical application, we obtain the following conclusions: Households accumulate wealth at least in two periods in life. In younger stages, the savings are to guard to uncertainty about potential income downturn, whereas when they are older, household save for retirement and bequests. Furthermore, public health programs have a negative impact on precautionary savings. Finally, by comparing educational levels we obtain that households with low education levels are more risk averse than those with higher levels.*

Keywords. *Panel data; varying coefficient models; endogeneity; fixed effects; nonparametric techniques*

3.31

Model-based clustering for high-dimensional regression data

E. Devijver[1]

[1] *INRIA Select, Université Paris-Sud, 91405 Orsay Cedex; emilie.devijver@math.u-psud.fr*

Abstract. *Finite mixture regression models are useful for modeling the relationship between response and predictors, arising from different subpopulations. We will work with a finite mixture regression model. In each class, the model considered is linear.*

We look at high-dimensional predictors and high-dimensional response. We propose two procedures to deal with this high-dimensional issue. The main idea is to construct a model collection, varying sparsity, and refitting to have better estimations. We conclude by selecting the best model using the slope heuristic, developped in Birgé (2007). The first procedure uses the Lasso estimator to take into account the coefficient sparsity. This generalizes the paper of Maugis (2012). The second one uses, furthermore, a penalty on the rank, to take into account the matrix structure. This follows the work of Giraud (2011) in the mixture case.

The main algorithm developp for this model is an extension of the EM algorithm. From this perspective, we generalize in the multivariate case the approach of the article of Städler (2010) for mixture regression model.

We extend these procedures to the functional case, where predictors and responses could be functions. For this, we use a wavelet-based approach, to consider the wavelet coefficients instead of the discretization of the function. We will evaluate our procedures on simulated datasets to approve each step, and illustrate on real datasets.

This work is supervized by Pascal Massart and Jean-Michel Poggi.

Keywords. *Model-based clustering ; Regression ; High-dimension ; Functional data.*

References

Städler, N. et al. (2010). ℓ_1-penalization for mixture regression models. *Test* **19**, 209-256.

Maugis, C. and Meynet, C. (2012). A sparse variable selection procedure in model-based clustering. *Rapport de Recherche INRIA*

Giraud, C. (2011). Low rank multivariate regression. *Electronic Journal of Statistics* **5**, 775-799.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**.

3.32

A new data-driven method for estimating the support

Paula Saavedra-Nieves[1,*] and Alberto Rodríguez-Casal[1]

[1] *University of Santiago de Compostela; paula.saavedra@usc.es, alberto.rodriguez.casal@usc.es*
[*] *Paula Saavedra-Nieves*

Abstract. *This work deals with the problem of estimating the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X from independent and identically distributed observations, $\mathcal{X}_n = \{X_1, ..., X_n\}$, taken in it.*

Several proposals for reconstructing S have been considered in the literature. For instance, Devroye and Wise (1980) proposed

$$S_n = \bigcup_{i=1}^{n} B_{\epsilon_n}[X_i]$$

as an estimator of S, where $B_{\epsilon_n}[X_i]$ denotes the closed ball with center X_i and radius ϵ_n depending only on n. More sophisticated estimators can be used if we have some additional information on the set. For example, if we know that S is convex then the convex hull of the sample is a natural support estimator. But convexity assumption may be too restrictive in practice. For instance, if S is not connected. Then, it is necessary to consider a more flexible shape restrictions such as $r-$convexity. A closed set $A \subset \mathbb{R}^d$ is said $r-$convex, for $r > 0$, if $A = C_r(A)$, where

$$C_r(A) = \bigcap_{\{B_r(x):B_r(x)\cap A=\emptyset\}} (B_r(x))^c$$

denotes the $r-$convex hull of A and $B_r(x)$, the open ball with center x and radius r. Furthermore, if A is $r-$convex then it is $\overline{r}-$convex for all $\overline{r} \leq r$. So, if we assume that S is $r-$convex then a natural support estimator will be the $r-$convex hull of the sample points, $C_r(\mathcal{X}_n)$. In Rodríguez-Casal (2007), it is proved that if r is correctly chosen, the $r-$convex hull of the sample achieves the same convergence rates (in Hausdorff and distance in measure) as the convex hull. But in practice, S is unknown and, consequently, the real value of the smoothing parameter r too.

In this work, we propose an almost sure consistent estimator of the largest value of r such that S

is $r-$convex, under the assumption that the distribution of X is uniform on S. The estimator of r is based on the uniformity test proposed by Berrendero, Cuevas and Pateiro-López (2012). The support estimator obtained from this smoothing parameter, which is fully automatic given the random sample, is able to achieve the same convergence rates as the convex hull for estimating convex sets. An application to a real data example is considered. It is analyzed if the water area in the Aral Sea is decreasing.

Keywords. *Support estimation; Geometric properties; $r-$convexity; Uniform distribution*

3.33

Algorithms to estimate the λ-medial axis

Antonio Cuevas[1], Pamela Llop[2] and Beatriz Pateiro-López[3,*]

[1] *Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain; antonio.cuevas@uam.es*
[2] *Facultad de Ingeniería Química (UNL) and Instituto de Matemática Aplicada del Litoral (UNL - CONICET), Argentina; pllop@santafe-conicet.gov.ar*
[3] *Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain; beatriz.pateiro@usc.es*
[*] *Corresponding author*

Abstract. *This work deals with the statistical problem of estimating a modified version of the medial axis, called λ-medial axis, introduced in Chazal and Lieutier (2005). The whole approach relies on a simple plug-in idea using methods of set estimation. The consistency of the proposed estimators under some, not too restrictive, regularity assumptions on C is derived. We also propose two algorithms to compute the exact λ-medial axis of sets whose shape is given by a union of balls (such as the Devroye-Wise estimator) or by the complement of a union of balls (such as the r-convex hull estimator).*

Keywords. *Medial axis; Set estimation; r-convexity*

References

Amenta, N., Kolluri, R.K. (2001). The Medial Axis of a Union of Balls. *Comput. Geom. Theory Appl.*, ***20****, 25–37.*

Chazal, F. and Lieutier, A. (2005). The "λ-medial Axis". J. Graphical Models, ***67****, 304–331.*

3.34

Adaptive estimation of convex polytopes and convex bodies

Victor-Emmanuel Brunel[1]

[1] CREST-ENSAE, Malakoff, France. University of Haifa, Israel; victor.emmanuel.brunel@ensae-paristech.fr

Abstract. *We are interested in two models. The first one consists of observing a sample of i.i.d. random variables, with uniform distribution on some unknown subset of $\mathbb{R}^d, d \geq 1$. The second one consists of a regression setup, where the regression function is the indicator on some unknown subset of $\mathbb{R}^d, d \geq 1$.*

In both case, we estimate the unknown set under two possible hypotheses. First, we assume that the unknown set is a convex polytope with r vertices, and $r \geq d+1$ is a known integer. If r is not known, we propose an adaptive estimator which achieves the same rate of convergence as in the known r case. Second, we assume that the unknown set is any convex body, and we give the corresponding minimax rate of convergence.

3.35

Regularitation techniques in recovering structural breaks in nonparametric regression

Matúš Maciak[1,*] and Ivan Mizera[2]

[1] T.G.Masaryk Water Research Institute, Prague, Czech Republic; maciak@ualberta.ca
[2] University of Alberta, AB, Canada; mizera@ualberta.ca
[*] Corresponding author

Abstract. *More and more data are collected every day forcing statisticians to deal with far more complex data structures than before. By implication, more sophisticated models are assumed and different types of structural breaks (change-points) are more common to occur.*

We propose a new approach to such change-point estimation in regression. It combines nonparametric regression estimation with different concepts of an L_1-norm regularization. The main advantage of our method is that it introduces a fully data-driven approach with no requirement on a prior knowledge of any kind. This is usually not true for classical methods used in such situations: indeed, other methods are either not fully automatic and they involve a kind of multiple stage estimation or, they require a prior knowledge for change-point positions instead. We discuss various alternatives motivated by different practical situations. A proper statistical inference is given and theoretical results are derived. Finite sample performance is investigated using simulated data and a real example as well.

3.36

Claim reversing using distance-based generalized linear models

T. Costa[1,*] and E. Boj[2]

[1] *Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Avinguda Diagonal 690, 08034 Barcelona, Spain; tcosta@ub.edu*
[2] *Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Avinguda Diagonal 690, 08034 Barcelona, Spain; evaboj@ub.edu*
[*] *Corresponding author*

Abstract. *As is well known, generalized linear models (GLM) can be considered as a stochastic version of the classical Chain-Ladder method of claim reserving in non-life insurance. We refer, e.g., to England (1999) and England and Verrall (2002) for a detailed description. In particular, the deterministic Chain-Ladder model is reproduced when a GLM is fitted using overdispersed Poisson error distribution and logarithmic link.*

Our aim is to propose the use of distance-based generalized linear models (DB-GLM) in the claim reserving problem. We refer to Boj et al. (2012) where the main characteristics of the DB-GLM are studied. DB-GLM can be considered a generalization of the classical GLM to the distance-based analysis. The only information required to fit these models is a predictor distance matrix. DB-GLM can be fitted using the dbstats package for R (Boj et al., 2013).

It is important to point out that DB-GLM contains as a particular instance ordinary GLM. Then it can be considered too as a stochastic Chain-Ladder claim reserving method. To complement the methodology and estimate reserve distributions and standard errors we develop a bootstrap technique adequate to the DB-GLM.

This research is part of the project: Semiparametric and distance-based methodologies with applications in bioinformatics, finance and risk management (grant MTM2010-17323).

Keywords. *Claim reserving; Generalized linear model; Bootstrap; Chain-Ladder; Distance-based prediction.*

References

Boj, E., Delicado, P., Fortiana, J., Esteve A. and Caballé, A. (2012). Local distance-based generalized linear models using the dbstats package for R. *Documentos de Trabajo de la Xarxa de Referència en Economia Aplicada (XREAP)*, XREAP2012-11, 2012 (Submitted to the Journal of Statistical Software, in Second Revision).

Boj, E., Caballé, A., Delicado, P. and Fortiana, J. (2013). *dbstats: Distance-Based Statistics (dbstats)*. `R` package version 1.0.3, 2013, URL `http://CRAN.R-project.org/package=dbstats`.

England, P.D. and Verrall, R.J. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics* **25**, 281–293.

3.37

Change-point for regression derivative with non-stationary errors.

S. Ben Hariz [1,*] and M. Zorgu[2]

[1] *Laboratoire Manceau de Mathématiques. Université du Maine; France ; shariz@univ-lemans.fr*
[2] *Université de Kairouan , Kairouan, Tunisie; zorgui3@yahoo.fr*
[*] *Corresponding author*

Abstract.
We consider the regression model

$$Y_i = g(x_i) + \varepsilon_i, \quad i = 0, 1, 2..., n,$$

where the regression function derivative has a jump point at an unknown position θ*. We propose a nonparametric Kernel-based estimator of the jump location* θ*. Assume that* $\sup_{|i-j|\geq k} |Cov\,(\varepsilon_i, \varepsilon_j)| \leq Ck^{-\rho}$ *for* $0 < \rho \leq 1$*. Under very general conditions, we prove the* $(nh)^{\frac{-\rho}{2}}$ *convergence rate of the estimator, where* h *is the window of the kernel. This includes short-range dependent as well as long-range dependent and even non-stationary errors. Finally, we gives conditions on the windows* h *to obtain the best rate of convergence. The obtained rate is known to be optimal for i.i.d. errors as well as for LRD errors*

Keywords. *Change-point; Kink estimation; Nonparametric regression; Rate of convergence; Long-range dependent.*

References

Ben Hariz, Samir; Wylie, Jonathan J.; Zhang, Qiang Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. Ann. Statist. 35 (2007), no. 4, 1802–1826.

Raimondo, Marc. Minimax estimation of sharp change points. Ann. Statist. 26, (1998), no. 4, 1379–1397.

3.38

Selection of the smoothing parameter for sparse regularization of linear inverse problems

C. Giacobino[1,*], S. Sardy[1]

[1] *University of Geneva, Section of Mathematics; caroline.giacobino@unige.ch, sylvain.sardy@unige.ch*
[*] *Corresponding author*

Abstract. *Statistical linear inverse problems arise in many scientific settings, ranging from medical imaging to astronomy. They pertain to situations where one is interested in estimating an unknown object based on noisy observations of a linear transform of that object. We consider sparsity regularization in linear inverse problems using the lasso. The selection of the regularization parameter is investigated: we first derive an unbiased risk estimate based on Stein (1981) and then the quantile universal threshold based on DJ (1994). The deblurring problem (e.g. the Hubble telescope) is considered as an example.*

Keywords. *Lasso; Wavelets; Sparsity.*

References

Donoho, D.L. and Johnstone, I.M. (1994). Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika* **81**, 425–455.

Stein, C.M. (1986). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* **9**, 1135–1151.

3.39

A statistical solution to the illumination inverse problem from a single image

M. Vimond[1], N. Klutchnikoff[1] and S. Geaffray[2,*]

[1] *Centre de Recherche en Economie et STatistique (Ensai); mvimond@ensai.fr, nklutchnikoff@ensai.fr*
[2] *Université de Strasbourg; geffray@math.unistra.fr*
[*] *Corresponding author*

Abstract. *The appearance of a scene is determined to a great extent by the illumination conditions. The presence of illumination artifacts is obviously undesirable especially when automatic measurement from digital images is the final goal, as it is required in Scanning Electron Microscopy. Retinex Theory (Land and McCann, 1971) attempt to simulate and explain how the human visual system perceives color. Then several Retinex algorithms have been developed ever since, see for example Hou (2006).*
The illumination artifact is used to be modeled by a function L which is "smooth" enough, in a sense to be precised in the mathematical developments, and which acts in a multiplicative way on the original signal R. We also assume the presence of an additional additive noise ϵ so that the observed image Y is actually linked to the original signal R by the equation: $Y(x) = R(x)L(x) + \epsilon(x)$, for x describing the pixel's domain. In this framework, our aim consists of estimating R from the observation of Y. To this aim, we improve a semi-parametric regression strategy proposed by Tasdizen et al. (2008). First using local polynomials (Ruppert and Wand, 1994), we estimate consistently the gradient of the log*-signal. Then we project it on a suitable finite dimensional subspace so that we get an estimation of $\log L$ and deduce an estimation of R. The procedure quality is studied from a theoretical point of view through the rate of convergence of uniform risk. At last, an application to real electron microscopy images is presented.*

Keywords. *SEMIPARAMETRIC REGRESSION; STATISTICAL IMAGE PROCESSING; RETINEX THEORY; COLOR PERCEPTION.*

References

E.H. Land and J.J. McCann. (1971). Lightness and the retinex theory. *J. Opt. Soc. Am.* **61**, 1–11.

Z. Hou. 2006. A review on mr image intensity inhomogeneity correction. *International Journal of Biomedical Imaging.*

T. Tasdizen, E. Jurrus, and R.T. Whitaker. 2008. Non-uniform illumination correction in transmission electron microscopy. *Preprint*

D. Ruppert and M. P Wand. 1994. Multivariate locally weighted least squares regression. *The annals of statistics* **22 (3)**, 1346–1370.

3.40

Nonparametric and semiparametric inference for signal/image symmetries

Mirosław Pawlak

Department of Electrical & Computer Eng., University of Manitoba, Canada
pawlak@ee.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.*

Keywords. *Symmetry detection, symmetry estimation, radial polynomials, limit distributions, degree of symmetry.*

3.41

Nonparametric estimation in interval censored data using the Bezier curve

Choongrak Kim[1,*] and Whasoo Bae[2]

[1] *Department of Statistics, Pusan National University; crkim@pusan.ac.kr*
[2] *Department of Data Science/ Institute of Information, Inje University; statwbae@inje.ac.kr*
[*] *Corresponding author*

Abstract. *In this paper we propose a Bezier curve method to estimate the survival function and the median survival time in interval-censored data. We compare the proposed estimator with other existing methods such as the parametric method, the single point imputation method, and the nonparametric maximum likelihood estimator through extensive numerical studies, and it is shown that the proposed estimator performs better than others in the sense of mean squared error and mean integrated squared error. An illustrative example based on a real data set is given.*

Keywords. *Imputation method; Kaplan-Meier estimator; Median survival time; Survival function*

3.42

In sample forecasting with local linear survival densities

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

[1] *Cass Business School, City University London, 106 Bunhill Row, London EC1Y8TZ, U.K.; munir.hiabu.1@cass.city.ac.uk, Maria.Miranda.1@city.ac.uk, Jens.Nielsen.1@city.ac.uk*
[2] *Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim;emammen@rumms.uni-mannheim.de*
[*] *Corresponding author*

Abstract. *In-sample forecasting is in this paper defined as forecasting a structured density to sets where the density is not observed. The structured density consists of one-dimensional in sample components that identify the density on such sets. This paper focus on the multiplicative density structure that recently have been seen as the underlying structure of non-life insurance forecasts. In non-life insurance the in-sample area is defined as one triangle and the forecasting area as the triangle that added to the first triangle produces a square. Recent approaches estimate two one-dimensional components by projecting an unstructured two-dimensional density estimator down on a multiplicative space. This paper shows that a simple time-reversal reduces the problem to two one-dimensional problems. On the reversed time axis the one-dimensional data is left truncated and a one-dimensional survival density estimator is needed. This paper* *uses the local linear density smoother with a weighted do-validated bandwidth selector. Full asymptotic theory of the weighted do-validated bandwidth selector is given along with finite sample studies and real life applications to non-life insurance.*

Keywords. *Aalen's multiplicative model; Cross-validation; Do-validation; Density estimation; Local linear kernel estimation; Survival data.*

3.43

Nonparametric estimation of a distribution function from doubly truncated data under dependence

C. Moreira[1,2,*], J. de Uña-Álvarez[1] and R. Braekers[3]

[1] *University of Vigo - Department of Statistics and O.R. Lagoas - Marcosende, 36 310, Vigo, Spain; carlamgmm@gmail.com; jacobo@uvigo.es*
[2] *Center of Mathematics and Department of Mathematics and Applications - Campus de Azurém, 4800-058 Guimarães, Portugal;*
[3] *University of Hasselt, Campus Diepenbeek - Center for Statistics, 3590 - Diepenbeek, Belgium; roel.braekers@uhasselt.be*
[*] *Corresponding author*

Abstract. *The NPMLE of a distribution function from doubly truncated data was introduced in the seminal paper of Efron and Petrosian Efron and Petrosian (1999). The consistency of the Efron-Petrosian estimator depends however on the assumption of independent truncation. In this work we introduce and extension of the Efron-Petrosian NPMLE when the lifetime and the truncation times may be dependent. The proposed estimator is constructed on the basis of a copula function which represents the dependence structure between the lifetime and the truncation times. Two different iterative algorithms to compute the estimator in practice are introduced, and their performance is explored through an intensive Monte Carlo simulation study. The asymptotic properties of the proposed estimator will be explored. Several applications to medical data are included for illustration purposes.*

Keywords. *Copula function; Dependent truncation; Double truncation*

References

Efron, B. and V. Petrosian (1999): Nonparametric methods for doubly truncated data, *Journal of the American Statistical Association*, **94**, 824 - 834.

3.44

A general approach for cure regression models in survival analysis

V. Patilea[1,*] and I. Van Keilegom[2]

[1] *CREST & Ensai; valentin.patilea@ensai.fr*
[2] *ISBA/UCL; ingrid.vankeilegom@uclouvain.be*
[*] *Corresponding author*

Abstract. *Cure regression models are a special topic in time-to-event statistical analysis. Such models take into account situations where a proportion of subjects will never experience the event under study. In such a case the time to event is considered infinite. For instance, medical studies could reveal a proportion of patients for whom the disease under surveillance will never recur, and these patients could be considered as cured. A well studied topic in Labor Economics is the time to get a new job after a permanent layoff. Quite often a proportion of the labor force will withdraw and never get a new job. In most of the applications the statistical analysis of cure models is made more difficult by the presence of a finite random right censorship. Indeed, the censoring prevent from knowing whether a censored observation has a finite or infinite time to event. In this paper we introduce new explicit representations of the conditional probability of being cured and of the conditional law of the time to event as functions of the law of the observations. We derive these representations under minimal conditional independence assumptions that are required for the identification of the law of the time variable of interest. Such explicit representations bring valuable insight on the assumptions behind the existing cure regression models, namely the two-component mixture cure models. Moreover, the explicit representations allow us to state general identification results and to derive a new general maximum likelihood estimation approach.*

Keywords. *Cure regression models; Nonparametric estimation; Random censoring; Semiparametric maximum likelihood; Asymptotic theory.*

References

Tsodikov, A.D., J. G Ibrahim, J.G. & Yakovlev, A.Y. (2003). Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models. *J. Amer. Statist. Assoc.* **98(464)**, 1063–1078.

Zheng, D., Yin, G. & Ibrahim, J.G. (2006). Semiparametric Transformation Models for Survival Data With a Cure Fraction. *J. Amer. Statist. Assoc.* **101(474)**, 670–684.

3.45

Prediction and variable selection in AFT cure models with interval censored data.

