A.Daouia

Title: On Projection-Type Estimators of Multivariate Isotonic Functions

Authors:
* Abdelaati Daouia
Institute of Statistics, Catholic University of Louvain,
Belgium (abdelaati.daouia@uclouvain.be)

** Byeong U. Park
Department of Statistics, Seoul National University, South Korea
(bupark@stats.snu.ac.kr)

Abstract:
Let M be an isotonic real-valued function on a compact subset of $\mathbb{R}^d$ and let $\hat{M}_n$ be an unconstrained estimator of M. A feasible monotonizing technique is to take the largest (smallest) monotone function that lies below (above) the estimator $\hat{M}_n$ or any convex combination of these two envelope estimators. When the process $r_n(\hat{M}_n - M)$ is asymptotically equicontinuous for some sequence $r_n\to\infty$, we show that these projected estimators are r_n-equivalent in probability to the original unrestricted estimator.
Our first motivating application involves a monotone estimator of the conditional distribution function that has the distributional properties of the local linear regression estimator. Applications also include the estimation of econometric (probability-weighted moment, quantile) and biometric (mean remaining lifetime) functions.

Adityanand Guntuboyina

Title: Planar Convex Set Estimation from Support Function Measurements

Abstract: Motivated by an application to reflective tomography, we study the problem of estimating a convex set in the plane from finitely many noisy measurements of it support function. We present a solution that is new, simple and theoretically optimal.

Bounds on expected generalized order statistics

Agnieszka Goroncy
Nicolaus Copernicus University, Poland

Abstract

Ordered random variables are widely used in statistical models and inference. They have many different interpretations and interesting applications, e.g., in reliability theory.

Generalized order statistics have been introduced by Kamps (*J. Statist. Plann. Inference* 48: 1-23, 1995) and they serve as a unifying approach for several models of increasingly ordered random variables, e.g., order statistics, records, k-th records, sequential order statistics or progressively Type-II censored order statistics.

We establish the upper nonpositive and all the lower bounds on the expectations of generalized order statistics based on a given distribution function with the finite mean and central absolute moment of a fixed order. We also describe the conditions for which the bounds are attained.

The methods of deriving the lower nonpositive (upper nonnegative) and lower nonnegative (upper nonpositive) bounds are totally different. The first one, the greatest convex minorant method is the combination of the Moriguti and well-known Hölder inequalities and the latter one is based on the maximization of some norm on the properly chosen convex set.

The paper completes the results of Cramer et al. (*Appl. Math.* 29: 285-295, 2002).

Functional projection pursuit regression

F. Ferraty[(a)], A. Goia[(b)]*, E. Salinelli[(b)] and P. Vieu[(a)]

[(a)] Institut de Mathématiques de Toulouse, Équipe LSP, Université Paul Sabatier
118, route de Narbonne, F-31062 Toulouse Cedex, France

[(b)] Dipartimento di Studi per l'Economia e l'Impresa,
Università del Piemonte Orientale, Via Perrone, 18, 28100 Novara, Italy

Abstract

We introduce a flexible approach to approximate the regression function in the case of a functional predictor and a scalar response. Following the Projection Pursuit Regression principle, we derive an additive decomposition which exploits the most interesting projections of the prediction variable to explain the response. On the one hand this approach allows to avoid the well-known curse of dimensionality problem and on the other hand can be used as an exploratory tool for the analysis of functional dataset. The terms of such decomposition are estimated with an alternating optimization strategy combining a spline approximation and the one-dimensional Nadaraya-Watson kernel regression estimate. The good behaviour of our procedure is illustrated from theoretical and practical points of view. Asymptotic results state that the terms in the additive decomposition can be estimated without suffering from the dimensionality problem, while some applications to real and simulated data show the high predictive performance of our method.

References

[1] Ferraty F., Goia A., Salinelli E. and Vieu P. (2012). Functional projection pursuit regression. *Preprint*.
[2] Ferraty F. and Vieu P. (2006). *Nonparametric functional data analysis*, Springer, New York.
[3] Friedman J.H., Stuetzle W., (1981). Projection Pursuit Regression, *Journal of the American Statistical Association*, 76, 817–823.
[4] Hall P. (1989). On projection Pursuit Regression, *The Annals of Statistics* **17**, 2, 573–588.
[5] James G.M. and Silverman B.W. (2005). Functional Adaptive Model Estimation, *Journal of the American Statistical Association*, **100**, 470, 565–576.
[6] Ramsay J.O. and Silverman B.W. (2005). *Functional Data Analysis*, 2nd edn., Springer-Verlag, New York.

Estimation in nonparametric location-scale models under dependent censoring

Aleksandar Sujica, Ingrid Van Keilegom

Institut de statistique, biostatistique et sciences actuarielles
Université catholique de Louvain
Voie du Roman Pays 20
1348 Louvain-la-Neuve
Belgium

Abstract

A common assumption when working with randomly right censored data, is the independence between the variable of interest Y (the "survival time") and the censoring variable C. This assumption, which is not testable, is however unrealistic in certain situations. In this paper we assume that for a given X, the dependence between the variables Y and C is described via a known copula. Additionally we assume that Y is the response variable of a heteroscedastic regression model $Y = m(X) + \sigma(X)\varepsilon$, where the explanatory variable X is independent of the error term ε, and the functions m and σ are 'smooth'. We propose an estimator of the conditional distribution of Y given X under this model, and show the asymptotic normality of this estimator. We also study the small sample performance of the estimator, and discuss the advantages/drawbacks of this estimator with respect to competing estimators.

The Variance Profile

Alessandra Luati*, Tommaso Proietti and Marco Reale

*University of Bologna, Department of Statistics

Abstract

The variance profile is defined as the power mean of the spectral density function of a stationary stochastic process. It is a continuous and non-decreasing function of the power parameter, p, which returns the minimum of the spectrum ($p \to -\infty$), the interpolation error variance (harmonic mean, $p = -1$), the prediction error variance (geometric mean, $p = 0$), the unconditional variance (arithmetic mean, $p = 1$) and the maximum of the spectrum ($p \to \infty$). The variance profile provides a useful characterisation of a stochastic process; we focus in particular on the class of fractionally integrated processes. Moreover, it enables a direct and immediate derivation of the Szegö-Kolmogorov formula and the interpolation error variance formula. The paper proposes a non-parametric estimator of the variance profile based on the power mean of the smoothed sample spectrum, and proves its consistency and its asymptotic normality. From the empirical standpoint, we propose and illustrate the use of the variance profile for estimating the long memory parameter in climatological and financial time series and for assessing structural change.

Keywords: Predictability; Interpolation; Non-parametric spectral estimation; Long memory.

Costationarity of Locally Stationary Time Series

A. Cardinali and G. P. Nason
University of Bristol

Two locally stationary time series are said to be costationary if there is exists a (usually time-varying) linear combination that is second-order stationary. It is trivial to obtain stationarity if the linear combinations are arbitrary, so we constrain the complexity of the linear combinations to obtain non-degenerate and interpretable combinations. The set of locally stationary processes is closed with respect to taking time-varying linear combinations, hence it is meaningful to enquire about their stationarity or lack of.

Costationary series imply a error-correction type of formula in which changes in the variance of one series are responded by simultaneous, balancing, changes in the other. We explain how we detect costationarity in practice using bootstrap tests of stationarity and multivariate techniques to determine common solution sets. The existence of costationarity between time series implies an interesting and interpretable relationship between the series.

There are four main reasons why discovering costationarity is important: (i) learning/discovery of any costationary relationship itself, (ii) estimating the strength of any such relationship, (iii) using the derived stationary series, in some applications in preference to either of the original series and (iv) using the relationship to learn about, say, X_t from data on Y_t or vice versa.

Costationary series possess a 'variance-correction' interpretation where series 'self-correct' if their current linear combination is to maintain stationarity. This is akin to the 'error-correction' model in cointegration theory. We also show that this implies bounds on the time-varying local covariance.

For costationary series a formula exists that permits the time-varying local covariance to be estimated indirectly in terms of the induced stationary series variance, and each series' time-varying variance. If multiple costationary solutions are available, we show that an 'indirect' local covariance estimator is available, which attains a greater efficiency rate in comparison to 'direct' estimators based on the local cross-periodogram. We illustrate this point through a simulation example. In addition, for the FTSE and SP500 series we show that our indirect estimate(s) compare favorably to the direct estimate.

To determine costationarity (or lack of) we adopted a projection pursuit type algorithm in combination with a bootstrap stationarity test. Our method searches for the simplest time-varying linear combinations of two (or more) series that results in a stationary series. For the bootstrap stationarity test we demonstrate uniform consistency for the power function and good empirical performances under very mild distributional assumptions.

We illustrate further the applicability of our methodology with one example of financial asset allocation. A mean-variance portfolio with (co)stationary factors is constructed using market index data and is shown to have superior Sharpe ratios to two established portfolio selectors.

Stable Nonparametric Signal Filtration in Nonlinear Models

A.V. Dobrovidov, K.K. Klyuchnikov,
Russian Academy of Sciences, Moscow, Russia

Abstract

A stationary bivariate markovian process $(X_n, S_n), n \geqslant 1$ is considered with the first component observable and the second one non-observable. The problem of filtering a usefull stochastic signal $(S_n), n \geqslant 1$ from the mixture with noise by observations $X_1^n = X_1, \cdots, X_n$ is solved under conditions of nonparametric uncertainty. This means that probabilistic parametric model of the useful signal (S_n) is assumed to be completely unknown. With such assumptions it is not possible to build an optimal Bayesian estimation in general case. Nevertheless for the particular class of observation models where conditional density $f(x_n|s_n, x_1^{n-1})$ belongs to conditionally-exponential family of distribution Bayesian estimation becomes the solution of some nonrecurrent equation which only depends on probabilistic characteristics of observable process (X_n). These unknown characteristics can be restored by observations X_1^n by using stable nonparametric estimation procedures adapted for dependent observations. In this paper nonlinear multiplicative observation model with nongaussian noise is considered in details. Numeric results show that a quality of nonparametric estimation built for nonlinear observation model turn out to be worse than a quality of Bayesian estimation but better than a quality of optimal linear estimation. The crucial step in the construction of stable non-parametric procedures is the choice of proper parameters of smoothing and regularization. In this paper we propose an optimal choice of these parameters which leads to automatic algorithms of non-parametric filtration.

Asymptotic equivalence of functional linear regression and a white noise inverse problem

Alexander Meister

Institut für Mathematik
Universität Rostock
D-18051 Rostock, Germany
email: alexander.meister@uni-rostock.de

We consider the statistical experiment of functional linear regression (FLR). Furthermore, we introduce a white noise model where one observes an Ito process, which contains the covariance operator of the corresponding FLR model in its construction. We prove asymptotic equivalence of FLR and this white noise model in LeCam's sense under known design distribution. Moreover, we show equivalence of FLR and an empirical version of the white noise model for finite sample sizes. As an application, we derive sharp minimax constants in the FLR model which are still valid in the case of unknown design distribution. This talk is mainly based on the following paper:

Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.* **39**, 1471–1495.