A. El Ghouch[1,*], S. Scolas[1] and C. Legrand[1]

[1] *Université catholique de Louvain; anouar.elghouch@uclouvain.be, scolas.sylvie@uclouvain.be, catherine.legrand@uclouvain.be.*
[*] *Corresponding author*

Abstract. *In this talk, we study an accelerated failure time (AFT) regression model when the data is subject to a general interval censoring scheme and in the presence of immune or cured individuals. We utilize an extended generalized gamma (EGG) distribution for the error term of the AFT model and a logistic regression for the cure proportion (i.e. the incidence part). We investigate some issues concerning this model, including estimation, variable selection, model simplification and prediction. We then apply this method to an Alzheimer disease database, which consists in 241 at-risk patients followed-up between 1998 and 2008 with regular checks for the appearance of mild cognitive impairment (MCI).*

Keywords. *Transformation; Likelihood; Numerical optimization; Adaptive LASSO.*

3.46

Recent advances in nonparametric methods for mixture cure models

Yingwei Peng

Department of Public Health Sciences, Department of Mathematics and Statistics, Cancer Care and Epidemiology at Cancer Research Institute, Queen's University; yingwei.peng@queensu.ca

Abstract. *Mixture cure models have long been used to model survival data with a cured fraction. Even though other types of cure models have been proposed as alternatives, the mixture cure model still received a great deal of attention in recent years. In this talk, I will present some recently proposed nonparametric methods that provide flexible ways to model covariate effects on cure rate and to model clustered survival data with a cured fraction. The performance of the methods will be illustrated with simulation studies and demonstrated in applications to cancer survival data sets.*

Keywords. *Correlation structures; EM algorithm; Estimating equations; Kernel smoothing.*

3.47

Partially linear transformation cure models for interval-censored data

T. Hu[1] and L. Xiang[2,*]

[1] *Capital Normal University, Beijing, China*
[2] *Nanyang Technological University, Singapore; lmxiang@ntu.edu.sg*
[*] *Corresponding author*

Abstract. *There has been considerable progress in the development of semiparametric transformation models for regression analysis of time-to-event data. However, most current work focuses on right-censored data. Significantly less work has been done for interval-censored data especially when the population contains a nonignorable cured subgroup. In this paper, we present a broad and flexible class of semiparametric transformation cure models for analyzing interval-censored data in the presence of a cure fraction. Our approach combines a logistic regression formulation for the probability of cure with partially linear transformation models for event times of susceptible subjects. We develop a spline-based sieve maximum likelihood estimation, which is computationally efficient and leads to estimators with appealing properties such as consistency, asymptotic normality and semiparametric efficiency. Furthermore, we propose a goodness-of-fit test for the proposed model based on the sieve likelihood ratio. Simulations and a real data analysis are provided for illustration of the methodology.*

Keywords. *Cure rate; Interval censoring; Semiparametric efficiency; Sieve likelihood ratio test; Transformation.*

3.48

Nonstationary cross-validation with applications to predictive regression

Valentina Corradi

Abstract. *Cross-validation is the most common data-driven procedure for choosing the bandwidth sequence in nonparametric regression. For the case of* i.i.d or strong mixing data, it is well-known that the bandwidth chosen by cross-validation is optimal with respect to the mean integrated squared error. However, the properties of cross-validated bandwidths in the context of nonstationary regressions have not yet been established. This is the subject of the current paper. For the case of β-recurrent (stationary or nonstationary) Markov chains, we show that the bandwidth chosen via cross-validation is optimal with respect to the average squared error. The accuracy of estimators based on cross-validated bandwidths is analyzed via a Monte Carlo study. The practical usefulness of cross-validated bandwidths in a highly-persistent, possibly nonstationary environment is illustrated by virtue of an application to nonlinear predictive regressions.

3.49

Robust econometric inference for stock return predictability

Alexandros Kostakis, Tassos Magdalinos and Michalis Stamatogiannis

Abstract. *This study examines stock return predictability via lagged financial variables with unknown stochastic properties. We conduct a battery of predictability tests for US stock returns during the period 1927-2012, proposing a novel testing procedure which: i) robustifies inference to the degree of persistence of the employed regressors, ii) accommodates testing the joint predictive ability of financial variables in multiple regression and iii) is easy to implement as it is based on a linear estimation procedure. We provide significant evidence in favor of short-horizon predictability in the full sample period. Nevertheless, this evidence considerably weakens in the post-1952 period.*

Keywords. *Stock returns; Predictability; Persistent regressors; Robust inference*

3.50

Unbiased QML estimation of log-GARCH models in the presence of zero returns

G. Sucarrat[1,*] and A. Escribano[2]

[1] *Department of Economics, BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway. Webpage: http://www.sucarrat.net/.; genaro.sucarrat@bi.no*
[2] *Department of Economics, Universidad Carlos III de Madrid (Spain); alvaroe@eco.uc3m.es*
[*] *Corresponding author*

Abstract. *A critique that has been directed towards the log-GARCH model is that its log-volatility specification does not exist in the presence of zero returns. A common "remedy" is to replace the zeros with a small (in the absolute sense) non-zero value. However, this renders Quasi Maximum Likelihood (QML) estimation asymptotically biased. Here, we propose a solution to the case where actual returns are equal to zero with probability zero, but zeros nevertheless are observed because of measurement error (due to missing values, discreteness approximisation error, etc.). The solution treats zeros as missing values and handles these by combining QML estimation via the ARMA representation with the Expectation-maximisation (EM) algorithm. Monte Carlo simulations confirm that the solution corrects the bias, and several empirical applications illustrate that the bias-correcting estimator can make a substantial difference.*

Keywords. *ARCH; Exponential GARCH; Log-GARCH; ARMA; Expectation-Maximisation (EM).*

3.51

A unified theory for time varying models: foundations and applications in the presence of breaks and heteroskedasticity

A. Paraskevopoulos[1],M. Karanasos[2] and S. Dafnos[3]

Abstract. *The paper develops an elegant approach to examine the dynamics of stochastic time series models with time dependent coefficients. We provide the closed form of the fundamental solutions for time varying autoregressive moving average models which is a long standing research topic. This enable us to characterize these models by deriving i) its multistep ahead predictor; ii) the first two unconditional moments; and iii) its covariance structure. In addition, capitalizing on the connection between linear difference equations and the product of companion matrices, we employ our general methodology to obtain an explicit formula for the latter. We also apply our method to obtain results on generalized continuant matrices. To illustrate the practical significance of our results we consider autoregressive models with multiple breaks and also apply our unified approach to a variety of processes such as i) periodic, cyclical and smooth transition autoregressive models, ii) time varying generalized autoregressive conditional heteroscedasticity specifications, and iii) generalized random coefficients autoregressive models*

3.52

Variable selection in high-dimensional exponential dispersion models

A. Antoniadis[1,*], I. Gijbels[2], S. Lambert-Lacroix[3] and J.-M. Poggi[4]

[1] *University Joseph Fourier, Grenoble, France; antonia@imag.fr;*
[2] *KUL, Leuven, Belgium; Irene.Gijbels@wis.kuleuven.be;*
[3] *University Pierre Mendes-France, Grenoble, France; Sophie.Lambert@imag.fr;*
[4] *University Paris-Sud, Orsay, France; Jean-Michel.Poggi@math.u-psud.fr;*
[*] *Corresponding author*

Abstract. *GLM models allow us to model responses which are not normally distributed, using methods closely analogous to linear methods for normal data. They assume an exponential family distribution for the response variable and are more general than linear models in that they accommodate a mean-variance relationship and choose an appropriate scale for modelling the mean on which the action of the covariates is approximately linear. In such models, the variance is assumed known up to a constant of proportionality, the dispersion parameter. A flexible extension of GLM models is the class of GAM models, allowing for arbitrary functions for modelling the influence of each covariate on an exponential family response in a multivariate regression setting. But often the observed data exhibit greater variability than the one implied by the mean-variance relationship and then, the loss of efficiency in estimating the parameters or additive components, using constant dispersion models when the dispersion is varying, may be* *substantial. Our work considers flexibly modelling the mean and variance functions within the framework of exponential dispersion models (and their additive extensions), a class of somehow over-dispersed generalized linear models. Our model describes the mean and dispersion parameters in terms of transformed additive functions of the predictors. Each of the additive terms can be either null, linear, or a fully flexible smooth effect. When the dispersion model is null we obtain a GLM, whereas with a null dispersion model and fully flexible smooth terms in the mean model we obtain a GAM. Whether or not to include predictors, and wether or not to model overdispersion at all is determined using a penalized likelihood-like variable selection approach with a possible diverging with the sample size number of parameters. We use a penalized criterion obtained by replacing the negative loglikelihood in the conventional penalized likelihood with Bregman divergence. With appropriate selection of the tuning parameters, we investigate the consistency of the variable selection procedure and asymptotic properties of the resulting estimators are established. The methodology is illustrated using real and simulated data.*

Keywords. *Exponential dispersion models; Variable selection; Penalized Bregman divergence.*

3.53

Testing additivity in nonparametric regression

E. Matzner-Løber

Univ Rennes and Agrocampus; eml@uhb.fr

Abstract. *In the talk we will discuss the possibility of using IBR for testing additivity*

Multivariate nonparametric smoothers, such as kernel based smoothers and thin plate splines smoothers, are adversely impacted by the sparseness of data in high dimension, also known as the curse of dimensionality. The common wisdom is to avoid all together general nonparametric smoothing with moderate sample size in dimensions higher than three. In this cases, it is usual practice in the statistical community to fit structurally constrained regression models such as additive models, MARS, projection pursuit models or additive L_2-Boosting.

Iterative bias reduction (IBR) is a practical and simple fully nonparametric multivariate smoothing procedure that adapts to the underlying smoothness of the true regression function. IBR is easily computed by successive application of existing base smoothers (without the need of selecting an optimal smoothing parameter), such as thin-plate spline, Duchon splines or kernel smoothers. In this talk, we will discuss the use of IBR to test the additivity of a function.

Keywords. *Nonparametric regression; iterative bias reduction; additive models; splines; kernel*

References

Cornillon, P.A. and Hengartner, N. and Jégou, N. and Matzner-Løber, E. (2014). Recursive Bias Estimation for multivariate regression *ESAIM* **36**, 326–329.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC.

3.54

Confidence interval for a function-valued predictor for non stationary processes

A. Antoniadis[1], X. Brossat[2], J. Cugliari[3] and J.-M. Poggi[4,*]

[1] *Univ. Joseph Fourier, Grenoble, France; Anestis.Antoniadis@imag.fr,*
[2] *EDF, Clamart, France; Xavier.Brossat@edf.fr,*
[3] *Univ. Lumière Lyon 2, Lyon, France; Jairo.Cugliari@univ-lyon2.fr,*
[4] *Univ. Paris-Sud, Orsay, France; Jean-Michel.Poggi@math.u-psud.fr*
[*] *Corresponding author*

Abstract.
The motivation of this work comes from electricity consumption forecasting and uses a function-valued time series representation of the discrete electricity records. The data are seen as a sequence of functions $Z_1(t), \dots, Z_n(t)$ with $t \in T$ representing daily load curves and typically the aim is to predict the function $Z_{n+1}(t)$ that corresponds to the next day load curve. If Z is stationary, then Antoniadis et al.(2006) propose the KWF (Kernel + Wavelet + Functional) predictor based on a non linear autoregressive model. The general principle of the forecasting model is to find in the past similar situations to the present and linearly combine their futures to build the forecast. Thus, using an appropriate similarity measure one can obtain the functional predictor in terms of weights defined through a kernel and a dissimilarity measure between curves based on wavelets. If the functional time series Z is non stationary, the KWF predictor fails to correctly predict we propose several strategies to take into account the various sources of non stationarity allowing to handle situations such that the mean level of the series changes over time or if there exists groups in the data that can be modeled as classes of stationarity. We study here the construction of a confidence interval for the prediction. The original prediction method, assuming that Z is stationary, uses a bootstrap re-sampling scheme to construct the confidence interval. We adapt some of the non stationarity corrections proposed for the pointwise prediction to construct a confidence interval for the prediction on the non stationary case.
See Antoniadis et al. (2012) and (2014) for details.

Keywords. *Confidence interval, Nonparametric forecasting, Functional data, Non stationary processes.*

References

Antoniadis, A., Paparoditis, E. and Sapatinas, T. (2006). A functional wavelet-kernel approach for time series prediction. *JRSS B* **68(5)**, 837–857.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2012). Prévision d'un processus à valeurs fonctionnelles en présence de non stationnarités. *J. SFdS* **153(2)**, 52–78.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J.-M. (2014). Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. *J. SFdS* **155(2)**, 202–219.

3.55

Nonparametric point source estimation in Cosmology with block total variation and block lasso

S. Sardy[1]

[1] *Section de Mathématiques, Université de Genève, Switzerland*

Abstract. *The XMM-Newton satellite has three telescopes taking images of the universe at three sensitivity levels. We propose to aggregate the information of the three telescopes by block total variation denoising to deblur and smooth the data. We give a technical description of XMM-Newton spacecraft, its telescope, data and specificities, describe the statistical model and discuss the block total variation estimator we are developping.*

Keywords. *Total variation; Inverse problem; Point source detection; Smoothing; Block lasso; Telescope*

3.56

Singular additive models for function to function regression

Hans-Georg Müller[1], Byeong U. Park[2,*] and Wenwen Tao[1]

[1] *University of California at Davis; hgmueller@ucdavis.ed, wtao@ucdavis.edu*
[2] *Seoul National University; bupark@stats.snu.ac.kr*
[*] *Corresponding author*

Abstract. *In various functional regression settings one observes i.i.d. samples of paired stochastic processes* (X, Y)*, and is interested in predicting the trajectory of* Y*, given the trajectory* X*. For example, one may wish to predict the future of a process from observing an initial segment of the trajectory. Commonly used functional regression models are based on representations that are obtained separately for* X *and* Y*. In contrast to these established methods, we base our approach a on a singular expansion of the paired processes* X, Y *with singular functions that are derived from the cross-covariance surface between* X *and* Y*. The motivation for this approach is that the resulting singular components are tuned towards reflecting the association between* X *and* Y*. The regression relationship is then based on the assumption that the singular components of* Y *follow an additive regression model with the singular components of* X *as predictors. The resulting singular additive model is fitted by smooth backfitting. We will discuss asymptotic properties of the estimates as well as their practical behavior in simulations and data analysis.*

Keywords. *Singular component analysis; Functional data; Additive models; Hilbert-Schmidt operators; kernel smoothing.*

3.57

Detecting smooth changes in locally stationary processes

M. Vogt[1,*] and H. Dette[2]

[1] *University of Konstanz, Department of Mathematics and Statistics, 78457 Konstanz, Germany; michael.vogt@uni-konstanz.de*
[2] *Ruhr-Universität Bochum, Department of Mathematics, 44780 Bochum, Germany; holger.dette@ruhr-uni-bochum.de*
[*] *Corresponding author*

Abstract. *In a wide range of time series applications, the stochastic properties of the observed process change over time. The properties can often be expected to be approximately the same for some time before they start to vary. In such situations, it is frequently of interest to locate the time point where the properties start to change. In this paper, we construct a procedure to estimate this time point. We set up a general method which allows to deal with a wide variety of stochastic properties including the mean, covariances and higher moments of the time series under consideration. In the theoretical part of the talk, we derive the asymptotic properties of our method. In addition, we illustrate the methodology by applications to temperature and financial return data.*

Keywords. *Local Stationarity; Empirical processes; Measures of time-variation.*

References

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* **25**, 1–37.

Mallik, A., Sen, B., Banerjee, M. and Michailidis, G. (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika* **98**, 887–900.

Dette, H., Preuß, P. and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. *Journal of the American Statistical Association* **106**, 1113–1124.

3.58

Optimal estimation of components in structured nonparametric models

Martin Wahl

Department of Economics, University of Mannheim, L7, 3-5, 68131, Germany; mawahl@mail.uni-mannheim.de

Abstract. *We consider the nonparametric random regression model $Y = f_1(X_1)+f_2(X_2)+\epsilon$ in which the function f_1 is the parameter of interest and the function f_2 is a nuisance parameter. We present a theory of estimating f_1 in settings where the second part is more complex than the first part. The proposed estimation procedure is based on the composition of two least squares criteria and can be written as an alternating projection procedure. Our main results are nonasymptotic risk bounds which reveal connections between the performance of our estimators of f_1 and the notions of minimal angles and Hilbert-Schmidt operators in the theory of Hilbert spaces. We show that, under additional regularity conditions on the design densities, these bounds can be further improved. As a consequence of these results, we find general assumptions under which the estimators of f_1 have up to first order the same sharp upper bound as the corresponding estimators in the model $Y = f_1(X_1) + \epsilon$. As an example we apply the theory to an additive model where the number of covariates is large or the nuisance components are considerably less smooth than f_1.*

Keywords. *Structured nonparametric models; Projection on sumspaces; Alternating projections; Additive models; Increasing number of covariates.*

3.59

A new nonparametric stationarity test of time series in time domain

Lei Jin[1], Suojin Wang[2,*] and Haiyan Wang[3]

[1] *Department of Mathematics & Statistics Texas A&M University-Corpus Christi, Corpus Christi, Texas 78412, USA; lei.jin@tamucc.edu*
[2] *Department of Statistics, Texas A&M University, College Station, Texas 77843, USA; sjwang@stat.tamu.edu*
[3] *Department of Statistics, Kansas State University, Manhattan, Kansas 66506, USA; hwang@k-state.edu*
[*] *Corresponding author*

Abstract. *In this talk, we present a new double order selection test for checking second-order stationarity of a time series. To develop the test, a sequence of systematic samples are defined via the Walsh functions. Then the deviations of the autocovariances based on these systematic samples from the corresponding autocovariances of the whole time series are calculated and the uniformly asymptotic joint normality of these deviations over different systematic samples is obtained. With a double order selection scheme, our test statistic is constructed by combining the deviations at different lags in the systematic samples. The null asymptotic distribution of the proposed statistic is derived and the consistency of the test is shown under fixed and local alternatives. Simulation studies demonstrate well-behaved finite sample properties of the proposed method. Comparisons with the test of Dette et al. (2011) in terms of power are given both analytically and empirically. In addition, the proposed method is applied to check the stationarity assumption of a chemical process viscosity readings data.*

Keywords. *Order selection; Stationarity test; Systematic samples; Time series; Walsh functions*

References

Dette, H., Preuß, P. and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. *Journal of the American Statistical Associaiton*, **106**, 1113–1124.

3.60

Short-term load forecasting: the similar shape functional time series predictor

Efstathios Paparoditis[1] and Theofanis Sapatinas[1]*

[1] *Department of Mathematics and Statistics, University of Cyprus, Cyprus; stathisp@ucy.ac.cy; fanis@ucy.ac.cy*
* *Corresponding author*

Abstract. *A novel functional time series methodology for short-term load forecasting is introduced. The prediction is performed by means of a weighted average of past daily load segments, the shape of which is similar to the expected shape of the load segment to be predicted. The past load segments are identified from the available history of the observed load segments by means of their closeness to a so-called reference load segment. The later is selected in a manner that captures the expected qualitative and quantitative characteristics of the load segment to be predicted. As an illustration, the suggested functional time series forecasting methodology is applied to historical daily load data in Cyprus. Its performance is compared to some recently proposed alternative methodologies for short-term load forecasting.*

Keywords. *Functional Kernel Regression; Short-Term Load Forecasting; Time Series, Wavelets.*

References

Paparoditis, E. and Sapatinas, T. (2013). Short-term load forecasting: the similar shape functional time series predictor. *IEEE Transactions on Power Systems*, **28**, 3818–3825.

3.61

Functional lagged regression

S. Hörmann[1,*], Ł. Kidziński[1] and P. Kokoszka[3]

[1] *Department of Mathematics, Université libre de Bruxelles; shormann@ulb.ac.be, lkidzins@ulb.ac.be*
[3] *Department of Statistics, Colorado State University; Piotr.Kokoszka@colostate.edu*
* *Corresponding author*

Abstract. *Consider a sequence (Y_k) of real responses, explanatory variables (X_k) taking values in some function space $\mathcal{F}$, iid noise (ε_k) which independent of (X_k), and a linear operator $\beta : \mathcal{F} \to \mathbb{R}$. The functional linear model $Y_k = \beta(X_k) + \varepsilon_k$ has received a great deal of attention over the last two decades. As in the usual linear regression, to derive inferential properties, the assumption imposed on the above model is that the pairs (Y_k, X_k) are independent and identically distributed. While this assumption is well justified in designed experiments, it may be called into question when the functions X_k form a functional time series.*

The objective of this talk is to develop estimation and testing methodology, and the underlying asymptotic theory, for the model

$$Y_\ell = a + \sum_{k \in \mathbb{Z}} b_k(X_{\ell-k}) + \varepsilon_\ell, \quad b_k : \mathcal{F} \to \mathbb{R}, \quad a \in \mathbb{R}. \tag{1}$$

Model (1) *is an extension of the* lagged regression model, *which is the most commonly used regression model in time series analysis.*

As with most functional procedures, the main challenge is a suitable dimension reduction technique and the need to deal with unbounded operators, difficulties not encountered in the scalar and vector theory.

Keywords. *Frequency domain methods; Functional regression; Functional time series*

3.62

Variable selection and dimension reduction criteria in functional regression models

S. Fremdt[1]

[1] *Ruhr-University Bochum, Department of Mathematics, Institute of Statistics, 44780 Bochum; stefan.fremdt@rub.de*

Abstract. *While variants of the LASSO regularization techniques have been introduced to select significant predictors in functional regression with scalar responses, most articles dealing with fully functional regressions have incorporated only a single predictor, and possibly its derivatives, into the model building process. To the best of our knowledge, no variable selection methods are currently available in the literature for this case. We consider approaches to select significant functional predictors for functional responses by obtaining an auxiliary multivariate linear model of functional principal component scores. In particular we discuss the construction of criteria which determine the significant predictors as well as the dimensionality of both response and predictor score vectors.redictor score vectors are determined through a novel automatic penalization criterion.*

Keywords. *Dimension reduction; Functional data analysis; Functional principal components; Multivariate linear models; Prediction error.*

3.63

Bootstrap based testing for functional data

Efstathios Paparoditis[1]* and Theofanis Sapatinas[1]

[1] *Department of Mathematics and Statistics, University of Cyprus, Cyprus;*
stathisp@ucy.ac.cy; fanis@ucy.ac.cy
* *Corresponding author*

Abstract. *We propose a bootstrap procedure to test hypotheses about the mean functions and/or the covariance operators for functional data. Our approach is simple and resamples the original data set of functional observations in such a way that the null hypothesis of interest is satisfied. It can be applied to a wide range of test statistics and also to the case where more than two groups of functional data are considered. The validity of the bootstrap-based testing procedures proposed for some commonly used test statistics in the literature are established. Simulation results demonstrate a very good performance of our bootstrap proposal in finite sample situations and also for testing problems and sample sizes where the asymptotic theory does not approximate appropriately the behavior of the test statistics considered. A real-life data set is also analyzed to illustrate the suggested bootstrap based methodology.*

Keywords. *Functional Data; Comparison of Mean Functions; Covariance Operators; Bootstrap.*

3.64

Sparse and functional principal components analysis

Genevera I. Allen

Departments of Statistics and Electrical and Computer Engineering, Rice University, & Jan and Dan Duncan Neurological Research Institute, Baylor College of Medicine and Texas Children's Hospital

Abstract. *Regularized principal components analysis, especially Sparse PCA and Functional PCA, has become widely used for dimension reduction in high-dimensional settings. Many examples of massive data, however, may benefit from estimating both sparse AND functional factors. These include neuroimaging data where there are discrete brain regions of activation (sparsity) but these regions tend to be smooth spatially (smoothness). Here, we introduce an optimization framework that can encourage both sparsity and smoothness of the row and/or column PCA factors. This framework generalizes many of the existing approaches to Sparse PCA, Functional PCA and two-way Sparse PCA and Functional PCA, as these are all special cases of our method. In particular, our method permits flexible combinations of sparsity and smoothness that lead to improvements in feature selection and signal recovery as well as more*

3.65

On estimation efficiency of the dimension reduction space

Yanyuan Ma[1] and Liping Zhu[2,*]

[1] *Department of Statistics, Texas A&M University; ma@stat.tamu.edu*
[2] *School of Statistics, Shanghai University of Finance and Economics; lzhu1@hotmail.com*

Abstract. *We investigate the estimation efficiency of the dimension reduction subspace in the framework of sufficient dimension reduction. We derive the semiparametric efficient score and study its practical applicability. Despite the difficulty caused by the potential high dimension issue in the variance component, we show that locally efficient estimators can be constructed in practice. We conduct simulation studies and a real-data analysis to demonstrate the finite sample performance and efficiency gain of the proposed estimators in comparison with several existing methods.*

Keywords. *Dimension reduction; Estimating equations; Nonparametric regression; Semiparametric efficiency; Sliced inverse regression.*

3.66

Conditional mean absolute deviation

T. Zhou[1] and L. Zhu[2,*]

Shanghai University of Finance and Economics; zhu.liping@mail.shufe.edu.cn.

Abstract. *We introduce conditional median absolute deviation to characterize how the local variability of one quantitative random variable varies with another one. A two-step estimation procedure is proposed and the resultant estimator possesses an adaptiveness property. Simulation indicates that this estimator is much more efficient than its competitors such as the conditional semi-interquartile range.*

Keywords. *Adaptiveness; least absolute deviation; median absolute deviation; quantile regression; robust estimation; semi-interquartile range.*

3.67

Estimation and testing of varying coefficients in quantile regression

Xingdong Feng[1], Liping Zhu[1,*]

[1] School of Statistics and Management, Shanghai University of Finance and Economics (SUFE), Shanghai 200433, China; feng.xingdong@mail.shufe.edu.cn, zhu.liping@mail.shufe.edu.cn
[*] Corresponding author

Abstract. *In this paper, we establish a novel connection between the null hypothesis* H_0 : $\mathbf{C}^T\boldsymbol{\beta}_0(t) = \mathbf{c}_0$ *and a rank-reducible varying coefficient models in quantile regression. We use* b*-spline to approximate the varying coefficients in the rank-reducible model, and reveal that the null hypothesis* H_0 *implies a unidimensional structure of a transformed coefficient matrix of* b*-spline bases. By evaluating the unidimensional structure, we alleviate the difficulty of testing such hypotheses commonly considered in varying coefficient models. We demonstrate through comprehensive numerical studies that the new method is much more powerful than the rank score test which is widely used in quantile regression literature.*

Keywords. *Dimension reduction; Hypothesis test; Quantile regression; Singular value decomposition.*

3.68

Quantile regression with cure rate model

Yuanshan Wu[1] and Guosheng Yin[2,*]

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China; shan@whu.edu.cn
[2] Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong; gyin@hku.hk
[*] Corresponding author

Abstract. *Censored quantile regression offers a valuable complement to the traditional Cox proportional hazards model for survival analysis. Survival times tend to be right-skewed, particularly when there exists a substantial fraction of long-term survivors who are either cured or immune to the event of interest. For survival data with a cure possibility, we propose cure rate quantile regression under the common censoring scheme that survival times and censoring times are conditionally independent given the covariates. In a mixture formulation, we apply censored quantile regression to model the survival times of susceptible subjects and logistic regression to model the indicators of whether patients are susceptible. We develop estimation methods using martingale-based equations, and also we discuss the possibility of using multiple imputation. We*

3.69

Bias correction using the SIMEX algorithm in the promotion time cure model with measurement error

A. Bertrand[1,*], R.J. Carroll[2], C. Legrand[1] and I. Van Keilegom[1]

[1] *Université catholique de Louvain, Louvain-la-Neuve, Belgium; aurelie.bertrand@uclouvain.be, catherine.legrand@uclouvain.be, ingrid.vankeilegom@uclouvain.be*
[2] *Texas A&M University, College Station, USA; carroll@stat.tamu.edu*
[*] *Corresponding author*

Abstract. *In many situations in survival analysis, it may happen that a fraction of individuals will never experience the event of interest : they are considered to be cured. The promotion time cure model is one of the survival models taking this feature into account.*
We consider the case where one or more explanatory variables in the model are subject to measurement error. This error should be taken into account in the estimation of the model, to avoid biased estimators of the model.
In the literature, several approaches to correct this bias have been proposed. The SIMEX algorithm is one of them: it is a method based on simulations which allows to estimate the effect of measurement error on the bias of the estimators and to reduce this bias. It has already been applied to many different models, but not to the promotion time cure model. For this model, Ma and Yin (2008) have suggested a corrected score approach.
We extend the SIMEX approach to the promotion time cure model. We show that the proposed estimator is asymptotically normally distributed. We also show via simulations that the suggested method performs well in practice by comparing it with the method proposed by Ma and Yin (2008), which is, as far as we know, the only paper that has studied this problem before in the literature. Finally, we analyze a database in cardiology: among the explanatory variables of interest is the ejection fraction, which is very likely to be measured with error.

Keywords. *Cure model; measurement error; SIMEX.*

References

Ma, Y., Yin, G. (2008). Cure Rate Model With Mismeasured Covariates Under Transformation. *Journal of the American Statistical Association* **103**, 743–756.

3.70

Dynamic cure models

Alex Tsodikov[1]

[1] University of Michigan; tsodikov@umich.edu

Abstract. *Usually cure models are induced by a frailty random variable with a mass at zero. A popular model assumes binary frailty variable effectively splitting the population into cured and non-cured subjects once and for all, at time zero. However, in situations when time-dependent covariates are present, the assumption that they only apply to the survival of the uncured subjects is too restrictive. For example, it would be natural to assume that treatment applied over time modifies the chance of cure for the subject. More broadly, the event of cure may be an outcome of a random process over time. We discuss some modeling and estimation approaches to the problem and illustrate them the using examples of cancer studies.*

Keywords. *Cure models; Semiparametric models; Dynamic frailty*

3.71

Bootstrap inference in the promotion time cure model

Anouar El Ghouch, François Portier* and Ingrid Van Keilegom

Institut de statistique, biostatistique et sciences actuarielles, Louvain-la-Neuve, Belgique
anouar.elghouch@uclouvain.be francois.portier@uclouvain.be ingrid.vankeilegom@uclouvain.be
* *Corresponding author*

Abstract. *During this talk, we shall focus on the asymptotic properties of the nonparametric maximum likelihood estimator (NPMLE) in the promotion time cure model. Since this model has some serious links with the Cox model, our study mimics approaches that have been developed in the case of the Cox model. First, we show that the NPMLE may be computed by a single maximisation over a set whose dimension equals the dimension of the covariates plus 1. Second, we derive the asymptotics by using a Z-estimator theorem for infinite dimensional parameter. In particular, we express simply the asymptotic variance of the finite dimensional parameter relying on profile likelihood. Since the variance of the NPMLE is difficult to estimate, we develop a general bootstrap strategy that allows for a consistent approximation of the asymptotic law. As in the Cox model, it turns out that suitable tools are the martingale theory for counting processes and the semiparametric efficiency theory. Finally, by means of simulations, we show the accuracy of the bootstrap with respect to the normal approximation.*

Keywords. *Promotion time cure model; Cox model; Asymptotic inference; Bootstrap; Semiparametric efficiency.*

3.72

Extreme value statistics for a stochastic process that is observed at discrete points only

H. Drees1[1], L de Haan[2,*] and K. F. Turkman[3]

[1] *University of Hamburg, Germany; holger.drees@math.uni-hamburg.de*
[2] *Erasmus University Rotterdam, The Netherlands and CEAUL, Portugal; ldehaan@ese.eur.nl*
[3] *University of Lisbon and CEAUL, Portugal; fturkman@fc.ul.pt*
[*] *Corresponding author*

Abstract. *When dealing with spatial extremes (extreme heat, rainfall, wind speed) the proper tool is a max-stable process i.e. the infinite-dimensional extension of extreme value theory. The main assumption then is that the underlying stochastic process is in the domain of attraction of a max-stable process. On the basis of this assumption the main features of the limiting max-stable process can be estimated. Asymptotic properties of the estimators have been derived (Lin e.a. Ann. Statist. 2003, 2006) under assumption that the entire underlying process can be observed. If the process is observed only at a limited number of points one can merely fit a max-stable process depending on finitely many parameters. We show that the general non-parametric case can be recovered if the observation points are close together.*

Keywords. *max-stable processes*

3.73

The block maxima method revisited applied to PWM estimators

A. Ferreira[1,*] and L. de Haan[2]

[1] *ISA and CEAUL, University of Lisbon, Portugal; anafh@isa.utl.pt*
[2] *Erasmus University Rotterdam, The Netherlands and CEAUL, Portugal; ldehaan@ese.eur.nl*
[*] *Corresponding author*

Abstract. *The block maxima method is a fundamental approach in Statistics of Extremes, where inferences are based on the 'k sample maxima' after dividing the sample into k blocks. We provide conditions under which this method can be justified. We restrict attention to the independent and identically distributed case and focus on the probability weighted moment (PWM) estimators of Hosking, Wallis and Wood (1985).*

Further, we present a theoretical comparison between the peaks-over-threshold and the block maxima methods.

Keywords. *Block maxima; Probability weighted moment estimators; Peaks-over-threshold.*

References

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985) Estimation of the Generalized Extreme-Value Distribution by the Method of Probability Weighted Moments. *Technometrics* **27**, 251–261.

3.74

Statistics of heteroscedastic extremes

John H.J. Einmahl[1,*], Laurens de Haan[2] and Chen Zhou[3]

[1] *Tilburg University; j.h.j.einmahl@uvt.nl*
[2] *Erasmus University Rotterdam; ldehaan@ese.eur.nl*
[3] *De Nederlandsche Bank; c.zhou@dnb.nl*
[*] *Corresponding author*

Abstract. *We extend classical extreme value theory to non-identically distributed observations. When the distribution tails are proportional much of extreme value statistics remains valid. The proportionality function for the tails can be estimated nonparametrically along with the (common) extreme value index. Joint asymptotic normality of both estimators is shown; they are asymptotically independent. We develop tests for the proportionality function and for the validity of the model. We show through simulations the good performance of tests for tail homoscedasticity. The results are applied to stock market returns. A main tool is the weak convergence of a weighted sequential tail empirical process.*

Keywords. *Extreme value statistics; Functional limit theorems; Non-identical distributions; Scale; Sequential tail empirical process.*

3.75

Conditional inference for territorial comparisons on the perception of odours

S. Bonnini

Department of Economics and Management, University of Ferrara (Italy); stefano.bonnini@unife.it

Abstract. *For comparing environmental odor perceptions in two different geographical areas, a conditional and data driven testing procedure is proposed. Data are related to environmental odors perceived by groups of sniffers. The response variable is nominal categorical (e.g. type of perceived odor) and the goal of the study consists in studying the mutability or, in other words, the diversity of the distributions, and testing whether the mutability of odor perceptions in one area is greater than the mutability of odor perceptions in another area. Such procedure is based on the comparison of the Pareto Diagrams of the observed frequency distributions and on* *approximate exchangeability of data under the null hypothesis, which allows the application of a permutation testing procedure similar to the permutation test for stochastic dominance (see (Arboretti, 2009)). The exchangeability is approximate because the true order of the unknown probabilities of the categorical distributions (necessary for the Pareto Diagram) is estimated with the observed frequencies (data driven ordering) and thus exchangeability is exact only asymptotically. Simulation studies prove that the test is well approximated and powerful. An application example is discussed.*

Keywords. *Conditional Inference; Nonparametric Test; Odor Perception; Mutability.*

References

Arboretti, G. R., Bonnini, S., Pesarin, F. (2009). A permutation approach for testing heterogeneity in two-sample problems. *Statitics and Computing* **19**, 2, 209–216.

3.76

A permutation approach for ranking of multivariate populations

Livio Corain1[1,*], Luigi Salmaso[1]

[1] *Department of Management and Engineering, University of Padova, Italy; livio.corain@unipd.it, luigi.salmaso@unipd.it*
[*] *Corresponding author*

Abstract. *The need to establish the relative superiority of each treatment/group when compared to all the others, that is ordering the effects with respect to the underlying populations, often occurs in many multivariate studies when there might be a "natural ordering" in which the responses are interpreted as "the higher the better" or "the lower the better". Within the framework of multivariate stochastic ordering (Pesarin and Salmaso, 2010), the purpose of this work is to propose a nonparametric permutation-based solution for the problem of ranking of multivariate populations, i.e. estimating an ordering related to the possible stochastic dominance among several unknown multivariate distributions. The method is metric-free in the sense that it can be applied to any kind of response variables, i.e. continuous/binary or ordered categorical or mixed (some continuous/binary univariate components and some other ordered categorical), and it is valid also in case the sample sizes are lower than the number of responses. It will be theoretically argue and numerically proved that our method controls the risk of false ranking classification under the null hypothesis of population homogeneity while under the alternatives we expect that the true rank can be estimated with satisfactory accuracy, especially for the 'best' and the 'worst' populations. Finally, to highlight the practical relevance of the proposed methodology, some real case studies are presented.*

Keywords. *Multivariate tests, Nonparametric combination, Pairwise comparisons, Permutation tests.*

References

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

3.77

Projection, residuals and permutation for comparing factors in complex designs and signals

O. Renaud[1,*], S. Kherad-Pajouh[2]

[1] *University of Geneva, Switzerland; Olivier.Renaud@unige.ch*
[2] *University of California, Berkeley, CA; kherad@berkeley.edu*
[*] *Corresponding author*

Abstract. *Permutation tests are well known for simple experiments, like for the comparison of two groups. In this talk, we will present a general principle using a projection that allows for the generalization of permutation tests to many situations, including all experimental designs (or ANOVA designs). We will also present the geometrical view of this principle and the similarities with the idea behind the restricted maximum likelihood (REML). We show the conditions to obtain an exact test and discuss the choice of the statistic. This principle can be applied to ANOVA designs with fixed effects only as well as to so-called mixed ANOVA designs, i.e. designs with fixed and random effects. It allows us to test any factor of the design without any restriction, including for example testing a main effect even if an interaction containing the corresponding factor has a large effect. Our approach can also be generalized to the comparison of signals obtained in different experimental conditions, providing tests that are simultaneous at all time points. In the case of signals, we will compare our method to TANOVA, which is a popular non-parametric method in neuroscience and to a method based on Gaussian random fields.*

Keywords. *ANOVA designs; Reduced residuals; exchangeability*

3.78

Testing model structure in high dimensional nonparametric regression

F. Giordano[1], S.N. Lahiri[2] and M.L. Parrella[1,*]

[1] *Department of Economics and Statistics - University of Salerno - Italy; giordano@unisa.it, mparrella@unisa.it*
[2] *Department of Statistics - North Carolina State University - USA; snlahiri@ncsu.edu*
[*] *Corresponding author*

Abstract. *In the context of nonparametric regression, we assume that the number of covariates tends to infinity but only some of these covariates are relevant for the model. The aim is to identify the relevant covariates and to obtain some information about the structure of the model.*

A new nonparametric procedure is proposed, called GRID (Gradient Relevant Identification Derivatives), in order to make variable selection and model structure discovering. The key idea is to use a modified local linear estimator in a non standard way together with Empirical Likelihood method to make variable selection without any additivity assumption. Under some regularity conditions, GRID automatically identifies the relevant covariates of the regression model, also distinguishing the nonlinear from the linear ones (a covariate is defined linear or nonlinear depending on the marginal relation between the response variable and such a covariate). Besides, the interactions between the covariates are automatically identified, without the necessity of considering some kind of stepwise selection method. In particular, GRID can identify the mixed terms of any order (two way, three way, ...) without increasing the computational complexity of the algorithm.

The GRID procedure is completely data-driven, so it is easily implementable for the analysis of real datasets. This makes the procedure appealing compared to its competitors, which are generally dependent on crucial regularization parameters. Moreover, an intuitive graphical tool, the GRID-plot, is proposed in order to make the output user-friendly.

The theoretical properties of the method have been investigates both theoretically and computationally. A simulation study compares the performance of the proposed method with the most direct competitors.

Keywords. *Variable selection; Model selection; Nonparametric regression; Empirical likelihood; High dimension.*

3.79

Novel methods for modelling multiple ranked lists

Michael G. Schimek[1,*], Peter Hall[2] 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*
[2] *The University of Melbourne, Australia; halpstat@ms.unimelb.edu.au*
[*] *Corresponding author*

Abstract. *In recent years there has been an increasing interest in the statistics of ranked lists, primarily stimulated by new Web applications, business intelligence, and biotechnologies. Typically, such lists comprise tens of thousands of objects (e.g. URLs representing locations of profile pages in rank order). However, only a comparably small subset of k top-ranked elements is informative and useful. These objects are characterized by a strong overlap of their rank positions when they are ranked by different instances of assessment (e.g. by different Web search engines). A central task is the estimation of an overall k^* for a number of ranked lists comprising the same set of elements, before one can fit a consolidated data model to the obtained sublists. For pairs of ranked lists, the estimation of k has already been addressed by Hall and Schimek (2012). Most recently, an exploratory approach with a graphical representation of the consolidated objets has been developed and implemented in the* `R` *package* `TopKLists`. *A* *computationally highly demanding task is the fitting of models to multiple lists, especially for those high-dimensional applications we have in mind. Conventional model-based approaches (e.g. Fligner and Verducci (1988)) are not practicable because the number of rankings is in our situation rather small compared to the lengths of such ranked lists. In this presentation we outline a most general simulation-based concept for model fitting of multiple lists.*

Keywords. *Nonparametric inference; Model fitting; R package; Simulation; Top-k ranked lists.*

References

Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *J. Amer. Statist. Assoc.* **83**, 892–901.

Hall, P. and Schimek, M. G. (2012). Moderate deviation-based inference for random degeneration in paired rank lists. *J. Amer. Statist. Assoc.*, **107**, 661–672.

3.80

Asymptotics for in-sample density forecasting

Young K. Lee[1,*], Enno Mammen[2], Jens P. Nielsen[3] and Byeong U. Park[4]

[1] *Kangwon National University, Korea; youngklee@kangwon.ac.kr*
[2] *Universität Mannheim, Germany; emammen@rumms.uni-mannheim.de*
[3] *Cass Business School, City University, UK; Jens.Nielsen.1@city.ac.uk*
[4] *Seoul National University, Korea; bupark2000@gmail.com*
[*] *Corresponding author*

Abstract. *This paper generalizes recent proposals of density forecasting models and it develops theory for this class of models. In density forecasting the density of observations is estimated in regions where the density is not observed. Identification of the density in such regions is guaranteed by structural assumptions on the density that allows exact extrapolation. In this paper the structural assumption is made that the density is a product of one-dimensional functions. The theory is quite general in assuming the shape of the region where the density is observed. Such models naturally arise when the time point of an observation can be written as the sum of two terms (e.g. onset and incubation period of a disease). The developed theory also allows for a multiplicative factor of seasonal effects. Seasonal effects are present in many actuarial, biostatistical, econometric and statstical studies. Smoothing estimators are proposed that are based on backfitting. Full asymptotic theory is derived for them. A practical example from the insurance business is given producing a within year budget of reported insurance claims. A small sample study supports the theoretical results.*

Keywords. *Density estimation; Kernel smoothing; Backfitting; Chain Ladder.*

3.81

Geometrically designed spline smoothing with applications in copula estimation

V.K. Kaishev[1,*]

[1] *Cass Business School, City University London; v.kaishev@city.ac.uk*
[*] *Corresponding author*

Abstract. *We present a new method of constructing what we call Geometrically Designed least squares splines (GeDS) with variable knots. It utilizes a novel geometric interpretation of the estimation of the spline parameters. The latter is based on, shape preserving properties of spline functions, combined with a data driven phase of recovering the underlying control polygon of the spline. The method produces simultaneously linear, quadratic, cubic (and possibly higher order) least squares spline fits with one and the same number of B-spline coefficients. Small/large sample properties of the GeDS estimator are explored. We demonstrate how the method is applied in the context of multivariate Archimedean copula estimation. The GeDS estimation procedure is further illustrated numerically, based on simulated and real data examples from actuarial modelling and materials science.*

This talk is based on joint work with D.S. Dimitrova, S. Haberman, R. Verrall, and S.I. Penev.

Keywords. *B-splines; Variable knot spline regression; Greville abscissae; Control polygon; Archimedean copulas.*

3.82

Nonparametric forecasting of the French load curve

V. Lefieux

RTE and UPMC-ISUP; vincent.lefieux@rte-france.com

Abstract.
RTE, the French electricity transmission system operator, is responsible for operating, maintaining and developing the high and extra high voltage network. RTE is required to guarantee the security of supply, so anticipating French electricity demand helps to ensure the balance between generation and consumption at all times, and directly influences the reliability of the power system.

Demand forecasts are carried out for several different timeframes: for the long-term, in the form of the Generation Adequacy Report or network development studies, for the medium-term (annual, monthly and weekly forecasts) and lastly on a day-ahead basis.