Flexible Multivariate Tolerance Zones based on Robust Quantization Techniques

L.A. García-Escudero, A. Gordaliza and A. Mayo-Iscar
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid. Spain

Butler (1982) introduced a technique for obtaining asymptotic distribution free tolerance intervals for real-valued random variables. These tolerance intervals are based on the non-discarded observations corresponding to the computation of the well-known Rousseeuw's Least Trimmed Squares location estimator. Later on, Butler et al.(1993) extended this technique to the multivariate setting for obtaining asymptotic distribution free tolerance ellipsoids based on the Minimum Covariance Determinant estimator.

In a similar fashion, tolerance zones can be obtained from the non-trimmed subset of observations after applying the trimmed k-means methodology. Trimmed k-means method was introduced in Cuesta-Albertos et al. (1997) with the aim of robustifying the classical k-means method and it has become a reference tool in Cluster Analysis (sample viewpoint) and in Quantization (population viewpoint) when outliers and background noise are expected to be present.

The proposed tolerance zones arise from trying to summarize the whole distribution through the use of a set of quantizers instead of merely considering summary functionals like the mean and scatter matrix estimators. The use of large values of k provides very flexible and adaptive tolerance zones that turn out to be very useful in complex problems in Statistical Quality Control.

References

[1] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997), "Trimmed k-means: An attempt to robustify quantizers," *The Annals of Statistics,* **25**, 553-576.

[2] Butler, R.W. (1982), Nonparametric tolerance interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann. Statist.* **10**, 197-204.

[3] Butler, R.W., Davies, P.L., and Jhun, M. (1993). Asymptotics for the minimum covarince determinant estimator. *Ann. Statist.* **21** 1385-1400.

Variable selection in high dimensional regression when variables are heavily correlated

Alois Kneip
University of Bonn

Joint work wth Pascal Sarda, Université Paul Sabatier, Toulouse

Abstract:
The talk considers linear regression problems where the number of predictor variables is possibly larger than the sample size. The basic motivation of the study is to combine the points of view of model selection and functional regression.

Model selection procedures like Lasso can be used if the predictor vector can be decomposed into a sum of two independent random components reflecting common factors and specific variabilities of the explanatory variables. However, the usual assumption of sparseness of coefficients is restrictive in this context. Common factors may possess a significant influence on the response variable which cannot be captured by the specific effects of a small number of individual variables. We therefore propose to include principal components as additional explanatory variables in an augmented regression model. We give finite sample inequalities for estimates of these components. It is then shown that model selection procedures can be used to estimate the parameters of the augmented model and we state theoretical properties of the estimators.

In a second part of the talk functional explanatory variables are considered. We study a generalization of the classical functional linear regression model. It is assumed that there exists an unknown number of "points of impact", i.e. a number of discrete observation times where the corresponding functional values possess significant influence on the response variable. Problems of identifiability and corresponding estimation procedures are discussed.

Pairwise Dynamic Time Warping for Event Data

Ana Arribas-Gil[1] and Hans-Georg Müller[2]

[1] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain

[2] Department of Statistics, University of California, Davis
Davis, CA 95616 USA

ABSTRACT

We introduce a new version of dynamic time warping for samples of observed event times that are modeled as time-warped intensity processes. We assume that for each experimental unit or subject within a sample of units or subjects, one observes a random number of event times locations. Since the number of observed events differs from subject to subject, usual landmark alignment methods which require the number of events to be the same across subjects are not feasible. We address this problem by applying dynamic time warping to align event times, initially by aligning the event times for pairs of subjects, regardless of whether the numbers of observed events within the considered pair of subjects match. We then utilize the information about pairwise alignments to extract an overall alignment of the events for each subject across the entire sample of subjects. This overall alignment provides a useful description of event data and can be used for subsequent analysis. The method is illustrated with a historical fertility study and on-line auction data.

AndrØMas

Title : Inverse problems aspects in the functional linear regression

Abstract :
The functional linear regression model with functional ouptuts is introduced. The estimation procedure comes down to solving in a Hilbert space a special inverse problem whose features will be highlighted. Two results are derived : the exact asymptotic risk and a CLT for the predictor. A surprising negative result is also proved : the (minimax) solution of the inverse problem cannot converge in distribution for the usual norm.

An M-Estimator of Tail Dependence in Arbitrary Dimensions

Andrea Krajina (speaker), *Göttingen University, Germany*, Andrea.Krajina@mathematik.uni-goettingen.de

John H.J. Einmahl, *Tilburg University, The Netherlands*
Johan Segers, *Université catholique de Louvain, Belgium*

Consider a random sample in the max-domain of attraction of a multivariate extreme value distribution. Under the assumption that the tail dependence structure (here modelled by the stable tail dependence function) belongs to a parametric model, we propose an M-estimator of the unknown parameter. The estimator is defined as the value of the parameter vector that minimizes the distance between a vector of weighted integrals of the stable tail dependence function and empirical counterparts of these integrals. Under minimal conditions, this minimization problem has (with probability tending to one) a unique, global solution. The estimator is consistent and asymptotically normal. The asymptotic behaviour of the estimator relies on the asymptotic normality of the nonparametric estimator of the stable tail dependence function in arbitrary dimensions, which we also show. Since no assumptions on the differentiability of the stable tail dependence function are made, the method can be used for discrete models as well.

Using machine learning algorithms for sufficient dimension reduction

Andreas Artemiou

Abstract

Recently, Principal Support Vector Machine (PSVM) was introduced as a new method for sufficient dimension reduction (SDR). The main advantage of PSVM is that it can be used for both linear and nonlinear sufficient dimension reduction under a unified framework. The basic idea is to divide the response variables into slices and use a modified form of support vector machine to find the optimal hyperplanes that separate them. These optimal hyperplanes are then aligned by the principal components of their normal vectors.

In this talk, we extend this idea by using different SVM algorithms, for example L2SVM. We introduce Principal L2 SVM (PL2SVM) and we compare its performance to PSVM. More importantly, we present the asymptotic for the gradient function, the Hessian matrix and the influence function. The simpler format of these three results, make PL2SVM more attractive for the development of inferential tools, i.e. the sequential tests to determine the dimension of the Central Dimension Reduction Subspace.

Title: Valid Post-Selection Inference

Speaker: Andreas Buja

Affiliation: Wharton, UPenn

Joint with: Richard Berk, Larry Brown, Kai Zhang, Linda Zhao

Abstract:

It is common practice in statistical data analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid ‘‘post-selection inference” by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels.
By purchasing ‘‘simultaneity insurance” for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. Importantly the inference does not depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models.

Partial Identification, Random Sets and Instrumental Variable Models[1]

Andrew Chesher and Adam Rosen

University College London

First Conference of the International Society for NonParametric Statistics
Chalkidiki, Greece, June 15th - 19th 2012

Abstract

We use results from the theory of random sets to characterise the identifying power of a rich class of models in which observed and latent variables can be scalar or multi-dimensional and discrete or continuous along different dimensions.

The models place restrictions on the structural relationships that deliver values of endogenous variables given values of observed and unobserved exogenous variables and on the co-variation of observed and unobserved exogenous variables.

Structural relationships may be incomplete in the sense that they can deliver non-singleton sets of values of endogenous outcomes. Such cases arise in models of games, auctions and other strategic interactions with multiple equilibria and in nonparametric instrumental variable models.

One application is covered in detail in the paper "An Instrumental Variable Model of Multiple Discrete Choice", Andrew Chesher, Adam Rosen and Konrad Smolinski, CeMMAP Working Paper 39/11, http://cemmap.ac.uk/wps/cwp3911.pdf.

[1]Financial support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and from the European Research Council (ERC) grant ERC-2009-StG-240910-ROMETA is gratefully acknowledged.

Dating Medieval English Charters

Andrey Feuerverger

University of Toronto

Deeds (or charters) dealing with property rights, provide important documentation for historians studying the evolution of social, economic and political changes. However, at least one million charters (written in Latin) dating from the tenth through fourteenth centuries in England were left undated. Correctly dating these documents is vital for the study of English medieval history. We consider the statistical problem of computer-automated methods for dating such document collections with the goal of reducing the considerable efforts required to date them manually. This problem turns out to encompass a wide variety of statistical applications, questions, and techniques.

QUANTILE REGRESSION IN VARYING-COEFFICIENT MODELS USING P-SPLINES

Y. Andriyana[1], I. Gijbels[1], and A. Verhasselt[2]

[1] Department of Mathematics and Leuven Statistics Research Center (LStat), Katholieke Universiteit Leuven, Belgium
[2] Department of Mathematics and Computer Science, Universiteit Antwerpen, Belgium

Abstract. As a generalization of median regression, quantile regression has been widely used in statistical modeling. Meanwhile, varying-coefficient models (VCM) have been developed as an important tool to create a flexible model since the regression coefficients in VCM change smoothly with the value of other variables such as 'time'. In this work we are interested in quantile regression for varying-coefficient models applied to longitudinal observations. We approximate each coefficient function by means of P-Splines. The limiting distributions of the estimators are derived under some regularity conditions. In the statistical estimation procedure the optimization problem is transformed into a linear programming problem for which we then use the *Frisch-Newton* interior point method. The performance of the estimation method is investigated in a simulation study.

Keywords and phrases: Quantile regression, varying-coefficient models, P-splines, *Frisch-Newton* interior point method, flexible modeling.

Anestis Antoniadis

Smoothing and variable selection using P-splines

In this talk we focus on nonparametric penalized estimation in additive varying coefficient models. The feature of varying coefficient models, a very useful extension of linear models, is to allow the coefficients varying, which can be used to explore the dynamics of the impacts of the covariates on the response variable. A main interest is also on variable selection for such models. Particular attention goes to recent variable selection procedures such as grouped Lasso, grouped SCAD, COSSO, but also to the nonnegative garrote method introduced originally for variable selection in a multiple linear regression model. We show how the latter method combined with P-splines estimation leads to an estimation and variable consistent method in the settings of varying coefficient models. The performances of this and other related selection procedures are investigated in a simulation study and illustrations on real data examples are provided.

This talk is based on joint works with Irene Gijbels, Sophie Lambert-Lacroix and Anneleen Verhasselt.

We derive the Maximum Likelihood estimator of an unknown probability mass function, for which the correct labeling is unknown, based on frequency counts. We show that the ML estimator is almost surely consistent in L^1-norm, and derive rates that are uniform over certain classes of probability mass functions.