From a short term point of view, RTE uses a complex nonlinear parametric regression model

with around one thousand coefficients estimated twice a year, and also a SARIMA model. If this process currently provides good forecasts, the context of the smart grids and the energy transition will lead to more variability in the load curve.
In order to obtain an adaptive forecasting model, nonparametric methods have already been tested without real success. We have used alternative methods (non or semiparametric) against the curse of dimensionality. In this talk, we present two different methods applied to the French electricity consumption: the IBR (Iterative Bias Reduction) method which iteratively corrects the bias initial estimator by an estimate of the latter obtained by smoothing the residuals, and a method based upon sparse approximations of the signals with a dictionary of functions (using LOLA algorithm).

Keywords. *Nonparametric forecasting; Iterative bias reduction; Sparse regression*

References

Mougeot, M. and Picard, D. and Tribouley, K. and Lefieux, V and Maillard-Teyssier, L (2013). Sparse approximation and fit of intraday load curves in a high dimensional framework. *Advanced in Adaptive Data Analysis* **5**.

Cornillon, P.A. and Hengartner, N. and Jégou, N. and Matzner-Løber, E. (2014). Recursive Bias Estimation for multivariate regression *ESAIM* **36**, 326–329.

3.83

Statistical models for electricity load forecasting

Yannig Goude [1]

[1] *EDF R&D*

Abstract. *Electricity load forecasting faces rising challenges due to the advent of innovating technologies such as smart grids, electric cars and renewable energy production. For utilities, a good knowledge of the future electricity consumption stands as a central point for the reliability of the network, investment strategies, energy trading, optimizing the production etc. Many statistical models have been investigated recently at EDF (Electricité de France) to forecast electricity consumption at different geographical scale and at temporal horizon. Among them stand regression on functional data, additive models, kernel based methods and random forest regression. We will dress a panorama of them on real data and show how a combining approach could improve the forecasting accuracy of these models.*

3.84

Automatic component selection in additive modelling of electricity load through the use of penalized regression methods

N1. Antoniadis[1], N2. Brossat[2], N3. Goude[2], N4. Poggi[3] and N5. Thouvenot[2,3,*]

[1] *Univ Joseph Fourier, Grenoble, France; Anestis.Antoniadis@imag.fr,*
[2] *EDF, Clamart, France; xavier.brossat@edf.fr, yannig.goude@edf.fr, vincent.thouvenot@edf.fr,*
[3] *Univ Paris-Sud, Orsay, France; Jean-Michel.Poggi@math.u-psud.fr*
[*] *Corresponding author*

Abstract. *Because capacity in storing and discharging electricity is very limited and costly, the French energy company Électricité de France (EDF) needs efficient tools for forecasting the consumption of its customers. Moreover, in recent years we have seen the arrival of new technologies involving collection of data on local networks leading to the recording of a large amount of time series (few thousands to about ten thousands). In order to analyze and forecast such a huge amount of time series a key tool is to model electricity loads via regression involving a large amount of potentially influential covariables. Efficient implementation requires dimension reduction. In this work we therefore use additive regression models (Hastie and Tibshirani, 1990) for electricity load forecasting since they are known to perform dimension reduction and to also perform flexible modeling. Inspired by recent work on automatic selection and estimation of additive components (Antoniadis , Gijbels and Verhasselt , 2012) we adopt a combination of penalized regression methods for electricity load forecasting, using a variable selection via the Group LASSO (Yuan and Lin, 2006), which is known to be biased, and a regularization via P-splines (Eilers and Marx, 1996) which is known to produce efficient nonparametric estimates. We further discuss on how to automatically implement simultaneously these methods and illustrate the resulting methodology on simulated and some electricity real data provided by EDF.*

Keywords. *Additive Model; Electricity Load Forecasting; Group LASSO; Model Selection; Penalized Spline*

References

A. Antoniadis I. Gijbels and A. Verhasselt (2012). Variable Selection in Additive Models Using P-Splines. *Technometrics* **54 (4)**, 425–438.

Eilers and Marx (1996). Flexible smoothing with B-splines and penalties. *Journal of Machine Learning Research* **11 (2)**, 89–121.

Hastie and Tibshirani. (1990). *Generalized Additive Models.* New York: Chapman and Hall.

Yuan and Lin (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

3.85

PARMA time series for modeling and prediction of energy market data

Anna Dudek[1,*]

[1] *AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland; aedudek@agh.edu.pl*
[*] *Joint work with H. Hurd and W. Wójtowicz*

Abstract. *Periodic autoregressive-moving-average models (periodic ARMA, PARMA) are used to model nonstationary time series with periodic structure. They are similar to ARMA except the coefficients are periodic in time with a common period. They are widely applied in climatology, hydrology, meteorology and economics data. We present how to model electricity demand using PARMA models. The demand for energy, which determines the price of the energy, changes periodically during the day. Therefore, periodic extension of ARMA can simultaneously catch periodic structure of the data and reflect the strongest dependencies between observations, providing accurate forecast. We describe all the standard steps of the usual model fitting procedure e.g. identification, estimation and diagnostics. We use methods and procedures implemented by Dudek et al. (2013) in the R software.*

Keywords. *Energy market; PARMA time series; Periodically correlated sequences;*

References

Dudek, A. E., Hurd H., and Wójtowicz W. (2013). R package, perARMA: Package for Periodic Time Series Analysis, http://cran.r-project.org/web/packages/perARMA.

Dudek, A. E., Hurd H., and Wójtowicz W. (2014). PARMA models with applications in R - submitted.

3.86

Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator

R. Douc 1[1], E. Moulincs[2,*]

[1] *Institut Mines-Telecom; Telecom SudParis; randal.douc@telecom-sudparis.eu*
[2] *Telecom ParisTech; eric.moulines@telecom-paristech.fr*
[*] *Corresponding author*

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

Keywords. *Time series of counts; Observations drive models; Asymptotically weak-Feller Markov chains; Misspecified models*

References

Fokianos, K., Rahbek, A., Tjøstheim, D., 2009. Poisson autoregression. J. Am. Statist. Assoc. 104 (488), 1430–1439, with electronic supplementary materials available online.
http://dx.doi.org/10.1198/jasa.2009.tm08270

Fokianos, K., Tjøstheim, D., 2011. Log-linear poisson autoregression. J. of Multivariate Analysis 102 (3), 563–578.

Gray, R., 2009. Probability, Random Processes, and Ergodic Properties. Springer, London.

Hairer, M., Mattingly, J., 2006. Ergodicity of the 2d navier-stokes equations with degenerate stochastic forcings. Ann. Math. 164, 993–1032.

Henderson, S. G., Matteson, D., Woodard, D., 2011. Stationarity of generalized autoregressive moving average models. Electronic Journal of Statistics 5, 800–828.

Hernández-Lerma, O., Lasserre, J.-B., 2003. Markov chains and invariant probabilities. Vol. 211 of Progress in Mathematics. Birkhäuser Verlag, Basel.

Kedem, B., Fokianos, K., 2002. Regression models for time series analysis. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

Meyn, S. P., Tweedie, R. L., 1993. Markov Chains and Stochastic Stability. Springer, London.

Neumann, M. H., Nov 2011. Absolute regularity and ergodicity of Poisson count processes. Bernoulli 17 (4), 1268–1284.

3.87

Identifiability conditions for partially observed Markov chains

R. Douc[1,*], F. Roueff[2] and T. Sim[2]

[1] *Department CITI, CNRS UMR 5157, Telecom Sudparis, Evry, FRANCE ; randal.douc@telecom-sudparis.eu*
[2] *Department LTCI, CNRS UMR 5141, Telecom Paristech, Paris, FRANCE ;roueff@telecom-paristech.fr, sim@telecom-paristech.fr*
[*] *Corresponding author*

Abstract. *This paper deals with a parametrized family of partially-observed bivariate Markov chains. We establish that the limit of the normalized log-likelihood is maximized when the parameter belongs to the equivalence class of the true parameter, which is a key feature for obtaining consistency the Maximum Likelihood Estimators in well-specified models.*
The novel aspect of this work is that the result is based on the unicity of the invariant distribution of the Markov chain associated to the complete data, regardless its rate of convergence to the

equilibrium. This is in deep contrast with existing results on the identifiability problem in these models, which assume exponential separation of measures or geometric ergodicity. This is obtained in a general framework including both fully dominated or partially dominated models, and thus, the result may be applied to both Hidden Markov models or Observation-Driven times series.

Keywords. *Consistency; Ergodicity; Hidden Markov models; Maximum likelihood; Observation-driven models; Time series of counts.*

3.88

Quasi-likelihood for observation driven models

K. Fokianos[1]

[1] *University of Cyprus; fokianos@ucy.ac.cy*

Abstract. *Observation driven models are specified by the existence of a hidden process that determines the dynamic behavior of the observed time series. There are numerous examples of such models including GARCH models and models for time series of counts. After reviewing their properties we establish the asymptotic behavior of the quasi maximum likelihood estimator. Our contribution unifies several existing results. This is a joint work with R. Douc and E. Moulines.*

Keywords. *asymptotic normality; consistency; ergodicity; mispecification; stationarity*

4.01

Functional modelling and forecasting of electricity load

P. Raña[1,*], G. Anciros[1] and J. Vilar[1]

[1] *Department of Mathematics, Faculty of Computer Science, Universidade da Coruña, Spain; paula.rana@udc.es, ganeiros@udc.es, juan.vilar@udc.es.*
[*] *Corresponding author*

Abstract.

The aim of this paper is to predict the daily electricity load curves using functional techniques. An empirical comparative study among different methods used in the electrical context is reported.

Predictions of electricity demand in the Spanish electricity market are computed. Due to the *presence of outliers in the electricity demand curves time series, some techniques to detect outliers in functional data are applied in order to clean the data used to develop the predictions. The behaviour of the different days of the week shows weekly seasonality, principally due to the fact that in the weekend the electrical consumption decreases. For this reason it is necessary to estimate different models depending on the day of the week.*

The comparative study includes nonparametric and semiparametric functional models. In both cases, two kinds of response are considered. On the one hand, scalar response is used, taking a model for each hour of the day as in Vilar et al. (2012). On the other hand also functional response is considered as Aneiros et al. (2013), taking the entire day as a curve. The proposed methods extend the work of Vilar et al. (2012) introducing exogenous variables in semiparametric models and considering the case of functional response. Naïve and ARIMA methods are widely known and used in the electrical context and are also included in the comparison.

Keywords. *Time series forecasting; Functional data; Electricity market; Nonparametric models; Semiparametric models.*

References

Aneiros, G., Vilar, J.M., Cao, R. and Muñoz, A. (2013). Functional prediction for the residual demand in electricity spot markets. *IEEE Transactions on Power Systems* **28(4)**, 4201–4208.

Vilar, J.M., Cao, R. and Aneiros, G. (2012). Forecasting next–day electricity demand and price using nonparametric functional methods. *International Journal of Electrical Power and Energy Systems* **39**, 48–55.

4.02

Nonparametric functional prediction of the unabsorbed flux continuum in the Lyman-α forest of quasar spectra

M. Ciollaro[1,*], J. Cisewski[1], P. E. Freeman[1], C. R. Genovese[1], R. O'Connell[2], L. Wasserman[1]

[1] *Statistics Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh (PA), 15213 - United States; ciollaro@cmu.edu, cisewski@stat.cmu.edu, pfreeman@stat.cmu.edu, genovese@stat.cmu.edu, larry@stat.cmu.edu*

[2] *Department of Physics, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh (PA), 15213 - United States; rcoconne@andrew.cmu.edu*

[*] *Corresponding author*

Abstract. *We present a novel approach to the prediction of the unabsorbed flux continuum in the Lyman-α forest portion of the light spectra of high redshift quasars. The unabsorbed flux continuum in this particular portion of the light spectrum is unobservable because of the presence of light-absorbing neutral hydrogen clouds lying along the lines of sight between the observers and the quasars. However, a number of recent cosmological analyses rely on quantities that depend on the flux continuum before the absorption due to neutral hydrogen, thus making its prediction a crucial statistical challenge. We demonstrate how the nonparametric regression model for functional predictor and functional response proposed in Ferraty, Van Keilegom, Vieu (2012)* *provides a natural framework to interpret and solve the problem of predicting the unabsorbed flux continuum in the Lyman-α forest of high redshift quasar spectra using low redshift spectra (for which the unabsorbed flux continuum in the Lyman-α forest is instead observable). Our results suggest that the model has the potential to produce accurate predictions of the unabsorbed flux continuum in the Lyman-α forest of both simulated and real quasar spectra. The nonparametric functional regression approach thus represents an appealing alternative to other methods that are traditionally used by the astronomical community, such as PCA-based methods.*

Keywords. *Nonparametric functional regression; Functional data analysis; Prediction; Lyman-α forest; Quasar spectra.*

References

Busca, N. et al. (2013). Baryon Acoustic Oscillations in the Lyman-α forest of BOSS quasars. *Astron. Astrophys.* **552**.

Dawson, K. S. et al. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. *Astron. J.* **145**.

Eisenstein, D. J. (2005). Dark energy and cosmic sound. *New Astron. Rev.* **49**, 360–365.

Ferraty, F., Van Keilegom, I., Vieu, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.* **109**, 10–28.

Ferraty F., Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. *J. Nonparametr. Statist.* **16**, 111–125.

4.03

Testing for lack of fit in functional regression models

S. Maistre[1] and V. Patilea[1]

[1] *Crest-Ensai; samuel.maistre@ensai.fr, patilea@ensai.fr*

Abstract. *We consider regression models with a response variable taking values in a Hilbert space, of finite or infinite dimension, and hybrid covariates. That means there are two sets of regressors, one of finite dimension and a second one functional with values in a Hilbert space. The problem we address is the test on the effect of the functional covariates. This problem occurs in many situations: testing the effect of the functional covariate in a semi-functional partial linear regression with scalar responses, significance test for functional regressors in nonparametric regression with hybrid covariates and scalar or functional responses, testing the effect of a functional covariate on a scalar or functional outcome. We propose a new test based on univariate kernel smoothing. Inspired by Fan and Li (1996), the test statistic is asymptotically standard normal under the null hypothesis provided the smoothing parameter tends to zero at a suitable rate. The one-sided test is consistent against any fixed alternative and detects local alternatives a la Pitman approaching the null hypothesis at suitable rate. In particular we show that neither the dimension of the outcome nor the dimension of the functional covariates influences the theoretical power of the test against such local alternatives.*

Keywords. *Functional data, Lack-of-fit test, Regression, U-statistics .*

References

Fan, J. and Li, Q. (1996). Consistent Model Specification Tests : Omitted Variables and Semiparametric Functional Forms. *Econometrica* **64**, 865–890.

4.04

Hyperspectral image segmentation based on functional kernel density estimation

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

[1] *MAPMO, Université d'Orléans, UFR Sciences Bâtiment de mathématiques*
Rue de Chartres B.P. 6759 - 45067 Orléans cedex 2 FRANCE.;
laurent.delsol@univ-orleans.fr, cecile.louchet@univ-orleans.fr
[*] *Corresponding author*

Abstract. *Splitting a picture into a set of homogenous regions is a common problem, called segmentation, in image analysis. The detection of such regions is usually a relevant way to identify specific parts of the scene. Various methods have been proposed to segment gray-level or multispectral images. The maximum a posteriori approach, based on Potts random field as prior and density estimation on each region, is an interesting use of Bayesian statistics in that domain. On the other hand, a great variety of functional statistical methods are nowadays available to deal with data sets of curves. The kernel density estimator has been adapted to such data. In this talk we focus on hyperspectral images for which each pixel is described through a curve (discretized on a thin grid) and discuss the way functional kernel density estimation and maximum a posteriori approach may be combined.*

Keywords. *Functional data; Hyperspectral image; Segmentation; Kernel smoothing; Bayesian statistics.*

4.05

Comparision of multivariate distributions using depth-based quantile-quantile plots and related tests

Subhra Sankar Dhar

IIT Kanpur, India; subhra@iitk.ac.in

Abstract. *The univariate quantile-quantile (Q-Q) plot is a well-known graphical tool for exam-*

ining whether two data sets are generated from the same distribution or not. It is also used to determine how well a specified probability distribution fits a given sample. In this talk, we will develop and study a multivariate version of Q-Q plot based on spatial quantiles (see Chaudhuri (1996), JASA), which is based on the spatial depth in a certain sense. The usefulness of the proposed graphical device will be illustrated on different real and simulated data, some of which have fairly large dimensions. We will also develop certain statistical tests that are related to the proposed multivariate Q-Q plots and study their asymptotic properties. The performance of those tests compared to some other well-known tests for multivariate distributions will be discussed also. This is a joint work with Biman Chakraborty (University of Birmingham, UK) and Probal Chaudhuri (Indian Statistical Institute, Calcutta, India).

Keywords. *Characterization of distributions; Contiguous alternatives; Gaussian process; Pitman efficacy; Spatial quantiles; Tests for distributions; The level and the power of test.*

4.06

Elliptical quantiles and their generalizations

Daniel Hlubinka[1,*], Miroslav Šiman[2].

[1] *Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague; daniel.hlubinka@mff.cuni.cz*
[2] *Institute of Information Theory and Automation of the ASCR; siman@utia.cas.cz*
[*] *Corresponding author*

Abstract. *In our contribution, we are going to present some multivariate quantiles defined by means of quantile regression. We start with introducing a concept of elliptical quantiles in the convex optimization framework. Then we drop the convexity assumption, modify the loss function, and extend the definition to produce a whole class of (still) elliptical quantiles with some desirable features, and show how these quantiles can further be generalized to various parametric and nonparametric regression frameworks as well as to non-elliptical (and potentially non-convex) shapes. As for all the multivariate quantiles mentioned in our presentation, we will also discuss their properties, compare them to their competitors, describe their non-trivial computation, and point out the difficulties connected with both their use and analysis. We hope that our research provides a meaningful (regression) quantile concept for centrally symmetric (conditional) distributions when parametric approaches suffer from the lack of information and nonparametric methods cannot benefit from the apriori information regarding the symmetry.*

Keywords. *Multivariate quantile; Quantile regression; Elliptical quantile*

References

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

4.07

Shape depth

Germain Van Bever[1,*] and Davy Paindaveine[2].

[1] *Université libre de Bruxelles; gvbever@ulb.ac.be*
[2] *Université libre de Bruxelles; dpaindav@email*
[*] *Corresponding author*

Abstract. *In many problems from multivariate analysis (principal component analysis, testing for sphericity, etc.), the parameter of interest is not the scatter matrix but the so-called* shape *matrix, that is, a normalized version of the corresponding dispersion matrices. In this paper, we propose, under elliptical assumptions, a depth concept for shape. If shape matrices are normalized to have determinant one, our shape depth results from the parametric depth construction in Mizera (2002). For other normalizations, however, defining a proper shape depth requires a semiparametric extension of this construction, which is likely to have applications in other contexts. We show that the proposed shape depth does not depend on the normalization adopted and is affine-invariant. We also establish consistency, in the sense that shape depth is maximized at the true shape value. Finally, we consider depth-based tests for shape, and investigate their finite-sample performances through simulations.*

Keywords. *Elliptical distributions; Shape matrix; Statistical depth functions; Tangent depth; Tests for sphericity.*

References

Mizera, I. (2002). On depth and deep points: a calculus. *Ann. Statist.* **30** (6), 1681–1736.

4.08

Nonparametric confidence regions for the central orientation of random rotations

B. Stanfill[1], Ulrike Genschel[1,*] and Heike Hofmann[1]

[1] *Department of Statistics, Iowa State University; stanfill@iastate.edu, ulrike@iastate.edu, hofmann@iastate.edu*
[*] *Corresponding author*

Abstract. *Three-dimensional orientation data, with observations as 3×3 rotation matrices, have applications in many areas such as computer science, kinematics and materials sciences where it is often of interest to nonparametrically estimate a central orientation parameter* $\boldsymbol{S}$

represented by a 3×3 rotation matrix. A well-known estimator of this parameter is the projected arithmetic mean, and based on this statistic, two nonparametric methods for setting confidence regions for $\boldsymbol{S}$ exist. Both methods involve large-sample normal theory, with one approach based on a data-transformation of rotations to directions (namely, four-dimensional unit vectors in) prior to analysis. However, both of these nonparametric methods may result in poor coverage accuracy in small samples. As a remedy, we consider two bootstrap methods for approximating the sampling distribution of the projected mean statistic and calibrating nonparametric confidence regions for the central orientation parameter $\boldsymbol{S}$. As with normal approximations, one bootstrap method is based on a data-transformation of directions, and both bootstraps are shown to validly approximate the sampling distribution of the projected mean statistic. We then conduct a simulation study to explore and compare the performance of existing and newly developed confidence regions for $\boldsymbol{S}$, using popular data-generating models for symmetric orientations (Cayley, circular-von Mises and matrix-Fisher) with differing amounts of data concentration. Coverage rates corresponding to both bootstrap methods are close to the nominal level and provide an improvement over normal theory approximations, especially for small sample sizes. The bootstrap methods are illustrated with a real data example from materials science.

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

4.09

Local empirical likelihood for smooth coefficients panel data models

F. Bravo

[1] *University of York, York, YO10 5DD, United Kingdom; francesco.bravo@york.ac.uk*

Abstract. *This paper considers local empirical likelihood estimation and inference for semiparametric smooth coefficients dynamic panel data models. The paper derives the asymptotic distribution of the empirical likelihood estimator for the unknown smooth coefficients for both the small T and large N, and the large T and N cases. The paper also proposes an empirical likelihood ratio statistic for testing the smooth coefficients. Monte Carlo simulations illustrate the finite sample properties of the proposed method.*

Keywords. *Alpha-mixing; Endogeneity; Kernel estimation*

4.10

A nonparametric estimator and bootstrap confidence bands for the Kolmogorov canonical measure

G. Basulto[1,*], M. Nakamura[2] and V. Pérez-Abreu[2]

[1] *Iowa State University; basulto@iastate.edu*
[2] *Centro de Investigación en Matemáticas, México; nakamura@cimat.mx, pabreu@cimat.mx*
[*] *Speaker*

Abstract. *Levy processes (e.g., Brownian motion, compound Poisson processes, stable Levy processes) can be represented with a trend parameter and Kolmogorov canonical measure, which appear in the characteristic function. We propose a nonparametric estimator of this measure for a Levy process with finite second moment. A sieves-type approximation to the Kolmogorov canonical measure is considered, which depends on parameters that represent jumps. The estimator is the result of minimizing the distance between the empirical characteristic function and the characteristic function based on the parameters of the sieves-type approximation of the Kolmogorov canonical measure; both characteristic functions are evaluated at several points. Some representative examples are shown to illustrate the performance of this estimator as well as bootstrap confidence bands.*

Keywords. *Nonparametric; Kolmogorov-canonical-measure; Characteristic-function; Levy-process; Bootstrap-confidence-bands.*

4.11

Frequency domain empirical likelihood based tests of spatial structures

M. V. Hala[1], S. Bandyopadhyay[2,*] and D. J. Nordman[1]

[1] *Department of Statistics, Iowa State University, Ames, IA USA 50011; mvanhala@iastate.edu, dnordman@iastate.edu*
[2] *Department of Mathematics, Lehigh University, Bethlehem, PA USA 18015; sob210@lehigh.edu*
[*] *Corresponding author*

Abstract. *Modeling spatial data often relies on various assumptions, such as isotropy and separability of the covariance functions. These assumptions are important because they simplify the structure of the model and its inference and ease the possibly extensive computational burden associated with spatial data sets. However, for the researchers it is often a hard task to decide which assumptions they should choose due to the complex structure of the data. We have investigated and developed frequency domain empirical likelihood based testing procedures to check for different assumptions for the spatial data.*

Keywords. *Discrete Fourier Transform; Estimating equations; Hypotheses testing; Periodogram; Spectral moment conditions.*

4.12

The generalised autocovariance function

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

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

Abstract. *The generalised autocovariance function is defined for a stationary stochastic process as the inverse Fourier transform of the power transformation of the spectral density function. Depending on the value of the transformation parameter, this function nests the inverse and the traditional autocovariance functions. A frequency domain non-parametric estimator based on the power transformation of the pooled periodogram is considered and its asymptotic distribution is derived. The results are employed to construct classes of tests of the white noise hypothesis, for clustering and discrimination of stochastic processes and to introduce a novel feature matching estimator of the spectrum.*

Keywords. *Stationary processes; Spectral estimation; White noise tests; Feature matching; Discriminant analysis.*

4.13

Estimation of semiparametric models by smoothed profile likelihood

Dennis Kristensen[1]

[1] *Department of Economics, University College London*

Abstract. *A new class of kernel-based estimators of semiparametric models is proposed. The estimators are based on a smoothed version of the global objective function that is normally used to estimate semiparametric models. This modified version is simply defined as the sum of the local, kernel-smoothed ones that defines the estimator of the nonparametric component. By using the one and the same objective function to construct estimators of both components, we obtain a regular profile likelihood with the added benefits that this generates. In particular, the semiparametric estimator will suffer from fewer biases and variances compared to standard two-step kernel-based estimators. Also, no undersmoothing is needed in the estimation of the nonparametric component, and so standard bandwidth selection methods can be employed.*

Keywords. *semiparametric estimation; two-step; kernel estimation.*

4.14

Consistent estimation of covariance matrix spectrum in large dimension

O. Ledoit[1] and M. Wolf[1,*]

[1] *Department of Economics, University of Zurich; olivier.ledoit@econ.uzh.ch, michael.wolf@econ.uzh.ch*
[*] *Corresponding author*

Abstract. *The covariance matrix is arguably one of the most important object in Statistics, with countless applications across finance, neuroimaging, genomics, wireless communications, and several other fields. In many applications the sample size is not large enough relative to the number of variables to estimate the whole covariance matrix accurately. Thus the focus shifts to the estimation of the* spectrum *of the covariance matrix, which is the family of ordered eigenvalues (also called* principal components*). Not only is the spectrum an interesting object in and of itself, but recent research has shown that it is also a key ingredient in estimating the covariance matrix optimally using the technique called* nonlinear shrinkage*. In this paper we introduce a deterministic, finite-dimensional function called QuEST (for Quantized Eigenvalues Sampling Transform) that maps population eigenvalues into sample eigenvalues, and show how to invert it numerically in order to obtain a consistent estimator of the population spectrum. Extensive Monte Carlo simulations demonstrate that our estimator of the spectrum of the population covariance matrix behaves well in finite sample.*

Keywords. *Large dimensional asymptotics; Covariance matrix; Eigenvalues; Spectral decomposition; Principal components.*

4.15

Nonlinear shrinkage for portfolio selection: Markowitz meets Goldilocks

Oliver Ledoit[1] and Michael Wolf[1]

[1] *Department of Economics, University of Zurich; olivier.ledoit@econ.uzh.ch, michael.wolf@econ.uzh.ch*
[*] *Corresponding author*

Abstract. *Markowitz (1952) portfolio selection requires estimates of (i) the vector of expected returns and (ii) the covariance matrix of returns. Many proposals to address the first question exist already. This paper addresses the second question. We promote a new nonlinear shrinkage estimator of the covariance matrix that is more flexible than previous linear shrinkage estimators and has 'just the right number' of free parameters (that is, the Goldilocks principle). In a stylized setting, the nonlinear shrinkage estimator is asymptotically optimal for portfolio selection. In addition to theoretical analysis, we establish superior real-life performance of our new estimator using backtest exercises.*

Keywords. *Large-dimensional asymptotics; Markowitz portfolio selection; Nonlinear shrinkage.*

References

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

4.16

Circular scale spaces and early stem cell diversification

S.F. Huckemann[1] *, M. Sommerfeld[2] and A. Munk[1]

[1] *Felix Bernstein Institute for Mathematical Statistics in the Biosciences*
Georg-August-Universität Göttingen, Germany
[2] *Statistical and Applied Mathematical Sciences Institute*
Duke University, Durham, NC, USA
* *Corresponding author: huckeman@math.uni-goettingen.de*

Abstract. *We generalize the SiZer of Chaudhuri and Marron (1999, 2000) for the detection of shape parameters of densities on the real line to the case of circular data and show that under reasonable regularity only the circular Gaussian gives a variation diminishing semi-group. We introduce the concept of inferred persistence of shape features and apply a CiZer (circular SiZer) – based mode persistence diagram to the analysis of early differentiation in adult human stem cells from their actin-myosin filament structure.*

Keywords. *Semi-groups; variation-reducing; heat equation; circular Gaussian; mode persistence diagram.*

References

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**(447), 807–823.

Chaudhuri, P. and Marron, J.S. (2000). Scale space view of curve estimation. *The Annals of Statistics* **28**(2), 408–428.

4.17

Statistical analysis of projective shapes of curves

Robert Paige[1*], Vic Patrangenaru[2], and Mingfei Qiu[2]

[1] *Missouri S & T, Rolla, Missouri, U.S.A; paigero@mst.edu*
[2] *Florida State University, Tallahassee, Florida, U.S.A; vic@stat.fsu.edu, mingfeiqiu@stat.fsu.edu*
[*] *Corresponding author*

Abstract. *We consider three dimensional image analysis of curves which may be more or less planar. We first consider a nonparametric test of planarity for curves where regular camera pictures have been taken. Here the projective shapes of the curves are treated as points on infinite dimensional projective shape manifolds. As a practical exercise, we consider the three dimensional reconstruction of contours for leaves which are not planar. We illustrate our methodology on a data set consisting pictures of leaves suspended in midair.*

Keywords. *Projective Shape Analysis; Functional Data Analysis*

4.18

Beyond means: a global view of the Fréchet function

W. Mio[1,*] and F. Mémoli[2]

[1] *Department of Mathematics, Florida State University, Tallahassee, FL 32312; mio@math.fsu.edu*
[2] *Department of Mathematics, Ohio State University, Columbus, OH 43210; memoli@math.osu.edu*
[*] *Corresponding author*

Abstract. *The barycenter of a (Borel) probability measure in Euclidean space provides a simple, yet useful summary of the data landscape. The Fréchet function V provides a pathway to extension of the notion of barycenter to probability measures defined on more general metric spaces, including Riemannian manifolds and geodesic stratified spaces. In this general setting, the Fréchet mean is not necessarily unique. Moreover, unlike the classical case, rich additional information about the data distribution also is reflected in the global behavior of V. We present a topological approach to global properties of Fréchet functions associated with probability measures on compact metric spaces. We also examine stability and consistency, and discuss some foundational aspects of model reduction that enable visualization.*

Keywords. *Barycenter; Fréchet mean; Fréchet function.*

4.19

Cartan means and cartan anti-means on stratified spaces

V. Patrangenaru[1,*], H. Hendriks[2] and Mingfei Qiu[1]

[1] *Florida State University; vic@stat.fsu.edu, mingfeiqiu@stat.fsu.edu*
[2] *Radboud University Nijmegen; H.Hendriks@math.ru.nl*
[*] *Corresponding author*

Abstract. *Means on stratified spaces were motivated to analyze data on manifolds and on tree spaces (see Sqwerer et al (2014) and references therein). (Cartan, 1928) first introduced the concept of center of mass on a Riemannian manifold of nonpositive curvature (a notion that over time was called in various contexts Fréchet sample mean, Karcher mean, intrinsic or extrinsic sample mean, Riemannian center of mass, Ziezold mean etc. In this paper we extend the notion of Fréchet means on stratified spaces (see Patrangenaru et al (2013), Hotz et al (2014) or Ellingson et al (2013)). Here* Cartan means (anti-means) *are introduced as minimizers (respectively maximizers) of the Fréchet function associated with a probability measure on a stratified space. We discuss asymptotics, stickiness and other nonparametric aspects for Cartan means and Cartan anti-means.*

Keywords. *Stratified space; Fréchet function; Cartan means; Sticky means; Asymptotics.*

References

Rabi N. Bhattacharya, Marius Buibas, Ian L. Dryden, Leif A. Ellingson, David Groisser, Harrie Hendriks, Stephan Huckemann, Huiling Le, Xiuwen Liu, James S. Marron, Daniel E. Osborne, Vic Patrângenaru, Armin Schwartzman, Hilary W. Thompson, and Andrew T. A.Wood. (2013) Extrinsic data analysis on sample spaces with a manifold stratification. *Advances in Mathematics, Invited Contributions at the Seventh Congress of Romanian Mathematicians, Brasov, 2011*, Publishing House of the Romanian Academy (Editors: Lucian Beznea, Vasile Brîzanescu, Marius Iosifescu, Gabriela Marinoschi, Radu Purice and Dan Timotin), 241–252.

É. Cartan (1928). Lèçons sur la Géométrie des Espaces de Riemann. Gauthier-Villars, Paris.

L. Ellingson, V. Patrangenaru, H. Hendriks, P. S. Valentin (2013) CLT on Low Dimensional Stratified Spaces. *Revision submitted at Proceedings of the First INSPS Conference.*

Thomas Hotz, Stephan Huckemann, Huiling Le, James S. Marron, Jonathan C. Mattingly, Ezra Miller, James Nolen, Megan Owen, V. Patrangenaru and Sean Skwerer (2013). Sticky Central Limit Theorems on Open Books. *Annals of Applied Probability* **23**, 2238–2258.

Sean Skwerer, Elizabeth Bullitt, Stephan Huckemann, Ezra Miller, Ipek Oguz, Megan Owen, Vic Patrangenaru, and J.S. Marron (2014). Phylogenetic Treespace Methods for Analysis of Brain Artery Tree Data. *Accepted at JMIV. DOI 10.1007/s10851-013-0473-0*

4.20

Segmenting multiple time series by contemporaneus linear transformation

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

[1] *University of Melbourne, Australia; jinyuan.chang@unimelb.edu.au*
[2] *Peking University, China; guobin1987@pku.edu.cn*
[3] *London School of Economics, UK; q.yao@lse.ac.uk*
[*] *Corresponding author*

Abstract. *The goal of the paper is to seek for a contemporaneous linear transformation for a p-variate time series such that the transformed series is segmented into several lower-dimensional subseries, and those subseries are uncorrelated with each other both contemporaneously and serially. The method is based on both auto- and cross-correlations only, making no assumptions on underlying models. It also applies to the cases when the dimension p is large in relation to the sample size n. The asymptotic theory is established for both fixed p and diverging p when $n \to \infty$.*

Keywords. *Dimension reduction; Eigen-analysis; High-dimensional time series; Segmentation*

4.21

Statistical modeling of spatial functional data: application to biomedical data

S. Dabo-Niang[12]

[1] *Laboratory EQUIPPE, University Lille 3, Villeneuve d'Ascq, France ; sophie.dabo@univ-lille3.fr*
[2] *INRIA Lille Nord-Europe, MODAL-Team*

Abstract. *Spatial statistics includes any (statistical) techniques which study phenomenons observed on spatial sets. Such phenomenons appear in a variety of fields: epidemiology, environmental science, econometrics, image processing and many others. Complex issues arise in spatial analysis, many of which are neither clearly defined nor completely resolved, but form the basis for current researches.*

The literature on functional data estimation techniques, which incorporate spatial dependency is limited, see for instance Monestiez and Nerimi (2008), Petrone et al. *(2009), Delicado* et al. *(2010), Dabo-Niang* et al. *(2010), Giraldo* et al. *(2012),...*

The modelization of this kind of data has been selected by Ramsay (2008) among the eight most interesting research subjects in functional data analysis. This is motivated by the increasing number of situations coming from different fields of applied sciences for which the data are of spatial functional nature. This is the case for instance in epidemiology, where data are often spatial, and so spatial location can acts as a surrogate for risk factors.

More recently, there has been increased interest in statistical models for disease mapping in space and time, see for instance Elliott et al. *(2000), Waller and Gotway (2004).*
In functional spatial epidemiological data, many issues arise and form the basis for current talk. Namely, we are interested in spatial functional regression estimation (parametric and non-parametric) applied to epidemiological survey data. More precisely, we consider a regression function where the explanatory variable is a functional random field while the response variable is a real-valued random field. The asymptotic distribution of the proposed estimators are established under some complex sampling. The skills of the methods are illustrated on simulations and real data analysis.

Keywords. *Spatial Data - Epidemiology - Functional data*

References

Dabo-Niang, S, Yao, A-F, Pischedda, L, Cuny, P and Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, **24(4)**, 487-497.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, **21(3-4)**, 224-239.

Giraldo, R. Mateu, J. and Delicado, P. (2012). geofd: An R Package for Function-Valued Geostatistical Prediction. *Revista Colombiana de Estaística* **35**, 385-407.

4.22

Estimation of non-negative surfaces over complex domains using bivariate splines

S. Guillas[1,*], M.-J. Lai[2], J. Gaudart[3]

[1] *University College London, UK; s.guillas@ucl.ac.uk*
[2] *University of Georgia, USA; mjlai@math.uga.edu*
[3] *Université de Marseille, France; jean.gaudart@univ-amu.fr*
[*] *Corresponding author*

Abstract. *In this paper we consider the estimation of a non-negative surface over a complex domain with an irregular boundary and interior holes. We employ bivariate splines over triangulations (Lai and Schumaker, 2007) to build the surface. Then non-negativity condition can be satisfied by recasting the estimation of the spline's coefficients as a constrained penalized least squares problem over the spline basis. In addition, boundary conditions are enforced in accordance with the given application. In the absence of replicates of a given surface, the computation of spatial uncertainties in the estimation is here addressed by bootstrapping the original surface. As a result, we compute and display boxplots for the distribution of surfaces that rely on functional data depth (Lopez-Pintado and Romo, 2009; Sun and Genton, 2011). An illustration to the estimation of population density of a town separated by a river highlights the skills of the method.*

Keywords. *Bivariate Splines; Functional Data; Data Depth*

References

Lai, M. J. and Schumaker, L. L. (2007). *Spline Functions over Triangulations.* Cambridge University Press.

Lopez-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* **104**, 486–503.

Sun, Y. and Genton, M. G. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics* **20**, 316–334.

4.23

Bayesian smoothing of functional data

D. Cox[1,*], J. Yang[1] and H. Zhu[2]

[1] *Rice University MS-138, P.O.Box 1892, Houston TX, 77251 USA; dcox@rice.edu, jy13@rice.edu*
[2] *Virginia Tech, MC0439, 250 Drillfield Drive, Blacksburg, VA 24061 USA; hongxiao@vt.edu*
[*] *Corresponding author*

Abstract. *We consider Bayesian inference in the for the data model* $Y_{ij} = Z_i(t_j) + \epsilon_{ij}$ *where the* $Z_i(t)$ *are smooth functions (called signals) and the* ϵ_{ij} *are i.i.d. noise. We assume the* Z_i*s are independent realizations of a Gaussian process and put priors on the mean and covariance functions. Our results show that we obtain improved accuracy in estimating the signals over that obtained by smoothing each one individually. The estimation of the mean function* $\mu(t) = E[Z_i(t)]$ *is also an improvement over simply averaging the smoothed signal estimates.*

Keywords. *Nonparametric regression; Functional data analysis; Bayesian estimation.*

4.24

Bayesian lienar regression-related models with functional inputs

Y. Pokern[1,*], S. Guillas[1] and A. Y. Park[1]

[1] *University College London; y.pokern@ucl.ac.uk, s.guillas@ucl.ac.uk, a.y.park@ucl.ac.uk*

Abstract. *Linear regression with functional covariates is applied to emulators with functional inputs from a Hilbert space. A Gaussian prior measure is imposed on the unknown regression coefficient using a differential operator representation of the prior precision. Bayesian conjugacy enables fast computation on low dimensional domains using a finite element representation. Hyperparameters are selected rationally and the model is applied to various sample datasets and compared to other approaches including functional principal component analysis.*

Keywords. *Emulator; Functional Covariate; Bayesian Nonparametrics*

4.25

An empirical likelihood approach to goodness of fit testing

H. Peng[1] and A. Schick[2,*]

[1] Department of Mathematical Sciences, Indiana University Purdue University at Indianapolis, Indianapolis, IN 46202, USA; hpeng@math.iupui.edu
[2] Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902, USA; anton@math.binghamton.edu
[*] Corresponding author

Abstract. *Motivated by applications to goodness of fit testing, the empirical likelihood approach is generalized to allow for the number of constraints to grow with the sample size and for the constraints to use estimated criteria functions. The latter is needed to deal with nuisance parameters. The proposed empirical likelihood based goodness of fit tests are asymptotically distribution free. For univariate observations, tests for a specified distribution, for a distribution of parametric form, and for a symmetric distribution are presented. For bivariate observations, tests for independence are developed.*

Keywords. *Infinitely many constraints; Estimated constraint functions; Testing for a parametric model; Testing for symmetry; Testing for independence.*

4.26

Multidimensional lack of fit tests for linear regressions models using minimal weighted matchings

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

[1] Department of Mathematics, Kansas State University; frm@math.ksu.edu
[2] Department of Statistics, Kansas State University; jwneill@ksu.edu
[*] Corresponding author

Abstract. *In previous work the authors developed a graph theoretic representation of near replicate clusterings of statistical units to obtain lack of fit tests for linear regression models that have good power for detecting model inadequacy. First, the graph was used to determine a special collection of clusterings (the atoms consistent with the graph) and then an optimization procedure was applied(a maximin method or restricted least squares approach) to choose an optimal clustering. In the current work, we use a different special collection of clusterings consistent with the graph. These are the clusterings that group at most two vertices together, provided these two vertices form an edge of the graph, which is why we call them edge clusterings. They are called matchings in the field of combinatorial optimization. Edge clusterings possess special advantages for testing regression lack of fit, some of which were discussed in more recent* *work by the authors. However, they also allow efficient implementation in high dimensional models with a large number of predictor variables, which is the emphasis of the current work.*

Keywords. *Matchings; Multidimensional Regression; Lack of Fit.*

4.27

Residual-based empirical distribution functions and tests

Ursula U. Müller[1]

[1] *Department of Statistics, Texas A&M University; uschi@stat.tamu.edu*

Abstract. *We consider semiparametric regression models with independent errors and covariates. This covers parametric (linear and nonlinear) and nonparametric regression as special cases. We propose estimating the error distribution using a residual-based empirical distribution function (which requires suitable estimators of the regression function). We will identify various regression models where the residual-based empirical distribution function allows a simple expansion which, in particular, characterizes an efficient estimator of the error distribution. With this expansion at hand, the residual-based empirical distribution function can be used for distribution free lack-of-fit and goodness-of-fit tests, for example martingale transform tests as suggested by Khmaladze and Koul (Ann. Statist. 2004, 2009).*

These results extend to the more general model where the response variable is missing at random. We show that the complete case version of an efficient estimator of the error distribution function remains efficient in the semiparametric regression model with missing data. This generalizes a result by Chown and Müller (2013) for nonparametric regression. We also propose complete case analysis to perform residual-based tests, in particular distribution free tests, since they are powerful and easy to use.

This talk is based on joint work with Justin Chown, Hira Koul, Anton Schick and Wolfgang Wefelmeyer.

Keywords. *Semiparametric regression; Distribution free tests; Efficient influence function; Responses missing at random.*

References

Chown, J. and Müller, U.U. (2013). Efficiently estimating the error distribution in nonparametric regression with responses missing at random. *J. Nonparametr. Statist.* **25**, 665–677.

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

Müller, U.U. and Schick, A. (2014). Efficiency transfer for regression models with responses missing at random. In preparation.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2004). Estimating linear functionals of the error distribution in nonparametric regression. *J. Statist. Plann. Inference* **119**, 75–93.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution function in semiparametric regression. *Statist. Decisions* **25**, 1–18.