A Non-Parametric Causality Test:
Detection of direct causal effects in multivariate systems using Partial Corrected Transfer Entropy

Angeliki Papana[1], Dimitris Kugiumtzis[2], Katerina Kyrtsou[1, 3]

[1]University of Macedonia, [2]Aristotle University of Thessaloniki, [3]University of Strasbourg, BETA, University of Paris 10, Economix, ISC-Paris, Ile-de-France

Abstract
In a recent work we proposed the corrected transfer entropy (CTE), which reduces the bias in the estimation of transfer entropy (TE), a measure of Granger causality for bivariate time series making use of the conditional mutual information. Here, we extend this correction to the partial transfer entropy (PTE), the modification of TE to account for the presence of other time series when quantifying Granger causality through conditional mutual information. For the estimation of the Corrected Partial Transfer Entropy (CPTE) time shifted surrogates are used in order to quantify and correct the bias, and the estimation of the involved entropies of high-dimensional variables is made with the method of k-nearest neighbors. CPTE is evaluated on a coupled stochastic system with both linear and nonlinear interrelations and a nonlinear coupled deterministic system. Finally, we apply CPTE to financial data and investigate whether we can detect direct effects among some financial indices.

Generalized seasonal block bootstrap for time series

Anna Dudek
Dept. of Applied Mathematics
AGH University of Science and Technology
al. Mickiewicza 30, 30-059
Cracow, Poland
email: aedudek@agh.edu.pl

Jacek Leśkow
Department of Econometrics
WSB-NLU, Nowy Sącz, Poland

Efstathios Paparoditis
Dept. of Mathematics and Statistics
University of Cyprus
P.O.Box 20537
CY 1678 Nicosia, Cyprus

Dimitris N. Politis
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, USA

March 26, 2012

The new block bootstrap method for seasonal time series will be introduced. It is the generalized version of Periodic Block Bootstrap (Chan et. al. (2004)) and Seasonal Block Bootstrap (Politis (2001)). The theorem establishing consistency for the global and seasonal means will be presented. Moreover, the simulation study results will be discussed.

References

[1] Chan, V., Lahiri, S.N., and Meeker, W.Q (2004). Block bootstrap estimation of the distribution of cumulative outdoor degradation, *Technometrics*, 46, 215-224.

[2] Leśkow, J. and Synowiecki, R. (2007). Consistency and application of moving block bootstrap for nonstationary time series with periodic and almost periodic structure, *Bernoulli,* 13(4), 1151-1178.

[3] Leśkow, J. and Synowiecki, R. (2010). On bootstrapping periodic random arrays with increasing period, *Metrika*, 71, 253-279.

[4] Politis, D.N. (2001), Resampling time series with seasonal components, in *Frontiers in Data Mining and Bioinformatics: Proceedings of the 33rd Symposium on the Interface of Computing Science and Statistics*, Orange County, California, June 13-17, 2001, 619-621.

Dependent wild bootstrap for degenerate U-statistics

Anne Leucht[1] **and Michael H. Neumann**[2]

[1] *Department Mathematik, Universität Hamburg*
[2] *Institut für Stochastik, Friedrich-Schiller-Universität Jena*

U- and related von Mises- (V-)statistics play an important role in mathematical statistics. In the case of hypothesis testing, major interest is on degenerate statistics of this type since numerous test quantities can be reformulated as or approximated by statistics of this type under the null hypothesis. Well-known examples are the Cramér-von Mises and the χ^2 statistics.

Especially in the case of dependent random variables, the distribution of such a statistic as well as its asymptotics have quite involved forms and depend on characteristics of the underlying process in a complicated manner. For the determination of quantiles, we therefore propose new versions of a model-free bootstrap method that can be viewed as variants of the dependent wild bootstrap recently proposed by Shao (2010) for smooth functions of the mean. Here, we do not directly bootstrap the underlying random variables but the summands of the U-statistics. In order to verify the validity of our procedures, asymptotic theory for the original and the bootstrap statistics is derived under simple and easily verifiable conditions.

REFERENCES

Shao, X. (2010). The dependent wild bootstrap. *J. Amer. Statist. Assoc. Theory and Methods* **105**, 218–235.

Anne Vanhems
Professor of Statistics and Econometrics at Toulouse Business School, France
Researcher at Toulouse School of Economics (TSE)

Personal webpage: https://sites.google.com/site/vanhemsa/home

Semi parametric transformation model with endogeneity: a control function approach

We consider a semi parametric transformation model, in which the regression function has an additive nonparametric structure and the transformation of the response is assumed to belong to some parametric family. We suppose that endogeneity is present in the explanatory variables. Using a control function approach, we show that the proposed model is identified under suitable assumptions, and propose a profile likelihood estimation method for the transformation. The proposed estimator is shown to be asymptotically normal under certain regularity conditions.

P-spline estimation in generalized varying coefficient models

A. Verhasselt

Universiteit Antwerpen, Department of Mathematics and Computer Science

Abstract

We consider nonparametric smoothing and variable selection in generalized varying coefficient models. Generalized varying coefficient models are commonly used for analyzing the time-dependent effects of covariates on responses, which are not necessary continuous, but for example counts or categories. We present the P-spline estimator in this context and show its estimation consistency for a diverging number of knots, by using an approximation of the link function. The combination of P-splines with nonnegative garrote (which is a variable selection method) leads to a good estimation and variable selection technique.

Conditional Inference Functions for Mixed-Effects Models with Unspecified Random-Effects Distribution

Annie Qu, UIUC

Abstract: In longitudinal studies, mixed-effects models are important for addressing subject-specific effects. However, most existing approaches assume a normal distribution for the random effects, and this could affect the bias and efficiency of the fixed-effects estimator.
Even in cases where the estimation of the fixed effects is robust with a misspecified distribution of the random effects, the estimation of the random effects could be invalid. We propose a new approach to estimate fixed and random effects using conditional quadratic inference functions. The new approach does not require the specification of likelihood functions or a normality assumption for random effects. It can also accommodate serial correlation between observations within the same cluster, in addition to mixed-effects modeling. Other advantages include not requiring the estimation of the unknown variance components associated with the random effects, or the nuisance parameters associated with the working correlations. We establish asymptotic results for the fixed-effect parameter estimators which do not rely on the consistency of the random-effect estimators. Real data examples and simulations are used to compare the new approach with the penalized quasi-likelihood approach, and SAS GLIMMIX and nonlinear mixed effects model (NLMIXED) procedures. This is joint work with Peng Wang and Cindy Tsai.

Sequential Cross-Validated Bandwidth Selection Under Dependence and Anscombe-Type Extensions to Random Time Horizons

Ansgar Steland
Institute of Statistics
RWTH Aachen University

To detect changes in the mean of a time series, one may use previsible detection procedures based on nonparametric kernel prediction smoothers that cover various classic detection statistics as special cases. Bandwidth selection, particularly in a data-adaptive way, is a serious issue and not well studied for detection problems. To ensure data adaptation, we select the bandwidth by cross-validation, but in a sequential way leading to a functional estimation approach.

In this talk, asymptotic theory for the method under fairly weak assumptions on the dependence structure of the error terms, particularly covering GARCH(p, q) processes, is established in the sense of (sequential) functional central limit theorems for the cross-validation objective function and the associated bandwidth selector. Our gradual change-point model covers multiple change-points in that it allows for a nonlinear regression function after the first change-point possibly with further jumps and Lipschitz continuous between those discontinuities. The proof is based on Kurtz and Protter (1996)'s results on the weak convergence of Itô integrals and a diagonal argument.

In applications, the time horizon where monitoring stops latest is often determined by a random experiment, e.g. a first-exit stopping time applied to a cumulated cost process or a risk measure, possibly stochastically dependent from the monitored time series. Thus, we also study that case and establish related limit theorems in the spirit of Anscombe (1952)'s result. The result has various applications including statistical parameter estimation and monitoring financial investment strategies with risk-controlled early termination.

Anthony Gamst

"On the Bernstein-von Mises Theorem for Conditionally Parametric Models"

Inference in the symmetric location model: An empirical likelihood approach

Anton Schick
Department of Mathematical Sciences, Binghamton University

An empirical likelihood approach with an increasing number of estimated constraints is developed for inference in the symmetric location model. For the resulting empirical likelihood we obtain a Wilks' Theorem and a uniform local asymptotic normality condition. The former is used to obtain confidence regions and tests for the center of symmetry. The latter is used to show that the maximum empirical likelihood approach yields an asymptotically normal, semiparametrically efficient and adaptive estimator of the center of symmetry and to derive a goodness of fit test for symmetry.

Nonparametric priors for Bayesian inference with partially exchangeable data

Antonio Lijoi

Department of Economics and Business, University of Pavia, Italy

The proposal of models that can accommodate for more general forms of dependence than exchangeability has recently been the focus of a large body of literature in Bayesian nonparametric statistics. In this talk we will present an approach aiming at the construction of vectors of random probability measures that are used to define neutral to the right priors and random hazard rate mixtures. Some of their distributional properties will be highlighted, with a special emphasis on their posterior characterization. The latter is relevant for the actual implementation of the proposed models to the analysis of partially exchangeable survival data.

A high performance biomarker detection method for exhaled breath mass spectrometry data

Ariadni Papana Dagiasis, Yuping Wu, Raed A. Dweik

Department of Mathematics, Cleveland State University, Ohio, USA
Department of Mathematics, Cleveland State University, Ohio, USA
Pulmonary and Critical Care Medicine / Respiratory Institute, and Pathobiology / Lerner Research Institute, Cleveland Clinic, Ohio, USA

Selected ion flow tube mass spectrometry, SIFT-MS, technology seems nowadays very promising to be utilized for the discovery and profiling of biomarkers such as volatile compounds, trace gases and proteins from biological and clinical samples. A high performance biomarker detection method for identifying biomarkers across experimental groups is proposed for the analysis of SIFT-MS mass spectrometry data. Analysis of mass-spectrometry data is often complex due to experimental design. Although several methods have been proposed for the identification of biomarkers from mass spectrometry data, there has been only a handful of methods for SIFT-MS data. Our detection method entails a three-step process that facilitates a comprehensive screening of the mass spectrometry data. First, raw mass spectrometry data are pre-processed to capture true biological signal. Second, the pre-processed data are screened via a random-forest-based screening tool that utilizes importance values. Finally, a visualization tool is complementing the findings from the previous step. Our biomarker detection method has two key components; the first is that this method can be used for the analysis of SIFT-MS data as well as other mass spectrometry data and the second is that it is applicable to fixed-time and time-dependent multi group experimental data. Our high performance detection method was applied to various SIFT-MS exhaled breath data from various samples and diseases such as asthma, liver disorders, pulmonary hypertension, sleep apnea and influenza infection. In this paper, we present two applications of our method; a control-asthma case study and an H1N1 Flumist time-course case study.

Local Block Bootstrap Inference for Trending Time Series

Arif Dowla
Stochastic Logic Ltd.
www.stochasticlogic.com

Efstathios Paparoditis
Dept. of Mathematics and Statistics
University of Cyprus
P.O.Box 20537
CY 1678 Nicosia, Cyprus

Dimitris N. Politis
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, USA
email: dpolitis@ucsd.edu

Abstract

Resampling for stationary sequences has been well studied in the last couple of decades. In the paper at hand, we focus on nonstationary time series data where the nonstationarity is due to a slowly-changing deterministic trend. We show that the *local block bootstrap* (LBB) methodology is appropriate for inference under this locally stationary setting without the need of detrending the data. We prove the asymptotic consistency of the local block bootstrap in the smooth trend model, and complement the theoretical results by a finite-sample simulation.

Greenwood's formula for a multivariate Kaplan-Meier estimator

Arusharka Sen

Department of Mathematics and Statistics, Concordia University, Montreal, Quebec H3G 1M8, Canada, *asen@mathstat.concordia.ca*

(Joint work with Winfried Stute, Justus-Liebig University, Giessen, Germany, and Yulan Jin, Concordia University, Montreal, Canada)

Abstract

We first present an estimator of the multivariate survivor function under multivariate random censoring. It is obtained using a characterization of the former as an eigenfunction for an integral operator based on the multivariate hazard measure. (It reduces to the famous Kaplan-Meier estimator in dimension one.) We then propose an estimator for the asymptotic variance — in fact, the variance-covariance matrix — of this estimator, i.e., a multivariate analogue of Greenwood's formula. This involves solving a so-called *matrix equation* of the form

$$\mathbf{UVU}^T = \mathbf{W},$$

where $\mathbf{V}_{n\times n}$ is the unknown (n = sample size). We also discuss how to estimate the variance of a linear functional of the survivor function estimator. The results will be illustrated with simulated as well as real (bivariate) data available in the literature.

Profile identification via weighted related metric scaling: An application to dependent Spanish children

Irene Albarrán[(1)] Pablo Alonso[(2)] Aurea Grané[(1)]

(1) Statistics Department. Universidad Carlos III de Madrid.
(2) Statistics Department. Universidad de Alcalá.

Abstract

Disability and dependence (lack of autonomy in performing common everyday actions) affect health status and quality of life, therefore they are significant public health issues. The main purpose of this study is to establish the existing relationship among different variables (continuous, categorical and binary) referred to children between 3 and 6 years old and their functional dependence in basic activities of daily living. We combine different types of information via weighted related metric scaling to obtain homogeneous profiles for dependent Spanish children. The redundant information between groups of variables is modelled with an interaction parameter that can be optimized according to several criteria. In this paper, the goal is to obtain maximum explained variability in an Euclidean configuration. Data comes from the Survey about Disabilities, Personal Autonomy and Dependence Situations, EDAD 2008, (Spanish National Institute of Statistics, 2008).

Componentwise classification of functional data

Aurore Delaigle

University of Melbourne, Australia

A.Delaigle@ms.unimelb.edu.au

The infinite dimension of functional data can challenge conventional methods for classification. A variety of techniques have been introduced to address this problem, particularly in the case of prediction, but the structural models that they involve can be too inaccurate, or too abstract, or too difficult to interpret. We introduce approaches to adaptively choose components, enabling classification to be reduced to finite-dimensional problems. Our techniques involve methods for estimating classifier error rate, and for choosing both the number of components, and their locations, to optimise these quantities. A major attraction of this approach is that it allows identification of parts of the function domain that convey important information for classification. It also permits us to determine regions that are relevant to one of these analyses but not the other.

This is joint work with Peter Hall and Neil Bathia.

References

Delaigle, A. and Hall, P. (2012). Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society, B*, **74**, 267–286.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, **40**, 322–352.

Delaigle, A., Hall, P. and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, to appear.

Axel B cher

Multiplier Bootstrap of Tail Copulas

In the problem of estimating the lower and upper tail copula we propose two boot- strap procedures for approximating the distribution of the corresponding empirical tail copulas. The first method uses a multiplier bootstrap of the empirical tail cop- ula process and requires estimation of the partial derivatives of the tail copula. The second method avoids this estimation problem and uses multipliers in the two- dimensional empirical distribution function and in the estimates of the marginal distributions. For both multiplier bootstrap procedures we prove consistency.
For these investigations we demonstrate that the common assumption of the ex- istence of continuous partial derivatives in the the literature on tail copula estima- tion is so restrictive, such that the tail copula corresponding to tail independence is the only tail copula with this property. We solve this problem and prove weak convergence of the empirical tail copula process under nonrestrictive smoothness assumptions which are satisfied for many commonly used models.

ASYMPTOTIC LAWS FOR CHANGE POINT ESTIMATION IN INVERSE REGRESSION

Axel Munk

Institute for Mathematical Stochastics, Georgia Augusta Universität Göttingen

Abstract: We derive rates of convergence and asymptotic normality for the least squares estimator for a large class of parametric inverse regression models $Y = (\Phi f)(X) + \varepsilon$. Our theory provides a unified asymptotic tretament for estimation of f with discontinuities of certain order, including piecewise polynomials and piecewise kink functions. Our results cover several classical and new examples, including splines with free knots or the estimation of piecewise linear functions with indirect observations under a nonlinear Hammerstein integral operator. Furthermore, we show that hard thresholding leads to a consistent model selection, using techniques from empirical process theory. The asymptotic normality is used to provide confidence bands for f. Simulation studies and a data example from rheology illustrate the results.

This is joint work with Sophie Frick and Thorsten Hohage.

Key words and phrases: Statistical inverse problems, jump detection, asymptotic normality, change point analysis, penalized least squares estimator, sparsity, entropy bounds, confidence bands, Hammerstein integral equations, reproducing kernel Hilbert spaces.

Institut für Mathematische Stochastik, Universität Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

E-mail: sbruns@math.uni-goettingen.de

Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestrasse 16-18, 37083 Göttingen, Germany

E-mail: hohage@math.uni-goettingen.de

Institut für Mathematische Stochastik, Universität Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

E-mail: munk@math.uni-goettingen.de

Title: Semiparametric Models with Bundled Parameters

Authors: Ying Ding and Bin Nan

Affiliation: University of Michigan

Abstract: In many semiparametric models that are parameterized by two types of parameters – a Euclidean parameter of interest and an infinite dimensional nuisance parameter, the two parameters are bundled together, i.e., the nuisance parameter is an unknown function that contains the parameter of interest as part of its argument. For example, in a linear regression model with censored survival data, the unspecified error distribution function involves the regression coefficients. Motivated by developing an efficient estimating method for the regression parameters, we consider the sieve maximum likelihood estimation and propose a general M-theorem for such bundled parameters. The numerical implementation of the proposed estimating method can be achieved through the conventional gradient-based search algorithms such as the Newton-Raphson algorithm. We show that the proposed estimator for the linear regression model with censored survival data is consistent, asymptotically normal and achieves the semiparametric efficiency bound. Finite sample performance is evaluated by simulations.

We give a general formulation of nonlinear sufficient dimension reduction, and explore its ramifications and scope. This formulation subsumes recent work employing reproducing kernel Hilbert spaces, and reveals many parallels between linear and nonlinear sufficient dimension reduction. Using these parallels we analyze the population-level properties of existing methods and develop new ones. We begin at the completely general level of σ-fields, and proceed to that of measurable and generating classes of functions. This leads to the notions of sufficient, complete and sufficient, and central dimension reduction classes. We show that, when it exists, the complete and sufficient class coincides with the central class, and can be unbiasedly and exhaustively estimated by a generalized slice inverse regression estimator (GSIR). When completeness does not hold, this estimator captures only part of the central class (i.e. remains unbiased but is no longer exhaustive). However, we show that a generalized sliced average variance estimator (GSAVE) can capture a larger portion of the class. Both estimators require no numerical optimization, because they can be computed by spectral decomposition of linear operators. Finally, we compare our estimators with existing methods by simulation and on actual data sets.

Bodhisattva Sen

Title: Estimation of a Two-component Mixture Model with Applications to Multiple Testing

Abstract:

We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the other unknown distribution nonparametrically, given i.i.d. data from the mixture model. We use ideas from shape restricted function estimation and develop tuning parameter free estimators that are easily implementable and have good finite sample performance. We establish the consistency of our procedures. Distribution-free finite sample lower confidence bounds are developed for the mixing proportion. The identifiability of the model, and the estimation of the density of the unknown mixing distribution are also addressed. We discuss the connection with the problem of multiple testing and compare our procedure with some of the existing methods in that area through simulation studies. We also analyse two data sets, one arising from an application in astronomy and the other from a microarray experiment.

Title: A new view of Fisher's information as a tool for projection pursuit, independent component analysis, and more.

Name: Bruce Lindsay

Affiliation: Pennsylvania State University

Abstract: This work is based on collaboration with Guodong Hui and Weixin Yao.

Let our observation be vector valued X of dimension d. Our basic object of investigation is the Fisher's information for theta in the location family f(x - theta) where f is an known continuous density function and theta is an location vector. This matrix has the surprising feature that the information is minimized, in the matrix sense, when the density f is multivariate Gaussian.

This might lead one to conjecture that this matrix could be useful as a diagnostic for multivariate normality when the density function f is unknown. We show that this is true in a very rich way. We consider the linear transformations of the data generated by the eigenvectors of the information matrix, which we call the "new variables" and show that the new variables with small eigenvalues are closest to being "white noise", by which we mean variables that marginally normal and independent of orthogonal linear combinations. In contrast, the most informative new variables must be either very non-normal or highly dependent on orthogonal variables.

In our second stage of analysis, we turn to the most informative variables, and use matrix analysis of variance tools to determine whether their high information arises from dependence or from non-normality. The set of mathematical and practical tools developed is relevant to projection pursuit, independent component analysis, and multivariate goodness-of-fit.

Semi and nonparametric Bayesian inference for long range dependence stationary Gaussian processes.

Brunero Liseo, Sapienza MeMoTEF, Roma Italy
(joint work with Nicolas Chopin and Judith Rousseau, CREST, France)

We discuss some Bayesian semi-parametric and nonparametric procedures for the analysis of stationary long range dependent Gaussian time series.
In the semi-parametric approach we use frequency domain methods – based on Whittle's approximation - to partition the infinite dimensional parameter space into regions where genuine prior information on the form of the spectral density is available, and others where vague prior beliefs are adopted; the solution to the partition problem, which is equivalent to bandwidth choice from a frequentist point of view, is obtained via Bayes factors.
In the nonparametric approach we avoid Whittle's approximation and prove posterior consistency for both the parameter of interest d and the entire density g, under appropriate conditions on the prior distribution. We also establish the rate of convergence for a general class of priors, and apply our results to the family of fractionally exponential priors.
We discuss the computational issues of the two approaches

Nonparametric Regression with Doubly Truncated Data

Carla Moreira * Jacobo de Uña - Álvarez § Luís Meira - Machado ††

Abstract

In this paper nonparametric regression with a doubly truncated response is introduced. Local constant and local linear kernel-type estimators are proposed. Asymptotic expressions for the bias and the variance of the estimators are obtained, showing the deterioration provoked by the random truncation. To solve the crucial problem of bandwidth choice, two different bandwidth selectors based on plug-in and cross-validation ideas are introduced. The performance of both the estimators and the bandwidth selectors is investigated through simulations. A real data illustration is included. The main conclusion is that the introduced regression methods perform satisfactorily in the complicated scenario of random double truncation.