4.28

A variable selection method for spatial additive models with applications

Taps Maiti

Abstract. *We develop a variable selection technique, specifically adaptive group LASSO type of selection in additive models with spatially dependent Gaussian random error. We also consider the problem of consistently estimating non-zero components under the same model. We allow the number of components to be 'large' but the number of non-zero components is 'small' compared to the number of observations. To address both selection and estimation, we use adaptive group Lasso technique, where we first use a group Lasso method to reduce the dimension and then apply an adaptive group Lasso method to select the number of non-zero components. We validate the proposed theory and method by simulation studies and real data examples.*

4.29

Modelling the probability of crossing the glass ceiling

Maria Paz Espinosa

Universidad del Pais Vasco (UPV/EHU), Facultad de Ciencias Económicas y Empresariales Econometría y Estadística, Avenida Lehendakari Agirre No. 83, 48015 Bilbao; mariapaz.espinosa@ehu.es

Abstract. *We define the glass ceiling effect in terms of the probability of promotion in a hierarchical structure. We find that the glass ceiling effect is affected by the interaction between gender bias, optimal participation decisions and incumbency advantages. We show how these features interact with gender bias to produce a considerable degree of inequality in the long run, worse at the upper levels of management. A main result is that the glass ceiling effect does not require any difference between the behavior or abilities of male and female candidates but appears as a consequence of the dynamics of gender bias.*

Keywords. *Glass ceiling effects; Optimal decision; Conditional expectations*

4.30

On discrimination and binomial processes

Eva Ferreira

Universidad del Pais Vasco (UPV/EHU), Facultad de Ciencias Económicas y Empresariales Econometría y Estadística, Avenida Lehendakari Agirre No. 83, 48015 Bilbao; eva.ferreira@ehu.es

Abstract. *We use binomial processes to represent the dynamics of gender bias when abilities are not observable, and obtain the long run equilibrium for the rate of gender composition in committees. The model can be applied to biases of different nature (ethnic, race, minorities,...). The selection process is modelled as a Markov process, where the gender composition of actual committees determines the composition of the future committees through a gender difference in the perception of the decision makers.*

Keywords. *Discrimination models; Conditional binomial distribution*

4.31

Testing for glass ceiling effects

Winfried Stute

University of Giessen
Mathematical Institute
Arndtstr. 2
D-35392 Giessen, Germany;
Winfried.Stute@math.uni-giessen.de

Abstract. *In this work we discuss dynamic models designed to describe glass ceiling effects. Especially we develop tests for the hypothesis that such effects do not exist. Some finite sample and asymptotic theory are presented, which are based on discrete time martingale theory.*

Keywords. *Glass ceiling effects; Conditional Laplace transform; Testing*

4.32

The analysis of biased time-to-event data with a cured portion

Walter Faig and Ronghui Xu*

University of California, San Diego; wfaig@ucsd.edu, rxu@ucsd.edu

* *Corresponding author*

Abstract. *Our work was motivated by pregnancy outcomes such as spontaneous abortion or preterm delivery, in the context of observational studies of drug exposure. These are essentially binary endpoints. However, women can enter a study any time during their pregnancy. Not counting for such left truncation leads to bias in the estimated rates. In addition, a substantial portion of the women will not have the events of interest, a portion termed ÔcuredÕ in survival analysis. While left truncation is relatively easily dealt with in the Cox proportional hazards regression, with a cured proportion new methodology is needed. We investigate approaches using the exact semiparametric likelihood, an approximate likelihood, and a weighted (complete data) likelihood. Variance estimates are derived with closed-form expressions. Time permitting efficiency consideration will be discussed.*

Keywords. *Cure rate; left truncation; preterm delivery; spontaneous abortion; weighted likelihood.*

4.33

Not Necessarily sparse classification: Inference based on robust high confidence sets

Jelena Bradic

Abstract. *We propose an inference based on high confidence sets that is robust to sparsity assumption in the high-dimensional linear discriminant framework. High confidence sets have been proved to be extremely useful for linear regression problems, but little has been studied in the classification context with the goal of sparse recovery. We propose a novel inference method in the context of Fisher's linear discriminant, based on a new, robust and data adaptive high confidence set. Turning to the ideas of measurement-in-errors models, we work in the high dimensional setting where the sample size n can be much smaller than the number of possible regressors p, and where the within-cluster variance matrix can be non-sparse. We show that naive plug-in Bayes classifier based on observed data may often be too dense, hence not an ideal candidate to mimic in high dimensions. We construct latent Bayes classifier and show that the proposed method, while fully designed on the observed data, has excellent adaptivity to the sparse latent structure. Our main results are finite sample oracle risk inequalities and* *support recovery probability, bounding various distances of the proposed classifier to sparse, latent Bayes classifier. All the obtained results do not necessarily require sparsity structure on the observed data, hence are robust to it. We compare our estimator to a class of linear discriminant estimators, and show it to be uniformly closer within such class, in the Pitman sense, to the best sparse Bayes estimator. Moreover, we propose a new algorithm that is able to recover the solution path of our estimator for a continuum of regularization parameters.*

4.34

Fast goodness-of-fit tests based on the characteristic function

M. Dolores Jiménez-Gamero[1,*] and Hyoung-Moon Kim[2]

[1] *Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Spain; dolores@us.es*
[2] *Department of Applied Statistics, Konkuk University, Republic of Korea; hmkim@konkuk.ac.kr*
[*] *Corresponding author*

Abstract. *A class of goodness-of-fit tests whose test statistic is an L_2 norm of the difference of the empirical characteristic function of the sample and a parametric estimate of the characteristic function in the null hypothesis, is considered. The null distribution is usually estimated through a parametric bootstrap. Although very easy to implement, the parametric bootstrap can become very computationally expensive as the sample size, the number of parameters or the dimension of the data increase. This work proposes to approximate the null distribution through a weighted bootstrap. The method is studied both theoretically and numerically. It provides a consistent estimator of the null distribution. In the numerical examples carried out, the estimated type I errors are close to the nominal values. The asymptotic properties are similar to those of the parametric bootstrap but, from a computational point of view, it is more efficient.*

Keywords. *Characteristic function; Goodness-of-fit; Weighted bootstrap; Consistency.*

Tail index estimation based on survey data

P. Bertail[1], E. Chautru[2,*] and S. Clémençon[3]

[1] *Université Paris-Ouest, 200 av. de la République, 92000 Nanterre, France; patrice.bertail@gmail.com*
[2] *Université de Cergy-Pontoise, 2 av. Adolphe Chauvin, 95302 Cergy-Pontoise cedex, France; emilie.chautru@gmail.com*
[3] *Télécom ParisTech, 37/39 rue Dareau, 75014 Paris, France; stephan.clemencon@telecom-paristech.fr*
[*] *Corresponding author*

Abstract. *In many application fields of theoretical statistics, the available observations are not independent and identically distributed, but originate from a potentially complex survey scheme. Moreover, in the "Big Data" era, sampling can be viewed as a natural solution to the computational issues induced by the immoderate size of databases. Since ignoring the survey scheme can impede estimation by introducing a non-negligible bias (Bonnery et al., 2012), it is customary to weight the data with the inverse of their probability of inclusion in the sample. While a plethora of analyzes has already been conducted to provide unbiased and efficient estimators of average quantities, to our knowledge, such is not the case for phenomenons involving tails of distributions in the framework of extreme value theory. The analysis of extreme events is yet of major importance for risk management in a plurality of fields, ranging from biology or climatology to finance. In an attempt to conciliate both branches of statistics, we propose here a Horvitz-Thompson variant of the Hill estimator (Hill, 1975), which assesses the extreme value index when the observations are drawn according to a large entropy survey plan like the Poisson design (Hàjek, 1964; Berger, 1998). After having proved its consistency and asymptotic normality under a set of hypotheses involving the calculation of inclusion probabilities and the underlying superpopulation model, we illustrate our results on numerical experiments. It appears in particular that an appropriate choice of inclusion probabilities can neutralize the loss of efficiency due to the sampling phase.*

Keywords. *Survey sampling; Extreme value theory; Hill estimator; Extreme value index.*

References

Berger, Y.G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **67**, 209–226.

Bonnéry, D. and Breidt, J. and Coquet, F. (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli* **18**, 1361–1385.

Hàjek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics* **35**, 1491–1523.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**, 1163–1174.

4.36

On survey sampling and empirical risk minimization

Patrice Bertail[1], E. Chautru[2] and S. Clémençon[2*]

[1] *MODAL'X - Université Paris-Ouest; patrice.bertail@u-paris10.fr*
[2] *LTCI UMR Telecom ParisTech/CNRS No. 5141; emilie.chautru@gmail.com, stephan.clemencon@telecom-paristech.fr*
[*] *Corresponding author*

Abstract. *In certain, situations that shall be undoubtedly more and more common in the Big Data era, the datasets available are so massive that computing statistics over the full samples is hardly feasible, if not unfeasible. A natural approach in this context consists in using survey schemes and substituting the "full data" statistics with their counterparts based on the resulting random samples, of manageable size. It is the main pupose of this paper to investigate the impact of survey sampling on statistical learning methods based on Empirical Risk Minimization (ERM) through the standard binary classification problem, considered here as a "case in point". Precisely, we prove that use of the Poisson survey scheme does not affect much the learning rates, while reducing significantly the number of terms that must be averaged to compute the empirical risk functional with overwhelming probability. These striking results are next shown to extend to more general sampling schemes by means of a coupling technique, originally introduced by Hajek (1964).*

Keywords. *Survey; Statistical Learning Theory, Supervised Binary Classification.*

4.37

Recent advances in empirical processes for survey sampling

P. Bertail[1,*], E. Chautru[2] and Stéphan Clémençon[3]

[1] *Université Paris-Ouest, 200 ave de la République 92000 Nanterre; patrice.bertail@gmail.com*
[2] *Université de Cergy; Emilie.Chautru@gmail.com*
[3] *ENST, 37-39 rue Darreau, 75014 Paris; Stephan.clemencon@gmail.com*
[*] *Corresponding author*

Abstract. *This talk is devoted to the study of the limit behavior of extensions of the empirical process indexed by classes of functions (see van der Vaart and Wellner (2000)), when the data available have been collected through an explicit survey sampling scheme and is motivated by some problems linked to practical exploitation of big datas. Indeed, in many situations, statisticians have at their disposal not only data but also weights arising from some survey sampling plans. On the other hand, for big data, survey sampling may be an efficent tool to* *reduce computational costs involved by massive data. Our main goal is here to investigate how to incorporate the survey sampling scheme into the inference procedure dedicated to the estimation of a probability measure P on a measurable space (viewed as a linear operator acting on a certain class of functions F), in order to guarantee its asymptotic normality. Recent results have been obtained for stratified sampling plans, with an uniform with replacement sampling scheme in each stratas. Our approach follows that of Hajek (1964), extended next by Berger (1998), and is applicable to general sampling surveys, namely those with unequal first order inclusion probabilities which are of the Poisson type or sequential/rejective.*

We propose first to study an Horvitz-Thompson Poisson type empirical process adequately centered to take into account the variability of sampling size of the Poisson sampling scheme. The main result of the paper is then a Functional Central Limit Theorem (FCLT) for an Horvitz Thompson empirical process for sampling plans which are closed in term of total variation or in term of entropy to the Poisson sampling plan; this includes rejective sampling, successive sampling, Rao-Sampford sampling etc...

Keywords. *Survey sampling, empirical process, big data*

References

Hajek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *Ann. Math. Statist.*, **35**, 1419-1880

Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 209–226.

van der Vaart A.W., Wellner, J.A. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*,Springer Series in Statistics.

4.38

Approximation of rejective sampling inclusion probabilities and applications

N1. Anne Ruiz-Gazen[1,*], N2. Hélène Boistard[1] and N3. Hendrik Lopuhaä[2]

[1] *Toulouse School of Economics; anne.ruiz-gazen@tse-fr.eu*
[2] *Delft Institute of Applied Mathematics*
[*] *Corresponding author*

Abstract. *In the finite population context, asymptotic properties of estimators, such as asymptotic design unbiasedness, consistency and asymptotic normality, are usually derived under assumptions on high order inclusion probabilities of the sampling design. Some recent examples in the literature on survey sampling concern nonparametric and semi-parametric estimators in the model-assisted context when auxiliary information is available(see Breidt et al. (2000),Breidt et al. (2007),Wang (2009)). These assumptions on inclusion probabilities hold for some particular sampling designs such as the simple random sampling without replacement. In the present work, we prove that these assumptions also hold for the rejective sampling which is a particular unequal probabilities and without replacement design. The proof is based upon a generalization* *of the approximation result obtained by Hájek (1964) for the first and second order inclusion probabilities of rejective sampling. An approximation of the inclusion probabilities of any order is provided together with a more precise remainder term in the expansion. The results (Boistard et al., 2012) are applied to illustrate that the rejective sampling design satisfies conditions on higher order correlations imposed in the recent literature for deriving asymptotic results.*

Keywords. *Finite population; Model-assisted estimation; Nonparametric; Semi-parametric; Survey sampling*

References

Boistard, H., Lopuhaä, H. P. and Ruiz-Gazen, A. (2012), Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics*, **6**, 1967–1983.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics* **28, 4**, 1026–1053).

Breidt, F. J., Opsomer, J. D., Johnson, A. A. and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology* **33, 1**, 35-44.

Hájek, J. (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* **35**, 1491–1523.

Wang, L. (2009). Single-index model-assisted estimation in survey sampling. *Journal of Nonparametric Statistics* **21,4**, 485–504.

4.39

Change point inference

Klaus Frick[1], Axel Munk[2,*], and Hannes Sieling[2]

[1] *NTB Buchs, Switzerland; klaus.frick@ntb.ch*

[2] *Institute for Mathematical Stochastics, Göttingen University, Germany; munk@math.uni-goettingen.de, hsielin@uni-goettingen.de*

[*] *Corresponding author*

Abstract. *We introduce a new estimator SMUCE (simultaneous multiscale change-point estimator) for the change-point problem in exponential family regression. An unknown step function is estimated by minimizing the number of change-points over the acceptance region of a multiscale test. The probability of overestimating the true number of change-points K is controlled by the asymptotic null distribution of the multiscale test statistic. Further, we derive exponential bounds for the probability of underestimating K. Balancing these quantities allows to maximize the probability of correctly estimating K. All results are non-asymptotic for the normal case. Based on these bounds, we construct honest confidence sets for the unknown step function and its change-points. It is shown that SMUCE asymptotically achieves the optimal detection rate of vanishing signals in a multiscale setting. We illustrate how dynamic programming techniques can be employed for efficient computation of estimators and confidence* *regions. The performance of the proposed multiscale approach is illustrated by simulations and in several applications including ion channel recordings, CGH array analysis, and photoemission spectroscopy. This work is based on Frick (2013).*

Keywords. *Change Point Analysis; Simultaneous Inference; Detection Rate; CGH data; Dynamic Programing.*

References

Frick, K., Munk, A., Sieling, H. (2013). Change point inference *Journ. Royal Statist. Soc.* **Ser. B**, with discussion and rejoinder, to appear.

4.40

Multiple testing of local maxima fot detection of peaks of Gaussian random fields

D. Cheng[1] and A. Schwartzman[1,*]

[1] *Department of Statistics, North Carolina State University; dcheng2@ncsu.edu, aschwar@ncsu.edu.*
[*] *Corresponding author*

Abstract. *An important problem in image analysis is to find local significant regions, either for a single image or for the difference between two or more images, where the need is to make inferences about spatial features such as peaks rather than individual pixels or voxels. Examples include finding protein binding sites in ChIP-Seq genomic data (1D), finding fluorescent molecules in cell nanoscopy (2D), and finding regions of neural activation in brain imaging (3D). Here a formal procedure is proposed for detecting smooth peaks buried in stationary noise, where the number, height and location of the peaks are unknown. The procedure, involving kernel smoothing and testing of local maxima, is easy to implement and takes advantage of existing multiple testing procedures. Theory and simulations show that the false discovery rate of detected peaks is controlled asymptotically and that the optimal bandwidth corresponds to the "matched filter" principle, where the kernel size should be close to that of the peaks to be detected.*

Keywords. *Topological inference; False discovery rate; Kernel smoothing, Matched filter.*

4.41

Elastic functional regression analysis

J. D. Tucker[1], W. Wu[1] and A. Srivastava[1,*]

[1] *Department of Statistics, Florida State University; dtucker@stat.fsu.edu, wwu@stat.fsu.edu, anuj@stat.fsu.edu*
[*] *Corresponding author*

Abstract. *We study regression problems using functional predictors in situations where these functions contain both phase and amplitude variability. In other words, the functions are misaligned either due to errors in time measurements or some other phase variability, and these errors can significantly degrade both model estimation and prediction performance. The current techniques either ignore the phase variability, or handle it via pre-processing, i.e. use an off-the-shelf technique for functional alignment and phase removal. A third possibility of forming a naive least-square solution, by incorporating phase removal, can result in degenerate solutions. We derive a comprehensive approach that assumes nonparametric phase model and the model estimation is handled at the same time as the phase removal, using a mathematical representation called square-root slope functions (Srivastava et al. (2011a,b); Tucker et al. (2013)). These functions preserve $\mathbb{L}^2$ norm under simultaneous warping and are ideally suited for simultaneous estimation of regression and warping parameters. This estimation is performed using numerical optimization under appropriate constraints. Using both simulated and real world data sets, we demonstrate our approach for functional logistic regression and evaluate its predictions performance relative to some current ideas. In addition, we propose an extension to functional multinomial logistic regression.*

Keywords. *Functional logistic regression; elastic shape analysis; square-root slope functions; Warping; Functional Alignment.*

References

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn I. (2011). Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE Pattern Analysis and Machine Intelligence* 33(7):1415-1428.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of Functional Data Using Fisher-Rao Metric. *arXiv*:1103.3817v2.

Tucker, J. D., Wu, W., and Srivastava, A. (2013). Generative Models for Functional Data Using Phase and Amplitude Separation. *Computational Statistics and Data Analysis* 61:50-66.

4.42

Principal flows

Victor M. Panaretos[1,*]

[1] Department of Mathematics, EPFL; victor.panaretos@epfl.ch
[*] Corresponding author

Abstract. *We revisit the problem of extending the notion of principal component analysis (PCA) to multivariate data sets that satisfy non-linear constraints, therefore lying on Riemannian manifolds. Our aim is to determine curves on the manifold that retain their canonical interpretability as principal components, while at the same time being flexible enough to capture non-geodesic forms of variation. We introduce the concept of a principal flow, a curve on the manifold passing through the mean of the data, and with the property that, at any point of the curve, the tangent velocity vector attempts to fit the first eigenvector of a tangent space PCA locally at that same point, subject to a smoothness constraint. It is shown that principal flows can yield the usual principal components on a Euclidean space. By means of examples, it is illustrated that the principal flow is able to capture patterns of variation that can escape other manifold PCA methods. (Based on joint work with T. Pham, Melbourne, and Z. Yao, EPFL).*

Keywords. *PCA; Euler-Lagrange method; Riemmaninan manifold.*

References

[1] Panaretos, V. M., Pham, T., & Yao, Z. (2014). Principal Flows. *Journal of the American Statistical Association (Theory and Methods)*, to appear.

4.43

Big-log-convance distribution and regression functions

L. Dümbgen[1,*], P. Kolesnyk[1] and R. Wilke[2]

[1] University of Bern; duembgen@stat.unibe.ch, petro.kolesnyk@stat.unibe.ch
[2] University of York; ralf.wilke@york.ac.uk
[*] Corresponding author

Abstract. *A univariate function F with values in $[0,1]$ is called bi-log-concave, if both $\log F$ and $\log(1-F)$ are concave. This new shape-constraint is rather natural in many situations. For instance, any c.d.f. F with log-concave density $f = F'$ is bi-log-concave. But bi-log-concavity alone allows for multimodal densities. Various characterizations are provided. It is shown* *that combining any nonparametric confidence band for a distribution function F with the new shape-constraint leads to substantial improvements and implies nontrivial confidence bounds for arbitrary moments and the moment generating function of F. In addition we discuss briefly applications to binary regression.*

Keywords. *Hazard, Honest confidence region, Moment generating function, Moments, Reverse hazard*

4.44

Comparision of two nonparametric regression curves: test of superiority and noninferiority

S. Rakshit[1], M. J. Silvapulle [1,*] and P. K. Sen[2]

[1] *Monash University; suman.rakshit@uwa.edu.au, mervyn.silvapulle@monash.edu*
[2] *University of North Carolina at Chapel Hill; pksen@bios.unc.edu*
[*] *Corresponding author*

Abstract. *A new method is developed for formulating and testing the hypothesis that a new treatment is better than the standard when there is a covariate and the response is represented by a nonparametric regression curve. This is related to analysis of covariance, but without assuming parametric regression models. First, we choose two curves, called the noninferiority and superiority bounds. The former lies below and the latter lies above the mean regression function for the standard treatment. Now, the new treatment is defined to be overall better than the standard if the mean regression function for the new treatment lies above the noninferioirity bound at every value of the covariate and above the superiority bound at least at some values of the covariate. For statistical inference, simultaneous noninferiority and superiority is formulated as the alternative hypothesis. As an example, it may be desired to test against the hypothesis that a government policy is noninferior for every income level and that it is superior at least for some in the low-income group. Because the asymptotic test is conservative, a less conservative bootstrap test is proposed. We compute the critical values at the least null favourable configuration.*

Keywords. *Constrained inference; Noninferiority; Order restricted inference.*

Shape constrained estimation in the Cox model

H.P. Lopuhaä[1] and G.F. Nane[2,*]

[1] *Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD, The Netherlands; h.p.lopuhaa@tudelft.nl*
[2] *Center for Science and Technology Studies, Leiden University, P.O. Box 905, 2300 AX Leiden, The Netherlands; g.f.nane@cwts.leidenuniv.nl*
[*] *H.P.Lopuhaä*

Abstract. *We investigate nonparametric estimation of a monotone baseline hazard and a decreasing baseline density within the Cox model. Two estimators of a nondecreasing baseline hazard function are proposed. We derive the nonparametric maximum likelihood estimator and 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 limit distribution at a fixed point. Both estimators of a nonincreasing baseline hazard and their asymptotic properties are acquired in a similar manner. Furthermore, we introduce a Grenander type estimator for a nonincreasing 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. We show that this estimator is strong consistent and derive its asymptotic distribution at a fixed point.*

Keywords. *Breslow estimator; Cox model; Shape constrained nonparametric maximum likelihood.*

4.46

Isotonic regression in several dimensions

Dragi Anevski[1] and Wolfgang Polonik[2]

[1] *Mathematical Sciences, Lund University, Sweden, dragi@maths.lth.se,* [2] *Department of Statistics, UC Davis, USA, wpolonik@ucdavis.edu*

Abstract. *We discuss regression and density function estimation of regression functions and density functions that are ordered with respect to the partial order on $\mathbb{R}^2$. We discuss in particular distributional results.*

Keywords. *Bimonotone; Regression; Density function estimation.*

4.47

Accounting for non-ignorable drop-out in mixed latent Markov models with covariates

Francesco Bartolucci[1*] and Alessio Farcomeni[2]

[1] *Department of Economics, University of Perugia (IT); bart@stat.unipg.it*
[2] *Department of Public Health and Infectious Diseases, Sapienza - University of Rome (IT); alessio.farcomeni@uniroma1.it*
[*] *Corresponding author*

Abstract. *Mixed latent Markov (MLM) models represent an important tool of analysis of longitudinal data when response variables are affected by time-fixed and time-varying unobserved heterogeneity, in which the latter is accounted for by a hidden Markov chain (Maruotti, 2011; Bartolucci et al., 2013). In order to avoid bias when using a model of this type in the presence of non-ignorable drop-out, we propose an extension of the LM approach that may be used with multivariate longitudinal data, in which one or more outcomes of a different nature are observed at each time occasion. The component of the model used to account for missing data is based on sharing common latent variables with the longitudinal component of the model. In order to perform maximum likelihood estimation of the proposed model by the expectation-maximization algorithm, we extend the usual backward-forward recursions of Baum and Welch (Baum et al., 1970). The algorithm has the same complexity of the one adopted in cases of ignorable drop-out. We illustrate the proposed approach through simulations and an application based on data coming from a medical study about primary biliary cirrhosis in which there are two outcomes of interest, one is continuous and the other is binary.*

Keywords. *Discrete Latent Variables; Expectation-maximization Algorithm; Hidden Markov Models.*

References

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**, 164–171.