Key Words: Local polynomial regression, Kernel smoothing, bandwidth selection, random truncation, biased data, mean squared error.

*Faculty of Economics. Department of Statistics and O.R., University of Vigo, Lagoas - Marcosende, 36 310 Vigo, Spain. University of Minho, Department of Mathematics for Science and Technology, Campus de Azurém, 4800-058 Guimarães, Portugal, E-mail address: carlamgmm@gmail.com. Research supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, by the Grant 10PXIB300068PR of the Xunta de Galicia and by Grant SFRH/BPD/68328/2010 of Portuguese Fundação Ciência e Tecnologia.

§Faculty of Economics. Department of Statistics and O.R., University of Vigo, Lagoas - Marcosende, 36 310 Vigo, Spain. E-mail address: jacobo@uvigo.es. Research supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, by the Grant 10PXIB300068PR of the Xunta de Galicia.

††University of Minho, Department of Mathematics for Science and Technology, Campus de Azurém, 4800-058 Guimarães, Portugal.

THE MULTIVARIATE LINEAR PROCESS BOOTSTRAP FOR STATIONARY TIME SERIES OF POSSIBLY INCREASING DIMENSION

CARSTEN JENTSCH AND DIMITRIS N. POLITIS

Abstract. This paper reconsiders the linear process bootstrap (LPB) proposed by McMurry and Politis (2010, J. Time Ser. Anal.) for univariate time series. We extend the LPB in several directions. First of all, we make the procedure applicable to multivariate time series. Under rather general assumptions that go beyond the physical dependence condition used in McMurry and Politis (2010), we prove asymptotic validity for the sample mean. Additionally, we show that the multivariate linear process bootstrap (MLPB) works for spectral density estimation which is also a novel result in the univariate case. Due to current interest in high-dimensional problems, we show consistency of tapered covariance matrix estimators when the time series dimension is allowed to increase with the sample size. The validity of the MLPB in this case is not obvious, but we provide rates of the time series dimension that allow for an asymptotic validity result of the MLPB for the sample mean. Finally, we conclude with a small simulation study that demonstrates the superiority of the MLPB in some important cases.

Investigating functional datasets with the BAGIDIS semimetric

Catherine Timmermans

Institute of Statistics, Biostatistics and Actuarial sciences, Université catholique de Louvain, voie du Roman Pays 20, BE-1348 Louvain-la-Neuve, Belgium.

This presentation highlights a new method for investigating functional datasets. The method is centered on the definition of a new functional, data-driven and highly adaptive semimetric for measuring dissimilarities between curves. It is based upon the expansion of each curve of a dataset into a *different* wavelet basis, one that is particularly suited for its description. The expansions remain however comparable as they rely on a common notion of hierarchy in describing the curve. Measuring dissimilarities in such a way implies comparing not only the projections of the curves onto the bases but also the bases themselves. Therefore, the name of the method stands for *BAses Giving DIStances.*

Due to its above-mentioned properties, the BAGIDIS semimetric reveals really powerful when dealing with curves with sharp local features that might be affected simultaneously by horizontal shifts and vertical amplification. Furthermore, as we overcome the limitation of expanding all the curves of a dataset in the same basis, it provides for a new paradigm for curves comparison, which opens attractive prospects.

Nonparametric tests for serial independence and Granger non-causality: An overview

Cees G.H. Diks

CeNDEF, University of Amsterdam

This paper reviews recent work by my co-authors and me on nonparametric tests for serial independence for univariate time series and nonparametric tests for Granger non-causality among the elements of multivariate time series.

Two nonparametric tests for serial independence will covered, based on quadratic forms and on marginal redundancies, respectively. When describing the test based on quadratic forms, particular attention is being payed to describing the connections between quadratic forms and more traditional divergence measures such as those based on empirical distribution functions and L^2 norms. The test based on marginal redundancies is more closely related to tests based on Kullback-Leibler divergence and other information theoretical divergence measures. For both tests it is described how the bandwidth selection problem can be addressed by using a multiple bandwidth approach. The performance of the tests in terms of their size and power properties is investigated numerically for simulated time series processes. The test based on quadratic forms is illustrated empirically with an application to financial time series.

Next nonparametric tests for Granger non-causality are discussed. After describing our nonparametric test for Granger non-causality and its relation to some earlier tests, the main focus will be on applications of linear as well as nonlinear Granger causality tests to multivariate empirical time series data from the oil futures and spot market, the exchange rate market and the grains market. We test for causality while correcting for the effects of the other variables. To check if any of the observed causality is strictly nonlinear in nature, we also examine the causal relationships among the residuals of an estimated optimal multivariate linear model, which for the non-stationary data at hand either takes the form of a VAR model for first differences or a VECM model in the presence of cointegration. For all datasets considered, the results suggest that the structure of the dynamic relations playing a role within the market is changing over time. For instance, in the grains market a clear decrease of the number of significant linear couplings can be observed, while the number of significant nonlinear couplings increases over time. We interpret this as statistical evidence that the market is becoming more efficient in the sense that linear relations are being exploited better by market participants.

Projection-based nonparametric goodness-of-fit testing in functional regression

Valentin Patilea[1], César Sánchez-Sellero[2] and Matthieu Saumard[3]

[1]CREST (Ensai) & IRMAR, France, patilea@ensai.fr
[2]Universidad de Santiago de Compostela, Spain, cesar.sanchez@usc.es
[3]INSA-IRMAR, France, Matthieu.Saumard@insa-rennes.fr

Abstract

We study the problem of nonparametric testing for the effect of a random functional covariate on a real-valued error term. The covariate takes values in $L^2[0,1]$, the Hilbert space of the square-integrable real-valued functions on the unit interval. The error term could be directly observed as a response or *estimated* from a functional parametric model, like for instance the functional linear regression. Our test is based on the remark that checking the no-effect of the functional covariate is equivalent to checking the nullity of the conditional expectation of the error term given a sufficiently rich set of projections of the covariate. Such projections could be on elements of norm 1 from finite-dimension subspaces of $L^2[0,1]$. Next, the idea is to search a finite-dimension element of norm 1 that is, in some sense, the least favorable for the null hypothesis. Finally, it remains to perform a nonparametric check of the nullity of the conditional expectation of the error term given the scalar product between the covariate and the selected least favorable direction. For such finite-dimension search and nonparametric check we use a kernel-based approach. As a result, our test statistic is a quadratic form based on univariate kernel smoothing and the asymptotic critical values are given by the standard normal law. The test is able to detect nonparametric alternatives, including the polynomial ones. The error term could present heteroscedasticity of unknown form. We do no require the law of the covariate X to be known. The test performs well in simulations and real data applications.

Nonparametric regression for circular data

Charles C. Taylor

University of Leeds, Leeds LS2 9JT, UK

Starting with a review of existing parametric models, we put these into a common framework and discuss problems with estimation. Various non-paramtric models, wihch make use of circular kernels, are described, as well as their asymptotic behaviour and approaches to bandwidth selection.

Examples are used to illustrate the methods.

Title: Variable Selection in Joint Modelling of Mean and Variance for Multilevel Data

Abstract: We propose to extend the use of penalized likelihood based variable selection methods to hierarchical generalized linear models (HGLMs) for jointly modelling both the mean and variance structures. We are interested in applying these new methods on multilevel structured data, hence we assume a two-level hierarchical structure, with subjects nested within groups. We consider a generalized linear mixed model (GLMM) for the mean, with a structured dispersion in the form of a generalized linear model (GLM). In the first instance, we model the variance of the random effects which are present in the mean model, or in other words the variation between groups (between-level variation). In the second scenario, we model the dispersion parameter associated with the conditional variance of the response, which could also be thought of as the variation between subjects (within-level variation). To do variable selection, we use the smoothly clipped absolute deviation (SCAD) penalty, a penalized likelihood variable selection method, which shrinks the coefficients of redundant variables to 0 and at the same time estimates the coefficients of the remaining important covariates. Our methods are likelihood based and so in order to estimate the mixed effects in our models, we apply iterative procedures such as the Newton-Raphson method, in the form of the LQA algorithm proposed by Fan and Li (2001). We carry out simulation studies for both the joint models for the mean and variance of the random effects, as well as the joint models for the mean and dispersion of the response, to assess the performance of our new procedures against a similar process which excludes variable selection. The results show that our method increases both the accuracy and efficiency of the resulting penalized MLEs and has 100% rate in identifying the zero and non-zero components over 100 simulations. The efficiency of our estimates is also verified by the real data analysis of the Health Survey for England (HSE) 2004 data and the Integrated Circuit data.

Christina Steinkohl

Max-stable processes have proved to be very useful for the statistical modelling of spatial extremes. We use the idea of constructing max-stable random fields as limits of normalized and rescaled pointwise maxima of Gaussian random fields to construct a max-stable space-time random field.

Since multivariate distribution functions are usually intractable for such models pairwise likelihood methods are often used for statistical inference. In this talk we present a semiparametric procedure to estimate the parameters in our model. We use a two-step estimation procedure, where we first calculate nonparametric estimates for the extremogram in space and time, introduced in Davis and Mikosch [2009], and then use weighted regression to estimate the parameters. Since the asymptotic variance of the estimates is intractable, bootstrap based confidence intervals are constructed.

(This is joint work with Richard A.Davis and Claudia Kl ppelberg)

Specification testing in nonparametric instrumental quantile regression.

Christoph Breunig *

Universität Mannheim

May 14, 2012

In nonparametric instrumental quantile regression, the function of interest is the solution to a nonlinear operator equation. Depending on the joint distribution of the instrument and the regressors, a solution might not exist. In this paper we consider a test whether a solution to the operator equation exists. Our test statistic is asymptotically normally distributed under correct specification and consistent against any alternative model. Under a sequence of local alternatives the asymptotic distribution of our test is derived. Moreover, uniform consistency is established over a class of alternatives whose distance to the null hypothesis shrinks appropriately as the sample size increases.

*Lehrstuhl für Statistik, Abteilung Volkswirtschaftslehre, L7, 3-5, 68131 Mannheim, Germany, e-mail: cbreunig@staff.mail.uni-mannheim.de

Title: "Robust estimation of mean and dispersion functions in extended generalized additive models"

Presenter: Christophe Croux, University of Leuven, Belgium,.

Co-authors: Irene Gijbels and Ilaria Prosdocimi, University of Leuven, Belgium.

Abstract:

Generalized Linear Models are a widely used method to obtain parametric estimates for the mean function. They have been further extended to allow the relationship between the mean function and the covariates to be more flexible via Generalized Additive Models. However the fixed variance structure can in many cases be too restrictive. The Extended Quasi-Likelihood (EQL) framework allows for estimation of both the mean and the dispersion/variance as functions of covariates. As for other maximum likelihood methods though, EQL estimates are not resistant to outliers: we need methods to obtain robust estimates for both the mean and the dispersion function. In this paper we obtain functional estimates for the mean and the dispersion that are both robust and smooth. The performance of the proposed method is illustrated via a simulation study and some real data examples.

High-dimensional multivariate regression

Christophe Giraud

Ecole Polytechnique (France)

We will review some recent results in multivariate regression in high-dimensional setting under various structural assumptions : sparsity, low rank and both together. We will focus on computationally feasible estimation procedures for which we will present some non-asymptotic results. We will highlight some results for the Gaussian setting with unknown variance. Part of these results are linked to recent works in random matrix theory.

References

[1] Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. Ann. Stat. 39, 2, 12821309.

[2] Bunea, F., She, Y., and Wegkamp, M. H. (2011). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. arXiv:1110.3556v1

[3] Giraud, C. (2011a). Low rank multivariate regression. Electron. J. Stat. 5, 775799.

[4] Giraud, C. (2011b). A pseudo-rip for multivariate regression. Arxiv:1106.5599v1.

[5] Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. J. Multivariate Anal. 5, 248264.

[6] Klopp, O. (2011). High dimensional matrix estimation with unknown variance of the noise. Arxiv:1112.3055v1.

[7] Koltchinski, V., Lounici, K., and Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. Annals of Statistics 39, 5, 23022329.

[8] Lounici, K., Pontil, M., Tsybakov, A., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. Annals of Statistics 39, 4, 21642204.

[9] Marchenko V.A. and Pastur L.A. (1967). Distribution of eigenvalues for some sets of random matrices. Mat. Sb. (N.S.), 72(114), 507536.

[10] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 39, 2, 10691097.

[11] Rigollet, P. and Tsybakov, A. (2011). Sparse estimation by exponential weighting. arXiv:1108.5116v1

[12] Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. Proceedings of the International Congress of Mathematicians, Hyderabad, India.

Estimating Nonlinear Additive Models with Nonstationarities and Correlated Errors

Michael Vogt
University of Cambridge

Christopher Walsh
University of Mannheim

In this paper, we study a nonparametric additive regression model suitable for a wide range of time series applications. Our model includes a periodic component, a deterministic time trend, various component functions of stochastic explanatory variables, and an AR(p) error process that accounts for serial correlation in the regression error. We propose an estimation procedure for the nonparametric component functions and the parameters of the error process based on smooth backfitting and quasi-maximum likelihood methods. Our theory establishes convergence rates as well as asymptotic normality of our estimators. Moreover, we are able to derive an oracle type result for the estimators of the AR parameters: Under fairly mild conditions, the limiting distribution of our parameter estimators is the same as when the nonparametric component functions are known. Finally, we illustrate our estimation procedure by applying it to a sample of climate and ozone data collected on the Antarctic Peninsula.

Key words: semiparametric, nonstationary, smooth backfitting, correlated errors.
AMS 2010 subject classifications: 62G08, 62G20, 62F12, 62P12.

A generalization of the Youden index for multiple-class classification problems useful for cut-off point selection in ROC surface analysis

C T Nakas, Laboratory of Biometry, University of Thessaly
cnakas@uth.gr
T A Alonzo, University of Southern California
J C Dalrymple-Alford & T J Anderson, New Zealand Brain Research Institute

The Youden index (J) is widely used in ROC curve analysis for selection of the optimal cut-off point which can be used in practice for classification purposes. Recently, a generalization of the Youden index (J_3) for three-class classification problems was proposed. J_3 can be used for the selection of the optimal cut-off points in ROC surface analysis. In this work we propose a generalized index (J_k) for k-class classification problems, with k>3. We examine theoretical and geometric properties of the index J_k. Methods are applied in the assessment of cognitive tests when screening cognition in Parkinson disease.

We develop monitoring schemes for detecting structural changes in nonlinear autoregressive models by using a semi-parametric approach. We first approximate the regression function by a parametric class such as single layer feedforward neural networks, then use statistics based on cumulative sums of estimating functions for the unknown parameter in this parametric class. Finally, we obtain the limit distribution for the corresponding sequential size α-tests in the misspecified case, where the observed time series does not exactly follow the chosen parametric approximation. As a result the proposed monitoring schemes reject (asymptotically) the null hypothesis only with a given probability α but will detect a large class of alternatives with probability one both in the correct as well as misspecified case.

Claudia Kl ppelberg

Technische Universit t M nchen

Statistics for turbulence data with high Reynolds numbers

Many real world turbulent flows, e.g. boundary layer atmospheric turbulence, are characterized by a high Reynolds number (fully developed turbulence). We will outline stylized facts exhibited by the velocity of a fully developed turbulent flow, when observed at high frequency in the longitudinal direction at a fixed location. We base the investigation of such mean flow turbulence data on a stochastic intermittency model, which integrates a memory function by a stochastic process with uncorrelated and weakly stationary increments, and allows for modeling of intermittency effects. We aim at statistical inference for very high frequency data as for instance the Brookhaven data. In this talk we discuss the estimation of the memory function via the spectrum of the process in a non-parametric way. The estimated velocity spectrum confirms Kolmogorov's 5/3-law for a turbulence spectrum for the inertial range. After filtering out the driving process we are able to investigate its distributional properties and dependence structure.

Nonlinear estimation in a nonstationary environment

Dag Tjøstheim

Department of Mathematics, University of Bergen,
5008 Bergen, NORWAY

The talk consists of two parts. In the first part a review is given of the Markov splitting technique and its application to nonparametric estimation in (possibly) nonstationary Markov chains. In particular, connections to nonlinear cointegrating regression is pointed out. In the second part new results are reviewed for parametric estimation of nonstationary processes. Also for this part the splitting technique plays an important part. Unlike the stationary situation, there is a fundamental difference between the regressive and the autoregressive case. The talk is based on joint work with several coauthors.

A Non-standard Empirical Likelihood for Time Series

Daniel J. Nordman
Department of Statistics, Iowa State University

Standard blockwise empirical likelihood (BEL) for stationary, weakly dependent time series requires specifying a fixed block length as a tuning parameter for setting confidence regions. This aspect can be difficult and impacts coverage accuracy. As an alternative, this talk discusses a new version of BEL based on a simple, though non-standard, data-blocking rule which uses a data block of every possible length. Consequently, the method involves no block selection and is also anticipated to exhibit better coverage performance. Its non-standard blocking scheme, however, induces non-standard asymptotics compared to standard BEL. We shall present the large-sample distribution of log-ratio statistics from the new BEL method for calibrating confidence regions for mean or smooth function parameters of time series. This limit law is not the usual chi-square one, but is distribution-free and can be reproduced through straightforward simulations. Numerical studies indicate that the proposed method generally exhibits better coverage accuracy than standard BEL.

ROTATION SAMPLING FOR FUNCTIONAL DATA

DAVID DEGRAS

DePaul University, Chicago, IL, USA

Survey sampling methods provide highly cost-effective solutions for monitoring population mean levels over time. For this purpose, time-varying samples often yield better performances than fixed panels. Motivated by recent applications in sensor networks, we develop novel sampling designs in the context of functional data (that is, continuous signals). The proposed samples incorporate stratification and can be replaced at given times by rotation sampling. They are defined as Markov chains, which allows for an effective adaptation to transversal and longitudinal population variations based on recent observations. Considering the Horvitz-Thompson estimator of the mean temporal signal, we show that the variance of the Integrated Squared Error (ISE) can be dramatically reduced by increasing the frequency or intensity of sample replacements. Further, the average ISE can be decreased by suitably allocating the sample across strata at replacement times. An application to simulated electricity consumption data illustrates the good performances of our sampling designs relative to fixed panels.

Keywords: survey sampling, functional data, Markov chains, Horvitz-Thompson estimator, asymptotic theory.

Manuscript available at `http://arxiv.org/abs/1204.4494`

R-ESTIMATION IN INDEPENDENT COMPONENT ANALYSIS

DAVY PAINDAVEINE PAULIINA ILMONEN

UNIVERSITÉ LIBRE DE BRUXELLES

ABSTRACT. We consider semiparametric location-scatter models for which the p-variate observation is obtained as $X = \Lambda Z + \mu$, where μ is a p-vector, Λ is a full-rank $p \times p$ matrix, and the (unobserved) random p-vector Z has marginals that are centered and mutually independent but are otherwise unspecified. As in blind source separation and independent component analysis (ICA), the parameter of interest throughout the paper is Λ. On the basis of n i.i.d. copies of X, we develop estimation procedures for Λ under a symmetry assumption on Z. We exploit the uniform local and asymptotic normality (ULAN) of the model to define signed-rank estimators that are semiparametrically efficient under correctly specified densities. Yet our estimators remain root-n consistent under a very broad range of densities. We derive the asymptotic properties of the proposed estimators and investigate their finite-sample behavior through simulations.

REFERENCES

[1] A. Chen, and P. J. Bickel. Efficient independent component analysis. *Ann. Statist.*, 34:2825–2855, 2006.
[2] M. Hallin, and Werker, B. J. M. Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli*, 9:137–165, 2003.
[3] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.
[4] H. Oja, S. Sirkiä, and J. Eriksson. Scatter matrices and independent component analysis. *Austrian J. Statist.*, 35:175–189, 2006.
[5] P. Ilmonen, and D. Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *Ann. Statist.* 39:2448–2476, 2011.

A general method for robust estimation of time-varying parameters in diffusion models is proposed.
This allows for estimation of asset pricing models in unstable environments where the researcher has no strong priors regarding the nature of instability.
Under regularity conditions, the estimators are shown to be consistent and asymptotically normally distributed.
A key part of the asymptotic analysis is the approximation of the time-inhomogenous diffusion model by time-homogenous stationary versions.
This allows for the local approximation of the non-stationary sample path by a stationary version.
A simulation study shows the good finite-sample performance of the estimators.
An empirical study employs the developed framework and tools to investigate parameter instability in multifactor term structure models when fitted to the US yield curve.

Title : Mixed Effects Trees and Random Forests for Clustered Data

Authors: Denis Larocque (1), Ahlem Hajjem (2) and François Bellavance (1)

(1) Department of Management Sciences, HEC Montreal, Montreal, Canada
(2) Department of Marketing, Université du Québec à Montréal (UQAM), Montreal, Canada.