Bartolucci, F., Farcomeni, A. and Pennoni, F. (2013). *Latent Markov models for longitudinal data.* Chapman & Hall/CRC Press. Boca Raton, FL.

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review* **79**, 427–454.

4.48

Semiparametric methods with mixed measurement error and misclassification in covariates

Grace Yi[1]

[1] *University of Waterloo, Canada yyi@uwaterloo.ca*

Abstract. *Covariate measurement imprecision or errors arise frequently in many areas. It is well known that ignoring such errors can substantially degrade the quality of inference or even yield erroneous results. Although in practice both continuous covariates subject to measurement error and discrete covariates subject to misclassification can occur, research attention in the literature has mainly focused on addressing either one of these problems separately. Relatively little work has been done to accommodate both features simultaneously. In this talk, I will discuss inference methods for analyzing data with mixed measurement error and misclassification in covariates. Specifically, both likelihood based and semiparametric methods will be described. This is joint work with Yanyuan Ma, Donna Spiegelman and Raymond Carroll.*

Keywords. *Classification; Measurement error; Semiparametric models*

4.49

Semiparametric inference of high-dimensional graphical models

Lingzhou Xue[1,*]

[1] *Department of Statistics, Penn State University, lzxue@psu.edu*
[*] *Corresponding author*

Abstract. *Graphical models have been widely used to explore the underlying conditional dependence structure in large-scale networks. It is common to assume that the available data are fully observable and generated from a specified Markov random field. However, in real-world applications, observations are often non-Gaussian and contain missingness or truncation, which has* *a significant effect on statistical inference. In this work, we propose a unified semiparametric inference of high-dimensional graphical models to retain the appealing graphical interpretability under non-Gaussianity and certain missingness/truncation. Theoretical properties are established under the high-dimensional setting, and numerical properties are also demonstrated in both simulation studies and real applications. This work greatly extends the methodology and applicability of graphical modeling. This talk is based on several joint works with collaborators.*

Keywords. *Graphical model; Semiparametric inference; Incomplete data; Pseudolikelihood; Regularization.*

4.50

Nonparametric eigenvalue-regularized precision or covariance matrix estimator

Clifford Lam

London School of Economics and Political Science; C.Lam2@lse.ac.uk

Abstract. *Recently there are numerous works on the estimation of large covariance or precision matrix. The high dimensional nature of data means that the sample covariance matrix can be ill-conditioned. Without assuming a particular structure, much efforts have been devoted to regularizing the eigenvalues of the sample covariance matrix. We introduce nonparametric regularization of these eigenvalues through subsampling of the data. The subsampling idea for covariance matrix estimation is originally introduced in Abadir, Distaso and Žikeš (2010). We improve on their covariance estimator, and for the first time provides vigorous proof that our version enjoys asymptotic optimal nonlinear shrinkage of eigenvalues with respect to the Frobenius error norm. Coincidentally, this nonlinear shrinkage is asymptotically the same as that introduced in Ledoit and Wolf (2012). One advantage of our estimator is its computational speed when the dimension p is not extremely large. Our estimator also allows p to be larger than the sample size n, and is always positive semi-definite. We prove that with respect to the Stein's loss function, the inverse of our estimator is the optimal precision matrix estimator. We also showed that all the aforementioned optimality holds for data with a factor structure as well, which can be useful in portfolio allocation. Our method avoids the need to estimate the unknown factors and factor loadings matrix first, and directly gives the covariance or precision matrix estimator. We compare the performance of our estimators with other methods through extensive simulations and a real data analysis.*

Keywords. *High dimensional covariance matrix; Precision matrix; Regularized eigenvalues; Stieltjes transform; Subsampling.*

References

Abadir, K. M., W. Distaso and F. Žikeš (2010). Model-free estimation of large variance matrices. *The Rimini Centre for Economic Analysis* **WP**, 10–17.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics* **40(2)**, 1024–1060.

4.51

An iterative estimation procedure for generalised varying-coefficient models with unspecified link functions

Wenyang Zhang[1], Degui Li[2,*] and Yingcun Xia[1]

[1] *The University of York, UK; National University of Singapore; wenyang.zhang@york.ac.uk, staxyc@nus.edu.sg*
[2] *The University of York, UK; degui.li@york.ac.uk*
* *Corresponding author*

Abstract. *In this talk, the generalised varying-coefficient models with unspecified link functions will be addressed. A very weak identification condition will be presented for the generalised varying-coefficient models when their link function is unknown. Under the identification condition, I will introduce an iterative estimation procedure for the generalised varying-coefficient models with unspecified link function. An algorithm will also be introduced to implement the proposed estimation procedure. I will also show the asymptotic properties of the nonparametric estimators obtained by the proposed iterative estimation procedure, and some simulation study results. Finally, I will use the generalised varying-coefficient models with unspecified link function to analyse a real dataset.*

Keywords. *Generalised varying-coefficient models; Iterative estimation procedure; Kernel smoothing; Weighted least squares.*

4.52

Extending the scope of cube root asymptotics

M. Seo[1,*] and T. Otsu[1]
[1] London School of Economics; m.seo@lse.ac.uk, t.otsu@lse.ac.uk
*Corresponding author

Abstract. *This article extends the scope of cube root asymptotics for M-estimators in two directions: allow weakly dependent observations and criterion functions drifting with the sample size typically due to a bandwidth sequence. For dependent empirical processes that characterize criterions inducing cube root phenomena, maximal inequalities are established to derive the convergence rates and limit laws of the M-estimators. The limit theory is applied not only to extend existing examples, such as the maximum score estimator, nonparametric maximum likelihood density estimator under monotonicity, and least median of squares, toward weakly dependent observations, but also to address some open questions, such as asymptotic properties of the minimum volume predictive region, conditional maximum score estimator for a panel data discrete choice model, and Hough transform estimator with a drifting tuning parameter.*

Keywords. *Cube root asymptotics; Minimum volume prediction; Maximum score estimator; Hough transform estimator; Least median of squares*

4.53

Whittle likelihood for bivariate processes

Sofia C. Olhede[1,*], Adam M. Sykulski[1], Jeffrey J. Early[2], Jonathan M. Lilly[2] & Frederik J. Simons[3]

[1] *Department of Statistical Science, University College London, London; s.olhede@ucl.ac.uk, a.sykulski@ucl.ac.uk*

[2] *NorthWest Research Associates, 4118 148th Ave NE, Redmond, WA 98052; jearly@nwra.com, lilly@nwra.com*

[3] *Department of Geosciences Â· Princeton University Â· Guyot Hall Â· Princeton NJ 08544; fjsimons@princeton.edu*

[*] *Corresponding author*

Abstract. *We shall discuss using Whittle likelihood for estimation of bivariate processes. The Whittle likelihood is formulated in the frequency domain, and relies on a number of asymptotic results for applicability. Real data analysis challenges such assumptions, especially if the frequency domain understanding of shorter segments of time or smaller spatial domains is to be arrived at.*

We propose modifications to the Whittle likelihood that improve estimation. The first takes the form of a complex-valued representation of bivariate structure that is only evident by separating negative and positive frequency behaviour, see Sykulski (2013). Flexible inference methods for such parametric models are proposed, and the properties of such methods are derived.

Secondly we propose an adjustment to Whittle likelihood suitable for ameliorating sampling effects semi-parametrically, thus advancing the state-of-the-art in frequency domain modelling and estimation of time series in general, see Simins (2013); Sykulski (2013). This reduces small sample bias, and can be interpreted as extending Whittle likelihood to a composite likelihood method.

Keywords. *Multivariate time series; nonstationary time series, time series inference, Whittle estimation*

References

Simons, F. J. & Olhede, S. C. (2013). Maximum-likelihood estimation of lithospheric flexural rigidity, initial-loading fraction, and load correlation, under isotropy. *Geoph. J. Int.*, 193 (3), 1300–1342.

Sykulski, A. M., Olhede, S. C., Lilly, J. M. and Early, J. J. (2013). The Whittle Likelihood for Complex-Valued Time Series. *arXiv:1306.5993*, technical report.

Sykulski, A. M., Olhede, S. C., Lilly, J. M. and Danioux,E. (2013). Lagrangian Time Series Models for Ocean Surface Drifter Trajectories. *arXiv:1312.2923*, technical report.

4.54

Pair-copula constructions with nonparametric and mixture pair-copulas

G. Weiß[1]

[1] *TU Dortmund University, Faculty of Economics and Social Sciences, Otto-Hahn-Str. 6a, D-44227 Dortmund, Germany; gregor.weiss@tu-dortmund.de*

Abstract. *In this talk, I will present the results of recent studies on the use of (a) nonparametric Bernstein copulas and (b) mixtures of bivariate parametric copulas as building blocks in pair-copula constructions. Both approaches circumvent the error-prone problem of choosing the pair-copulas of a vine model from a set of parametric copulas. We show in both simulations and empirical applications that the proposed models significantly outperform a fully parametric benchmark model from the literature.*

Keywords. *Vine Copulas; Mixture Copulas; Bernstein Copulas.*

4.55

Conditional copula models with multiple covariates

E. Acar

Department of Statistics, University of Manitoba, Canada; elif.acar@umanitoba.ca

Abstract. *Conditional copula models provide a flexible framework to study covariate effects on dependence structures. A number of nonparametric estimation techniques have been recently proposed for these models in the case of a single covariate. These approaches, however, are not directly extendible to, or become impractical in, settings with multiple covariates.*

This talk will present a nonparametric modelling strategy that can accommodate multiple covariates in conditional copula models. We consider a semiparametric conditional copula model where the copula function belongs to a parametric copula family and the copula parameter varies smoothly with covariates. To alleviate the curse of dimensionality, we use an additive formulation of the copula parameter and estimate smooth component functions associated with each covariate via a local likelihood backfitting algorithm. The finite sample performance of the proposed approach will be demonstrated using simulated and real data. The talk will also address general identifiability restrictions and computational challenges.

Keywords. *Additive models; Conditional dependence; Covariate adjustment; Local likelihood.*

4.56

Model selection for copulas for right-censored event times

C. Geerdens[1,*], G. Claeskens[2] and P. Janssen[1]

[1] *Center for Statistics, I-BioStat, Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium; candida.geerdens@uhasselt.be, paul.janssen@uhasselt.be*
[2] *ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; gerda.claeskens@kuleuven.be*
[*] *Corresponding author*

Abstract. *Copulas are used to model the association pattern in grouped time-to-event data. For right-censored event times, we investigate the use of exchangeable as well as nested Archimedean copulas and we contrast them with the Joe-Hu copula family, which consists of mixtures of max-infinitely divisible bivariate copulas. These copulas are fit by a likelihood approach where the vast amount of copula derivatives present in the likelihood is approximated by finite differences.*

Given the amount of possible copula constructions, the question arises how to choose a copula that describes the data well. For right-censored time-to-event data, we give conditions under which a penalized likelihood based model selection criterion is either weakly consistent or consistent. Classical information criteria, like AIC and BIC, are included for appropriate choices of the penalty term.

A set of four-dimensional data on time-to-mastitis (veterinary medicine) is used to demonstrate the developed methodology. In particular we show that, for these data, the model selection criterion prefers the very flexible dependence modeling of the Joe-Hu copula models over the rather restrictive Archimedean copulas.

Keywords. *Multivariate right-censored data; Copulas; Model selection.*

References

Geerdens C., Claeskens G. and Janssen P. (2014). Copula based flexible modeling of associations between clustered event times. *Technical report (submitted for publication).*

4.57

Spline LASSO in high-dimensional linear regression

Bing-Yi JING [1]

[1] *Department of Mathematics. Hong Kong University of Science and Technology*

Abstract. *We consider a linear regression problem in a high dimensional setting where covariates are ordered and the number of covariates p can be much larger than the sample size n. Under sparsity assumptions, we propose a Spline-LASSO approach. It has a number of advantages over its competitor, i.e., the fussed LASSO. First, it can preserve the shape of the parameter values much better. Secondly, it is computationally efficient, as it can be easily modified to use LARS algorithms. Simulations justify our findings. We also mention some possible applications.*

4.58

The HEL: a hybrid empirical likelihood bridging from nonparametrics to parametrics

Nils Lid Hjort[1]

[1] *Department of Mathematics, University of Oslo; nils@math.uio.no*

Abstract. *Suppose data are observed from some distribution and that a certain parameter μ is of interest. The parametric type solution is to employ a suitable family for the underlying unknown distribution, indexed by some θ, leading to a likelihood function $L(\theta)$. Under the model, μ may be expressed as the appropriate $\mu(\theta)$, and inference proceeds in the usual fashion, under or outside model conditions. One may however also form the empirical likelihood $R_n(\mu)$, leading to nonparametric inference for μ; cf. Hjort, McKeague and Van Keilegom (2009). The main idea of this work, which represents joint and ongoing efforts with I. McKeague and I. Van Keilegom, is the construction and analysis of the hybrid empirical likelihood $H(\theta) = L(\theta)^{1-a}R(\mu(\theta))^a$. Here a is a tuning parameter, ranging from full trust in the parametric model ($a = 0$) to the fully nonparametric ($a = 1$). We prove asymptotic normality of the resulting maximum HEL estimator, leading to confidence intervals and confidence distributions, cf. Schweder and Hjort (2014). We also provide illustrations demonstrating that the HEL method may win in efficiency over the nonparametric method even when the parametric model is moderately incorrect, and devise a way of selecting the tuning parameter from data.*

Keywords. *Bridging parametrics and nonparametrics; Efficiency; Empirical likelihood; Hybrid likelihood*

References

Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics* **37**, 1079–1111.

Schweder, T. and Hjort, N.L. (2014). *Confidence, Likelihood, Probability.* Cambridge University Press, Cambridge.

4.59

Testing for uniform stochastic ordering via empirical likelihood

Hammou El Barmi[1] and Ian W. McKeague[2,*]

[1] *Baruch College; hammou.elbarmi@baruch.cuny.edu,*
[2] *Columbia University; im2131@columbia.edu*
* *Corresponding author*

Abstract. *This talk discusses an empirical likelihood approach to testing for the presence of uniform stochastic ordering (or hazard rate ordering) among univariate distributions based on independent random samples from each distribution. The proposed test statistic is formed by integrating a localized empirical likelihood statistic with respect to the empirical distribution of the pooled sample. The asymptotic null distribution of this test statistic is found to have a simple distribution-free representation in terms of standard Brownian motion. The approach is extended to the case of right-censored survival data via multiple imputation. Two applications are discussed: 1) uncensored survival time data of mice exposed to radiation, and 2) right-censored time-to-infection data from a human HIV vaccine trial comparing a placebo group with a vaccine group.*

Keywords. *Order restricted inference; Nonparametric likelihood.*

4.60

Empirical likelihood in high dimensions and confidence sets for functional parameters

S.N. Lahiri* and S. Sahoo

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA; snlahiri@ncsu.edu, ssahoo2@ncsu.edu
* *Corresponding author*

Abstract. *Inference for high dimensional data presents unique challenges. This talk considers some recent developments in Empirical Likelihood (EL) methodology for high dimensional data, where the naive extension of the standard EL is known to fail. The particular inference* *problem considered here is the construction of simultaneous confidence bands for functional parameters using an unbounded number of constraints. The methodology employs a suitable version of the penalized empirical likelihood. Asymptotic distribution of the log empirical likelihood is derived. Finite sample properties of the proposed method are investigated through a moderate simulation study.*

Keywords. *Penalized Empirical Likelihood, Subsampling*

4.61

A consistent jackknife empirical likelihood test for distribution functions

Xiaohui Liu[1], Qihua Wang[2]

[1] *School of Statistics, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China*
[2] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract. *In this paper, an improved jackknife empirical likelihood based approach is developed to test whether the underlying distribution is equal to a specified one. The limiting distributions of the proposed testing statistic are derived under both the null and alterative hypothesis, respectively. It is shown that the proposed test is consistent. Finally, simulation studies are constructed to illustrate the finite sample performance of the proposed test approach.*

Keywords. *Jackknife empirical likelihood; Estimating equations; CramÂ ´er-von Mises test*

4.62

Practical procedures to deal with common support problems in matching estimation

Michael Lechner[1], Anthony Strittmatter[2]

[1] *Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen*
[2] *Department of Applied Econometrics, Albert-Ludwigs-University Freiburg*

Abstract. *We assess the performance of different procedures that deal with common support issues in the context of treatment effects in a selection-on-observables setting. Based on an Empirical Monte Carlo simulation design, we evaluate their small and large sample performance. Without any adjustment, a lack of common support is found to increase the root mean squared error (RMSE) of all investigated parametric and semiparametric estimators. Dropping observations that are off support usually improves their performance. In most simulations, dropping treated observations above a specific propensity score appears to be the best procedure in terms of improvements in RMSE (and the simplicity of its implementation).*

Keywords. *Empirical Monte Carlo Study; matching estimation; common support; outlier; small sample performance*

4.63

The changes-in-changes model with covariates

B. Melly[1,*] and G. Santangelo[2]
[1] Bern University; blaise.melly@vwi.unibe.ch
[2] Universita di Roma La Sapienza; giuliasantangelo@libero.it
*Corresponding author

Abstract. *Differences-in-differences is a quasi-experimental technique used to estimate the effects of a treatment that is not affecting everyone at the same time. Athey and Imbens (2006, AI) suggest an alternative, the changes-in-changes model, that does not depend on the scale of the dependent variable and recovers the whole distribution of the counterfactual outcome. Estimation is relatively straightfoward in the absence of covariates. In the presence of covariates, AI suggest either a nonparametric strategy, which suffers from the curse of dimensionality, or a parametric strategy based on a separability assumption. We suggest a flexible semiparametric estimator that does not impose any separability assumption. We estimate the whole conditional outcome distributions using quantile or distribution regression. We apply the quantile-quantile and probability-probability transformations at the heart of the AI procedure conditionally on the covariates. Finally we integrate the conditional distributions over the empirical distribution of the covariates to obtain unconditional estimates. The conditional estimators satisfy central limit theorems and converge to Gaussian processes. Since all functionals of the conditional distributions are Hadamard differentiable for continuous dependent variables we can derive the asymptotic distribution of the estimator by applying the functional delta method. The validity of the bootstrap also follows.*

Keywords. *Difference-in-differences; Heterogenous treatment effects; Counterfactual distribution; Quantile regression; Distribution regression.*

4.64

The effects of training incidence and duration on labor market transitions

Bernd Fitzenberger[1], Aderonke Osikominu[2,*] and Marie Paul[3]

[1] *University of Freiburg, ZEW, IZA, ROA, IFS; bernd.fitzenberger@vwl.uni-freiburg.de*
[2] *University of Hohenheim, CESifo, IZA; a.osikominu@uni-hohenheim.de*
[3] *University of Duisburg-Essen, RGS Econ; marie.paul@uni-due.de*
* *Corresponding author*

Abstract. *Training programs are an important tool of active labor market policy. Yet, their effectiveness is controversial. A key issue that complicates the evaluation of training programs is methodological. Standard statistical models for treatment evaluation are static. In practice, the decisions whether to enrol and stay in a program are made dynamically. For instance, caseworkers tend to assign training programs to job-seekers who fail to find a job quickly. In* *such a case, a static evaluation is biased towards finding negative effects because unsuccessful job-seekers are over-represented in the treatment group, Fredriksson and Johansson (2008). Building on Robins (1997), we devise an evaluation framework in discrete time that takes the dynamics of program start and duration into account. First, we show how conditionally on time-varying observed covariates and time-constant unobserved heterogeneity causal effects can be identified under no-anticipation and sequential randomization conditions. In a next step, we identify the distribution of the unobserved heterogeneity relying on results for bivariate dynamic discrete choice models, Heckman and Navarro (2007).*

We specify a flexible bivariate random effects model for employment and training status that we estimate with Bayesian MCMC techniques. Based on our estimates we simulate the average effect on the treated as well as the effect of being assigned to programs with differing enrolment lengths. Our results suggest that training improves the employment probability of the participants by 6 to 13 percentage points 2.5 years after program start. Further, participants benefit from being assigned to programs with a longer planned enrolment length.

Keywords. *Dynamic treatment effects; Dynamic nonlinear panel data model; MCMC; Active labor market programs.*

References

Fredriksson, P. and P. Johansson (2008). Dynamic Treatment Assignment: The Consequences for Evaluations Using Observational Data. *Journal of Business and Economic Statistics*, **26**, 435-445.

Heckman, J.J. and S. Navarro (2007). Dynamic Discrete Choice and Dynamic Treatment Effects. *Journal of Econometrics*, **136**, 341-396.

Robins, J. (1997). Causal Inference from Complex Longitudinal Data. In M. Berkane (ed.), *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120)*. Springer. New York, 69-117.