This paper presents extensions of tree-based and random forest methods for the case of clustered data. Previous works extending tree methods to accommodate correlated data are mainly based on the multivariate repeated-measures approach. We propose mixed effects regression tree and forest methods where the correlated observations are viewed as nested within clusters rather than as vectors of multivariate repeated responses. The proposed method can handle unbalanced clusters, allows observations within clusters to be splitted, and can incorporate random effects and observation-level covariates. They are implemented using standard algorithms within the framework of the EM algorithm. Simulation results show that the proposed methods provides substantial improvements over standard trees and forests when the random effects are non negligible. The use of the method is illustrated to predict the first-week box office revenues of movies.

A nonparametric test for independence of bivariate random variables

Dimitrios Bagkavos[1] and Prakash Patil[2]

[1] Accenture, Greece, email: dimitrios.bagkavos@gmail.com,

[2] The University of Birmingham, Department of Mathematics, Birmingham, UK

ABSTRACT. A new nonparametric test is proposed to test the independence of bivariate random variables X and Y. The test statistics for the proposed test is based on the fact that under independence every quantile of Y given X = x is constant. This is in contrast to the most commonly used basis that the joint probability density or distribution function of X and Y equal to the product of their marginal probability density or distribution functions to develop a test statistics for the independence tests. The theoretical properties of the proposed test are investigated together with the concepts of application in small to moderate sample sizes. The test power is also investigated by use of distributional data and comparison with other procedures is provided via simulation studies.

DECONVOLVING DISTRIBUTION ESTIMATORS

Dimitrios Ioannides

University of Macedonia, Thessaloniki Greece

George Stamatelos

Aristotelian University, Thessaloniki Greece

Abstract: A nonparametric estimator for a smooth distribution function based on contaminated observations was first considered by Fan. This paper proposes a modified estimator of the Fan estimator and a method is developed to establish its asymptotic properties for weakly stochastic processes corrupted by some noise process. The asymptotic normality is obtained under very general assumptions on the error characteristic function , which generalizes previous conditions on this topic.

Keywords: deconvolution, nonparametric estimation, distribution function, noise distribution

Confidence and prediction intervals in nonparametric regression without an additive model

Dimitris N. Politis
Univ. of California—San Diego

Abstract:

The Model-free Prediction Principle was introduced by Politis (ISI Bulletin, 2007). Its application to nonparametric regression (Y vs. X) involves a particular transformation of the non-i.i.d. response vector Y into a vector with i.i.d. components. An additive regression model is not required; all that is required is that the conditional distribution of Y given X is continuous in Y, and a smooth function of X. Working in the transformed domain allows the use of the i.i.d. bootstrap in order to construct confidence intervals for the unknown regression function, as well as prediction intervals for a yet unobserved response Y.

Expert system forecasting: nonparametric bootstrap approach

Dmitry Sumkin

Moscow Institute of Physics and Technology

The bootstrap technique, which applied in expert systems, is discussed in the paper. It is concerned hierarchical tree structure (graph) of criteria in decision support systems, which is widely used in the field of forecasting. Expert judgments of criteria values (assigned to graph nodes) are assumed to be as the original sample. The criteria values of dependent and independent criteria (connected by ribs of graph) are assumed to be as dependent and independent variables correspondingly. The main idea is to implement nonparametric regression model to criteria tree of decision support system and under statistics of judgments via bootstrap technique obtain weights of graph ribs. The purpose is to describe the rules of experts (in form of criteria) used to make a judgment, formalize them in the form of nonparametric regression model and compare them. It could be done via aggregation of judgments by calculating weights of regression model with the help of bootstrap, because size of the sample is not large enough to determine the kind of distribution. The first advantage of this forecasting technique is that it is combination of methods. The technique could be applicable when experts give direct quantitative judgments to the criteria values, and should be useful when expert judgments have validity but data are scarce and where key factors do not change in the historical data. The initial model is tested for robustness using the random number generator. Moreover, the results of the bootstrap system accuracy and robustness are compared with neural network approach applied to the same criteria hierarchy and bias data trained on the same set of random data. The approach seems to be promising because studies showed that bootstrapping improved the quality of production decisions in companies, more accurate than unaided judgments and appropriate in complex situations.

Testing for the error distribution in GARCH models

M. Dolores Jiménez-Gamero

Dpto. Estadística e Investigación Operativa,
Universidad de Sevilla,
41012 Sevilla, Spain

Abstract

A class of goodness-of-fit tests for the innovation distribution in generalized autoregressive conditional heteroscedastic models based on the empirical characteristic function of the residuals is studied. Single and composite hypothesis are considered. The tests are consistent against any fixed alternative for suitable choices of the weight function involved in the definition of the test statistic. The bootstrap can be employed to estimate consistently the null distribution of the test statistic. The goodness of the bootstrap approximation and the power of some tests in this class for finite sample sizes are investigated by simulation.

Domenico Marinucci

Title: Stein-Malliavin Approximations for Wavelet Coefficients on Spherical Poisson Fields

Authors: Claudio Durastanti, Domenico Marinucci and Giovanni Peccati

Abstract: We prove some upper bounds for the Wassestein distance to a multivariate Gaussian distribution for the joint law of wavelets/needlets coefficients on spherical Poisson fields. More precisely, we develop some results from Peccati and Zheng (2011), based on Malliavin calculus and Stein's methods, to establish the rate of convergence to Gaussianity for triangular array of needlet coefficients with growing dimensions. The results are motivated by astrophysical and cosmological applications, in particular related to the search for point sources in Cosmic Rays data.

MODELING AND FORECASTING ELECTRICITY SPOT PRICES: A FUNCTIONAL DATA PERSPECTIVE

By Dominik Liebl
Universities of Cologne and Bonn

Classical time series models have serious difficulties in modeling and forecasting the enormous fluctuations of electricity spot prices. Markov regime switch models became one of the most often used models in the electricity literature. These models try to capture the fluctuations of electricity spot prices by using different regimes, each with own mean and covariance structure. Usually one regime is dedicated to moderate prices and another is dedicated to high prices. However, these models show poor performances and there is no theoretical justification for this kind of classification. The merit-order model however, the most important microeconomic pricing model for electricity spot prices, suggests a continuum of mean-levels with a functional dependence on electricity demand.
We propose a new statistical perspective on modeling and forecasting electricity spot prices that accounts for the merit-order model. In a first step, the functional relation between electricity spot prices and electricity demand is modeled by daily price-demand functions. In a second step, we parametrize the series of daily price-demand functions using a functional factor model. The power of this new perspective is demonstrated by a forecast study that compares the functional factor model with established classical time series models for electricity spot prices.

Copulas for dependence structure testing

Irène Gijbels, Dominik Sznajder*
KU Leuven, Belgium

In this talk we discuss tests for several classes of dependence structures, namely quadrant dependence, tail monotonicity and stochastic monotonicity. These kind of relations between random variables are of particular interest in financial, insurance and econometric studies and we illustrate it on several examples of data sets available in literature.

Often dependence between random variables in a random vector can be expressed as a feature of the underlying copula function. Such a copula function is a linking function between the joint distribution of a random vector and the marginal distributions.

We propose to build the test statistics as functional violation measures based on the empirical copula estimator. Furthermore, the statistical inference is based on the bootstrapped distribution of the test statistics. This requires resampling scheme under the null hypothesis and we propose a smooth constrained nonparametric copula estimation procedure as a remedy. It is based on the local polynomial smoothing of the initial constrained estimator and on transforming its partial derivatives by rearrangement technique.

The proposed methodology is generic and flexible and can be applied to other dependence concepts, which can be expressed as shape constraints on the copula function.

References

[1] I. Gijbels, D. Sznajder *Positive quadrant dependence testing and constrained copula estimation*, The Canadian Journal of Statistics, accepted, (2012).

[2] I. Gijbels, D. Sznajder *Testing tail monotonicity by constrained copula estimation*, submitted (2011).

[3] I. Gijbels, D. Sznajder *Constrained copula estimation in stochastic monotonicity testing*, manuscript (2011).

Nonparametric inference for directional–linear data

Eduardo García–Portugués, Rosa M. Crujeiras, and Wenceslao González–Manteiga

Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela.

A kernel density estimator for directional–linear data is proposed. The estimator is based on a directional–linear kernel product and expressions for bias, variance and mean square error are derived. Optimal smoothing parameters in terms of the asymptotic mean integrated square error criterion is also provided. Using the proposed estimator, an L^2 goodness–of–fit tests for directional–linear densities is also introduced. The estimator and test finite sample performances are explored throughout a simulation study for the circular–linear and spherical–linear cases. Finally, the methods are illustrated with a real data example.

Sufficient reductions for elliptically contoured distributions

Efstathia Bura[1] and Liliana Forzani[2]

[1]George Washington University
[2] Universidad Nacional del Litoral/Instituto Matematica Aplicada Litoral - CONICET

May 14, 2012

Abstract

There are two general approaches for determining sufficient reductions via inverse regression. The first is the moment-based approach (SIR, SAVE, DR) in which moments of the conditional distribution of $\mathbf{X}|Y$ are used to estimate a sufficient reduction. Despite the success of the moment-based methods to reduce the dimension, they depend upon distributional assumptions and are constructed to estimate linear reductions. Cook and Forzani (2008) introduced likelihood-based Sufficient Dimension Reduction (SDR): LAD (likelihood acquired directions) inherits properties and methods from general likelihood theory, provides exhaustive estimation of the central subspace under mild conditions and constitutes an asymptotically optimal dimension reduction method in terms of efficiency.Yet, this is so only when the conditional distribution of $\mathbf{X}|Y$ is normal. We extend both moment-based and model-based inverse regression dimension reduction to the general case where the conditional distribution of $\mathbf{X}|Y$ is elliptically contoured. We also extend Linear SDR to General SDR by exploring and identifying nonlinear sufficient reductions.

BOOTSTRAPPING LOCALLY STATIONARY PROCESSES

JENS-PETER KREISS AND EFSTATHIOS PAPARODITIS

Abstract. We propose a nonparametric method to bootstrap locally stationary processes which combines a time domain wild bootstrap approach with a nonparametric frequency domain approach. The method generates pseudo-time series which mimic (asymptotically) correct, the local second and to the necessary extent the fourth order moment structure of the underlying process. Thus it can be applied to approximate the distribution of statistics that are based on observations of the locally stationary process. We prove a bootstrap central limit theorem for a general class of statistics that can be expressed as functionals of the preperiodogam, the later being a useful tool in inferring properties of locally stationary processes. A real data example illustrate the capability of the bootstrap method proposed.

Technische Universität Braunschweig, Insitut für Mathematische Stochastik, Pockelsstrasse 14, D-38106 Braunschweig, Germany

University of Cyprus, Department of Mathematics and Statistics, P.O.Box 20537, CY-1678 Nicosia, Cyprus

Beyond Simplified and Towards General Pair Copula Constructions

ELIF F. ACAR

McGill University, Department of Mathematics and Statistics

Pair-copula constructions (PCCs) offer great flexibility in modeling multivariate dependence. For inference purposes, however, conditional pair copulas are often assumed to depend on the conditioning variables only indirectly. In this talk, it will be shown through examples that this assumption can be misleading. To assess its validity in trivariate PCCs, a visual tool based on a local likelihood estimator of the conditional copula parameter which does not rely on the simplifying assumption will be presented. The proposed technique will be demonstrated using simulated and real data. The talk will further address how to construct a formal test of the simplifying assumption in trivariate PCCs and outline inference for general PCCs in higher dimensions.