4.65

Modeling heterogeneity by varying coefficient models

S. Sperlich[1]

[1] *Affiliation of first and third author; first.author@email, third.author@email*
[2] *Affiliation of second author; second.author@email*
* *Corresponding author*

Abstract. *During the last two decades the varying coefficient models, the mixed effects models, and Bayesian modeling have been attracting an increasing attention in non- and semiparametric statistics. This was partly in order to bridge the gap between (additive) nonparametric models and the still dominating linear models in empirical economics. There, semiparametric methods found a way out of their shadowy existence only for some prediction issues in finance, and nowadays for matching methods in the treatment effect literature. In the context of the latter, there is an increasing awareness of problems that happen to arise with the identification and* *interpretation of linear coefficients and the quite popular IV estimators: if heterogeneity in returns is present, then IV estimators lose their identification capacity, and linear coefficients estimated by GMM or standard least squares are no longer the average return. Consequently, also the impact evaluation becomes harder - no matter whether ex-post or ex-ante, and the much more work is necessary to make structural model equations 'external valid'. We outline the above mentioned problems to show that (nowadays) standard semiparametric models offer excellent remedies for these problems. These work partly with varying-coefficient models (see Mammen et al. (2013) for a recent review), control functions (see Telser (1964) and Newey et al. (1999)) and LATE-type methods (cf. Imbens and Angrist (1994)).*

Keywords. *Varying-coefficient models; causality analysis; econometrics; semiparametric modeling.*

References

Imbens, G.W. and Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.

Mammen, E., Park, B.U., Lee, Y.K., and Lee, E.R. (2013). Varying-Coefficient Regression Models: A Review and New Developments. *International Statistical Review.*

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

Telser, L. (1964). Iterative Estimation of a Set of Linear Regression Equations. *Journal of the American Statistical Association*, **59**, 845–862.

Authors Index

A

Abadir, Karim M., 38
Acar, Elif, 220
Aguilera, Ana M., 57
Aguilera-Morillo, Carmen, 57
Ahkim, Mohamed, 145
Akritas, Michael, 66, 137
Albert, Mélisande, 29
Alcalá, José Tomás, 19
Allen, Genevera, 168
Alonzo, Todd A., 98
Amini, Arash, 43
Andréas Sperlich, Stefan, 226
Andriyana, Yudhie, 144
Aneiros Pérez, Germán, 184
Anevski, Dragi, 95, 215
Antoniadis, Anestis, 160, 162, 181
Aoshima, Makoto, 92
Arribas, Ana, 83
Asadi, Peiman, 72
Asgharian, Masoud, 115
Ausín, Concepción, 139
Austin, F., 50
Azarang, Leyla, 14
Azzimonti, Dario , 112

B

Bae, Whasoo, 154
Bagkavos, Dimitrios, 41
Bai, Lianyang, 67
Bandyopadhyay, Soutir, 191
Banerjee, Moulinath, 90
Bartolucci, Francesco, 215
Basulto-Elias, Guillermo, 190
Bect, Julien, 112
Ben Hariz, Samir, 151
Berrendero, José R., 12
Bertail, Patrice, 206, 207
Bertrand, Aurélie, 171
Bhattacharya, Sharmodeep, 45
Bickel, Peter J., 43, 45, 101
Birke, Melanie, 111
Boente, Graciela, 90
Boistard, Hélène, 208
Boj del Val, Eva, 150
Bonnini, Stefano, 174
Borrajo, María Isabel, 129
Bouezmarni, Taoufik, 31, 61
Bouret, Yann, 29
Bradic, Jelena, 204
Braekers, Roel, 155
Branscum, Adam, 99
Bravo, Francesco, 138, 190
Breheny, Patrick, 120
Brossat, Xavier, 105, 162, 181
Brown, Lawrence D. , 53
Brunel, Victor-Emmanuel, 149
Bühlmann, Peter, 22

C

Cai, Yun, 21
Camirand, Félix, 31, 61
Cao, Hongyuan, 65
Cao, Ricardo, 130
Carone, Marco, 115
Carroll, Raymond J., 171
Cascos, Ignacio, 135
Chacón, José Enrique, 50, 114, 129
Chang, Jinyuan, 197
Charpentier, Arthur, 62
Chaudhuri, Sanjay, 54
Chautru, Emilie, 206, 207
Chen, Aiyou , 43
Chen, Jia, 122
Chen, Kehui, 83
Chen, Song Xi, 93
Chen, Yuan, 67
Cheng, D., 210
Cheng, Guang, 96
Cherepanov, Oleg, 34
Chevalier, Clément, 112
Cho, Haeran, 105
Chochola, Ondrej, 117
Christmann, Andreas, 89
Chun, Hyonho, 32
Churpek, Matthew M., 65
Ciollaro, Mattia, 185
Cisewski, Jessi, 185
Claeskens, Gerda, 24, 50, 221
Clémençon, Stéphan, 206, 207
Coffman, Donna L., 18
Colling, Benjamin, 143
Conde-Amboage, Mercedes, 26, 27
Corain, Livio, 175
Cornea, Adriana, 38
Corradi, Valentina, 158
Costa, Teresa, 150
Cox, Dennis, 199
Cristóbal, Jose A., 19
Crujeiras, Rosa M., 142
Cuesta-Albertos, Juan A., 56
Cuevas, Antonio, 12, 114, 148
Cugliari, Jairo, 162

D

Dabo-Niang, Sophie, 23, 197
Dafnos, Stavros, 160
Dai, Zhijun, 67
Dalalyan, Arnak, 66
Dalla, Violetta, 70
Datta, Somnath, 117
Davison, Anthony C., 72
de Carvalho, Miguel, 98, 99
de Carvalho, Vanda Inácio, 98, 99
de Haan, Laurens, 173, 174
de Neve, Jan, 126
de Souza, Camila P.E., 104, 105
de Uña Álvarez, Jacobo, 14, 116, 155
del Barrio, Eustasio, 100
Delaigle, Aurore, 49
Delgado, Miguel A, 71
Delicado, Pedro, 83
Delsol, Laurent, 187
Dette, Holger, 164
Devijver, Emilie, 146
Días, Ronaldo, 104

Dobrovidov, Alexander, 39
Douc, Randal, 182, 182
Doukhan, Paul, 118
Drees, Holger, 110, 173
Dudek, Anna, 16, 182
Duembgen, Lutz, 96, 212
Dunson, David, 96
Duong, Tarn, 50
Dutta, Subhajit, 133

E

Early, Jeffrey J., 219
Eberts, Mona, 60
Efimov, Krill, 84
Einmahl, John, 174
El Barmi, Hammou, 223
El Ghouch, Anouar, 14, 115, 157, 172
El Karoui, Noureddine, 136
Engelke, Sebastian, 72
Escabias, Manuel, 57
Escanciano, Juan Carlos, 71
Escribano, Álvaro, 159
Espinosa, Maria Paz, 202

F

Faig, Walter, 204
Farcomeni, Alessio, 215
Febrero-Bande, Manuel, 56
Feng, Gang, 17
Feng, Xingdong, 170
Feridun Turkman, Kamil, 173
Ferreira, Ana, 173
Ferreira, Eva, 203
Feuerverger, Andrey, 30
Fine, Jason P., 65
Fitzenberger, Bernd, 225
Flachaire, Emmanuel, 62
Flenklach, Michael, 78
Fokianos, Konstantinos, 184
Fraiman, Daniel, 88
Fraiman, Ricardo, 88
Francisco-Fernández, Mario, 130
Fraser, Donald A. S., 68
Freeman, Peter, 185
Freitas Lopes, Hedibert, 139
Fremdt, Stefan, 167
Freyermuth, Jean Marc, 50, 103
Frick, Klaus, 209
Fromont, Magalie, 29, 131
Fuentes, Montserrat, 126
Funk, Benedikt, 61
Funtowicz, Silvio, 79

G

Gajecka-Mirek, Elzbieta, 34
Galeano, Pedro, 12
Galeano, Pedro, 139
Gamst, Anthony, 64, 101
García, Nancy, 104
García-Pérez, Alfonso, 31
García-Portugués, Eduardo, 142
Gaudart, Jean, 198
Geaffray, S., 152
Geerdens, Candida, 221
Gel, Yulia R., 47
Gennady, Koshkin, 40
Genovese, Chris, 185
Genschel, Ulrike, 189
Genton, Marc G., 133
Ghosh, Debashis, 18
Giacofci, Madison, 24
Gijbels, Irène, 22, 24, 136, 144, 160
Ginsbourger, David, 112
Giordano, F., 176
Golubev, Yuri, 78
González-Manteiga, Wenceslao, 26, 98, 142
Goude, Yanning, 105, 180, 181
Gu, Hong, 21
Guillas, Serge,13, 23, 198, 199
Guo, Bin, 197

H

Hall, Peter, 177
Hallin, Marc, 140
Handcock, Mark S., 48, 54
Hart, Jeff, 94
Hebiri, Mohamed, 66
Heckman, Nancy E., 104, 105
Hee Lee, Myung, 32
Heesen, Philipp, 131
Hendriks, Harrie, 196
Heuchenne, Cedric, 115
Hiabu, Munir, 154
Hidalgo, Javier, 79
Hlávka, Zdenek, 119
Hlubinka, Daniel, 188
Ho, Qirong, 42
Hofmann, Heike, 189
Hörmann, Siegfried, 166
Horova, Ivana, 128
Hu, T., 158
Hualde, Javier, 72
Huckeman, Stephan, 194
Hudecová, Šárka, 118
Hušková, Marie, 117, 118, 119

I

Iacone, Fabrizio, 72
Ioannides, Dimitris, 41

J

Jacobino, Caroline, 151
Jankowski, Hanna, 49
Jansen, Maarten, 24
Janssen, Arnold, 131
Janssen, Paul, 221
Jentsch, Carsten, 37, 73
Jiménez-Gamero, Maria Dolores, 205
Jin, Lei, 165
Jing, Bing-Yi, 222
Johannes, Jan, 87
Jokiel-Rokita, Alicja , 97
Joseph, E., 12

K

Kaiser, Mark S., 55
Kaishev, Vladimir, 179
Karanasos, Menelaos, 160

Keener, Robert W., 139
Kenney, Tob, 21
Kheifets, Igor, 142
Kherad-Pajouh, Sara, 176
Kidzinski, Łukasz, 166
Kim, Choongrak, 154
Kimm Hyoung-Moon, 205
Kirch, Claudia, 119
Kirschstein, Tomas, 33
Kleijn, Bas, 101
Klopp, Olga, 59
Klutchnikoff, Nicolas, 152
Kneip, Alois, 75, 76, 77
Kohli, Priya, 73
Kokoszka, Piotr, 166
Kolácek, Jan, 128
Kolesnyk, Petro, 212
Kong, Schengchun, 63
Kostakis, Alexandros, 159
Krampe, Jonas, 18
Kreiss,Jens-Peter, 16, 17, 18 , 73
Kristensen, Dennis, 192
Krivobokava, Tatyana, 111
Kyung Lee, Young, 178

L

Lafon, J., 59
Lahiri, S.N., 176, 223
Lai, Ming-Jun, 198
Lam, Clifford, 217
Lambert-Lacroix, Sophie, 160
Lang, Gabriel, 118
Laniado, Henry, 26, 35
Lechner, Michael, 224
Lederer, Johannes, 66
Ledoit, Olivier, 193
Lefieux, Vicent, 179
Legrand, Catherine, 157, 171
Lenzi, Amanda, 104
Lerasle, Matthieu, 131
Leskow, Jacek, 34
Letón, Emilio, 98
Leucht, Anne, 37, 118
Levina, Elizaveta, 43
Levine, Michael, 53
Li, Degui,121, 122, 218
Li, Gaorong, 67
Li, Hongjun, 63
Li, Jun, 93
Liang, Hua, 121, 122
Lid Hjort, Nils, 222
Liebl, Dominik, 77
Liebscher, Steffen, 33
Lillo, Rosa E., 12, 25, 35
Lilly, Jonathan M., 219
Linton, Olivier, 121
Little, Roderick J.A., 139
Liu, Dungang, 69
Liu, Regina, 69
Llop, Pamela, 148
Lockhart, Richard 11
Long, James, 136
López-Blázquez, Fernando, 21
López-de-Ullibarri, L.,20
López-Díaz, Miguel, 113
Lopuhaä, Hendrik, 208
Lopuhaä, Rik., 214
Loubes, Jean-Michel, 123
Louchet, Cecile, 187
Lu, Zudi, 121
Luati, Alessandra, 192
Lyubchich, Vysacheslav , 47

M

Ma, Yanyuan, 169
Maciak, Matús, 149
Magdalinos, Tassos, 159
Mahdizadeh, M., 15
Maistre, Samuel, 186
Maiti, Tapabrata, 202
Mammen, Enno, 154, 178
Mankad, Shawn, 91
Markovich, L., 39
Markovich, Natalia, 107
Marrron, Steve, 125
Martínez Miranda, Mª Dolores, 154
Martínez, Alejandra, 90
Martsynyuk, Yuliya, 137
Matsuda, Yasumasa, 55
Mattielo, Federico, 132
Matzner-Lober, Eric, 161
McKeague, Ian, 223
McMahan Jr., Jerry, 80
Mehdi, Zorgui, 151
Meintanis, Simos, 118, 119
Melly, Blaise, 225
Mémoli, Facundo, 195
Meneses, A, 20.
Meyer, Marco, 73
Meyer, Mery, 110
Mia Hubert, 11, 56
Michailidis, George, 90, 91, 92, 97
Mikosch, Thomas, 51
Miller, Forrest, 200
Miño,Salamanca, 21
Mio, Washington, 195
Mizera, Iván, 149
Molanes-López, Elisa M., 98
Molchanov, Ilya, 112
Mondal, Purba, 128
Moreira, Carla, 155
Mosler, Karl, 81
Moulines, Eric, 59, 182
Mozharovskyi, Pavio, 81
Mukhopadhyay, Subhadeep, 94
Müller, Hans-Georg, 83, 163
Müller, Ursula U.m 201
Munk, Axel, 124, 194, 209

N

Nagy, Stanislav, 134
Nakamura, Miguel, 190
Nan, Bin, 63
Nane, Tina, 214
Nath, Anannya, 128
Naya, Salvador, 20
Neill, James, 200

Neumann, Michael H., 118
Neumeyer, Natalie, 110
Niebuhr, Tobias, 16, 37
Nieto-Reyes, Alicia, 134
Noh, H., 28
Nordman, Dan, 55, 191
Novak, Sergey, 36

O

O'Connell, Ross, 185
Olhede, Sofia, 44, 47, 219
Osikominu, Aderonke, 225
Otneim, Håkon, 52
Otsu, Taisuke, 218
Ottoy, Jean-Pierre, 126
Oviedo de la Fuente, Manuel, 56

P

Packard, Andrew, 78
Paige, Robert, 195
Paindeveine, Davy, 58, 88, 189
Palma, Wilfredo, 102
Panaretos, Victor, 123, 212
Panov, Maxim, 86
Paparoditis, Efstathios, 16, 18, 166, 168
Paraskevopoulos, Alexandros G., 160
Pardo-Fernández, Juan Carlos, 98
Park, AhYeon, 13, 199
Park, Byeong Y., 28, 163, 178
Parrella, María Lucía, 176
Parzen, Emanuel, 94
Pateiro-López, Beatriz, 148
Patil, Popat N., 41
Patilea, Valentin, 85, 156, 186
Patrangenaru, Victor, 195, 196
Paul, Marie , 225
Paulauskas, Vygantas, 108
Pawlak, Miroslaw, 153
Peng, Hanxiang, 200
Peng, Heng, 67
Peng, Yingwei, 157
Pensky, Marianna, 59
Perch Nielsen, Jens, 154, 178
Pérez-Abreu, Víctor, 190
Picek, Jan, 28
Poggi, Jean-Michel, 160, 162, 181
Pokern, Yvo, 199
Politis, Dimitris, 37
Polonik, Wolfgang, 215
Portier, François, 172
Porzio, Giovanni G., 33
Poss, Dominik, 76
Pouet, Christophe, 50
Pourahmadi, Mohsen, 73
Pozzi, Luca, 115
Praškova, Zuzana, 117
Proietti, Tommaso, 192
Pulit, Michal, 97

Q

Qiu, Mingfei, 195, 196
Quessy, Jean-François, 31

R

Racine, Jeffrey, 63
Ragozini, Giancarlo, 33
Rakshit, Suman, 213
Ramírez, Lilia Leticia, 47
Ramírez-Cobo, Pepa, 60
Ramón Flores, 25
Ramsay, James, 75
Raña Míguez, Paula, 184
Rayner, John C.W, 127
Regier, Jeffrey C., 81
Renaud, Olivier, 176
Rendall, Michael, 54
Reyes, Miguel Ángel, 130
Reynaud-Bouret, Patricia, 131
Reynaud-Bouret, Patricia, 29
Rice, John, 136
Rippl, Thomas, 124
Ritov, Yaacov, 101
Rizzo, María L., 51
Rodríguez Poo, Juan, 71, 145
Rodríguez-Casal, Alberto, 129, 147
Rombouts, Jeroen, 62
Romo, Juan, 25, 83
Roueff, François, 183
Rousseeuw, Peter, 11
Ruiz-Gazen, Anne, 208
Ryung Lee, Eun, 28

S

Saavedra-Nieves, Paula, 147
Sahoo, Saswata, 223
Salibian-Barrera, Matías, 90
Salmaso, Luigi, 175
Salmon, Joseph, 59
Saltelli, Andrea, 79
Sánchez-Sellero, César, 26, 27
Sankar Dhar, Subhra, 128, 187
Santangelo, Giulia, 225
Sapatinas, Theofanis, 166, 168
Sarda, Pascal, 76
Sardy, Sylvain, 151, 163
Scharfstein, Daniel, 115
Schick, Anton, 200
Schimek, Michael G, 177
Schneider, Ulrike, 125
Schwartzman, Armin, 210
Schwarz, Katja, 111
Schweinberger, Michael, 48
Scolas, Sylvie, 157
Secchi, Piercesare, 82
Segaert. P, 11
Selk, Leonie, 110
Sen, Pranab K., 213
Seo, Myung, 218
Serfling, Robert, 141
Shaddick, Gavin, 106
Shmileva, Elena, 108
Shojaie, Ali, 91
Sieling, Hannes, 209
Silvapulle, Mervyn , 213
Sim, T., 183
Simakhin, V. A., 34

Šiman, Miroslav , 188
Simons, Frederik J., 219
Smagin, Valery, 40
Smith, Ralph, 80
Soberón, Alexandra, 71, 145
Sommerfeld, Max, 194
Spokoiny, Vladimir, 84, 85, 86
Srivastava, Anuj, 211
Stamatogiannis, Michalis, 159
Stanberry, Larissa, 49
Stanfill, Bryan, 189
Stark, Philip B., 81
Stehlik, Milan, 109
Steinebach, Josef, 117
Steinwart, Ingo, 60
Steland, Ansgar, 39
Strittmatter, Anthony, 224
Strzalkowska-Kominiak, Ewa, 15
Sturm, A., 124
Stute, Winfried, 14, 203
Subba Rao, Suhasini, 74
Sucarrat, Genaro, 159
Svendova, Vendula, 177
Sykulski, Adam M., 219
Székely, Gábor J., 51

T

Talamakrouni, Majda, 14
Tao, Wenwen, 163
Tarrio-Saavedra, J., 20
Taylor, Jonathan 11
Tecuapetla-Gómez, Inder, 102
Ternynck, Camille, 23
Thas, Olivier, 126, 127, 132
Thouvenot, Vicent, 181
Tibshirani, Robert, 11
Tibshirani, Ryan, 11
Tjoestheim, Dag Bjarne, 52, 53
Tone, Cristina, 100
Torrecilla, José Luis, 12
Torres, Raúl, 35
Tsodikov, Alex, 172
Tucker, J.Derek, 211

V

Van Bever, Germain, 58, 189
Van der Laan, Mark, 115
Van Hala, Matthew, 191
Van Keilegom,Ingrid, 14, 115, 142, 143, 156, 171, 172
Velasco, Carlos, 142
Veraverbeke, Noël, 116
Verbist, Bie, 132
Verdebout, Thomas, 88
Verhasselt, Anneleen, 144, 145
Vidakovic, Brani, 60
Viennet, Emmanuel, 46
Vilar, Juan, 184
Vimond, Myriam, 152
Virbickaite, Audrone, 139
Vogt, Michael, 164
Voorman, Arend, 91
Vrinssen, Inge, 136

W

Wagner, Heiko, 77
Wahl, Martin, 164
Wand, Matthew P.,50
Wang, Haiyan, 67, 165
Wang, Jiantian, 26
Wang, Lie, 53
Wang, Lifeng, 67
Wang, Qihua, 224
Wang, Suojin, 122, 165
Wang, Y., 101
Wasserman, Larry, 185
Weiss, Gregor, 220
Wellner, Jon A., 96
Wilke, Ralf, 212
Witten, Daniela, 91
Wolf, Michael, 193
Wolfe, Patrick J., 44, 45, 47
Wood, A.T.A, 41
Wu, Wei, 211
Wu, Yuanshan, 170

X

Xia, Yingeun, 218
Xiang, Liming, 158
Xiaohui Liu, 224
Xie, Jun, 120
Xie, Min-ge, 69
Xing, Eric, 42
Xu, Ronghui, 204
Xue, Lingzhou, 216

Y

Yajima, Yoshi, 55
Yang, Jingjing, 199
Yang, Seon J., 115
Yang, Yun, 96
Yao, Qwei, 105, 197
Yata, Kazuyoshi, 92
Yi, Grace, 216
Yin, Guosheng, 170
Yin, Junming, 42
Yuan, Zheming, 67
Yudovina, Elena, 90

Z

Zambom, Adriano Z., 137
Zangeneh, Sahar Z., 139
Zeng, Donglin, 65
Zhan, D., 94
Zhang, Hao, 101
Zhang, Jun, 67
Zhang, Wenyang, 218
Zhao, Yuwei, 51
Zhivotovskiy, Nikita, 85
Zhongi, Ping-Shou, 93
Zhou Zhou, 104
Zhou, Chen, 174
Zhou, Ding-Xuan, 89
Zhou, T., 169
Zhu, Hongxiao, 199
Zhu, Liping, 169, 170
Zhu, Lixing, 67
Zhu, Yeying, 18