Resampling methods for weakly dependent sequences

Elżbieta Gajecka-Mirek
State Higher Vocational School in Nowy Sacz, Poland

First conference of the International Society for NonParametric Statistics (ISNPS), Chalkidiki, Northern Greece, June 15-19 2012

Introduction

Many authors have used the mixing properties as a type of dependence in time series. Unfortunately many classes of time series do not satisfy any mixing condition.
In 1999 *Bickel* and *Bülmann* and simultaneously *Doukhan* and *Louhichi* introduced a new idea of dependence - weak dependence.

Attention

Mixing conditions are dependence conditions in terms of the $\sigma-$algebras generated by a random sequences. In that case we need to consider conditions, which are often unverified or very difficult to verify in practice.

Attention

Relaxation of mixing conditions and assuming the notion of weak dependence give us tools for the analysis of statistical procedures with very general data generating processes (e.g.: Bernoulli shifts or Markov processes driven by discrete innovations). The weak dependence is measured in terms of covariances of functions, and those are easier to estimate.

$\alpha-$mixing time series

Definition

Let $\{X_t : t \in \mathbb{Z}\}$ be time series. We define $\alpha-$mixing sequence as

$$\alpha_X(\tau) = \sup_{t\in\mathbb{Z}} \sup_{\substack{A\in\mathcal{F}_X(-\infty,t)\\ B\in\mathcal{F}_X(t+\tau,\infty)}} |P(A\cap B) - P(A)P(B)|,$$

where $\tau \in \mathbb{N}$.
We call the time series $\{X_t\}$ $\alpha-$mixing if $\alpha_X(\tau) \to 0$ for $\tau \to \infty$.

Attention

Intuition: mixing is "forgetting" in time series, it means that distant observations are almost independent random variables. In consequence we can obtain limits results.

Weakly dependent sequences

Let $(E, \|\cdot\|)$ be a normed space and let $\mathbb{L} = \bigcup_{u\in\mathbb{N}^*} \mathbb{L}^\infty(E^u)$ denote the set of numeric, measurable and essentially bounded functions on space E^u. A function $h : E^u \longrightarrow \mathbb{R}$ belongs to the class $\mathcal{L} = \{h : E^u \to \mathbb{R},\ \| h \|_\infty \le 1, Lip(h) < \infty\}$, where $Lip(h) = \sup_{x\neq y} \frac{|h(x)-h(y)|}{\|x-y\|_1}$ and $\| x \|_1 = \sum_{i=1}^u \| x_i \|$.

Definition

A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables taking values in $E = \mathbb{R}^d$ is $(\theta, \mathcal{L}, \Psi)-$weakly dependent if there exist
$\Psi : \mathcal{L}\times\mathcal{L}\times\mathbb{N}^*\times\mathbb{N}^*\to\mathbb{R}$ and a sequence $\{\theta_r\}_{r\in\mathbb{N}}$ $(\theta_r \to 0)$ such that for any $(f, g) \in \mathcal{L}\times\mathcal{L}$, and $(u, v, r) \in \mathbb{N}^{*2}\times\mathbb{N}$

$$|(f(X_{i_1},...,X_{i_u}), g(X_{j_1},...,X_{j_v}))| \le \Psi(f, g, u, v)\theta_r$$

whenever $i_1 < i_2 < ... < i_u \le r + i_u \le j_1 < j_2 < ... < j_v$.

Example 1

Weakly dependent sequence: Bernoulli shift

Let $X_n = H(\xi_n, \xi_{n-1}, ...)$, with $H(X) = \sum_{k=0}^\infty 2^{-(k+1)}\xi_{n-k}$. Where ξ_{n-k} is the $k-th$ digit in the binary representation of the uniformly chosen number $X_n = 0.\xi_n\xi_{n-1}... \in [0, 1]$.

$\{Xn\}$ is not mixing, because:
X_n is deterministic function of X_0, so the event $A = (X_0 \le \frac{1}{2})$ belongs to the $\sigma-$ algebras: $\sigma(X_t, t \le 0)$ and $\sigma(X_t, t \ge n)$. From definition: $\alpha(n) \ge | P(A\cap A) - P(A)P(A) | = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$.
But the $\{X_n\}$ is weakly dependent, because:

Lemmat

(Doukhan, Louhichi, 2008) Bernoulli shifts are $\theta-$weakly dependent with $\theta(r) \le 2\delta_{[r/2]}$, where $\{\delta_r\}_{r\in\mathbb{N}}$ is defined by: $E | H(\xi_{t-j}, j \in \mathbb{Z}) - H(\xi_{t-j}1_{|\le r}, j \in \mathbb{Z}) |$.

Example 2

Weakly dependent sequence: Markov processes

Let $Z_t = f(Z_{t-1}, \xi_t)$, $(t \in \mathbb{Z})$ be an $\mathbb{R}^D-$valued Markov process. $\{\xi_t\}_{t\in\mathbb{Z}}$ is i.i.d. with $E(\xi_0)$, ξ_t is independent from $\{Z_s; s < t\}$ and $f : \mathbb{R}^D\times\mathbb{R}\to\mathbb{R}^D$. Z_0 is independent of the sequence $\{\xi_t\}_{t\in\mathbb{N}}$.

Suppose that for some $0 \le c_i \le 1$, $E | f(0, \xi_1) | < \infty$ and $E | f(u, \xi_1) - f(v, \xi_1) | \le \sum_{i=1}^d c_i | u_i - v_i |$, where $c = \sum_{i=1}^d c_i < 1$ for all $u, v \in \mathbb{R}^D$. This kind of Markov process has a stationary distribution μ with finite first moment. If $c = \sum_{i=1}^d c_i < 1$ holds Markov chain is weak dependent with $\theta_r = c^r E | Z_0 |$.

Resampling

In resampling we consider the sample as a "population" and we draw many samples from it or construct many rearrangements of the obtained sample values. For each sample or rearrangement, we compute a statistic. The set of statistics constitutes the sampling distribution of that statistic, and we can use that sampling distribution to draw inferences about the model underlying the data.

Why resampling?

- Statistic inference for dependent data based on asymptotic distributions often fails
- Distribution of the estimators converges to asymptotic distributions is often slow
- The asymptotic distributions in many cases are very complicated
- In practice we often have problems with collection the data witch are long enough

Resampling methods can help. All we need to know is if there exists non-degenerated asymptotic distribution of the statistic (we do not have to know the form of the asymptotic distribution) - **we need to have CLT**.

Attention

Consistency of resampling methods

Using resampling methods we need to answer the question:

- If the empirical distribution (obtained from resampling methods) is close enough to real distribution?

We need to know if our method is consistence.

Resampling methods for weakly dependent models

In the poster there are considered 2 of the resampling methods:

- The Block bootstrap - Bootstrapping GMM (generalized method of moments) estimators
- Subsampling (Politis, 1999)

1. The Block bootstrap - bootstrapping GMM estimators

The block-bootstrap procedure (adapted to the time series $(X_i)_{i\in\mathbb{N}}$):

- Let $b = b(n)$ and $l = l(n)$ denote the number and the length of the blocks.
- Assume $b \cdot l = n$ and consider l blocks $(X_{(j-1)l+1}, ..., X_{jl})$ for $1 \le j \le b$.
- Bootstrap's blocks $(\widetilde{X}_{(j-1)l+1}, ..., \widetilde{X}_{jl})$ are randomly drawn among those l blocks.

Bootstrapping GMM estimators

Let the arg-min problem $J_n(\hat{\theta}_n) = \min_{\theta\in\Theta} J_n(\theta)$, where

$$J_n(\theta) = (\frac{1}{n}\sum_{i=1}^n g(X_i, \theta))' \Omega (\frac{1}{n}\sum_{i=1}^n g(X_i, \theta)).$$

$\Theta \subset \mathbb{R}^d$, $g(\cdot, \cdot)$ is a given function: $E_{\theta_0} g(X_1, \theta_0) = 0$, where θ_0 is the true parameter point.
$\hat{\theta}_n$ is a solution of the arg-min problem (GMM estimation procedures involve an estimate $\hat{\theta}_n$).

Theorem

CLT holds under standard mixing assumptions:

$$T_n(\theta) = n^{\frac{1}{2}}\Sigma_n^{-1}(\hat{\theta}_n - \theta) \to^{\mathcal{D}}_{n\to\infty} \mathcal{N}_d(0, I_d).$$

Let $\{\widetilde{X}_i^*\}_{1\le i\le n}$ denote a block-bootstrap sample. Let $g^*(x, \theta) = g(x, \theta) - E^* g(x, \hat{\theta}_n)$. The GMM estimate $\hat{\theta}_n^*$ solves the arg-min problem

$$J_n^*(\theta) = (\frac{1}{n}\sum_{i=1}^n g^*(X_i^*, \theta))' \Omega (\frac{1}{n}\sum_{i=1}^n g^*(X_i^*, \theta)).$$

Attention

Consistency of the bootstrap procedure proposed by Hall and Horowitz (1996) was incomplete (they have only assumption about strong mixing of the process).
To prove generally and rigorously the consistency of above bootstrap procedure we need to have the weak dependence assumption (Doukhan, Louhichi, 2008)

2. Subsampling in Politis and McElroy model

The advantage of subsampling is its insensitivity to the form of the asymptotic distribution.

Model of Politis and McElroy (2007) is the combination of two phenomena:

- long memory time series (it means that there is dependence between distant observations)
- heavy-tailed time series

The model

Let $\{X_t\}_{t\in\mathbb{Z}} : X_t = \sigma_t G_t + \eta$ be a strictly stationary time series, where:

- σ_t and G_t are independent
- σ_t are i.i.d. $\alpha-$stable random variables, $\alpha \in (1, 2)$.
- $E(\sigma_t) \neq 0$
- G_t is long memory time series with parameter $\beta \in (0, 1)$
- G_t is purely nondeterministic with the finished variance (Gaussian with long memory).

Attention

The Gaussian models with long memory do not have the property of mixing. But such a models are weakly dependent.

Theorem

CLT (Politis, 2007) Assume that conditions 1. - 5. and $LM(\beta)$, where $\beta \in [0, 1)$ are fulfill.
Let $\rho \in (0, 1)$. Let
$\widehat{LM}(\rho) = | \sum_{|h|=1}^{[n^\rho]} \frac{1}{n-|h|} \sum_{t=1}^{n-|h|} (X_t X_{t+h} - \bar{X}^2) |^{1/\rho}$. Then:

$$(n^{-\zeta}\sum_{t=1}^n (X_t - \eta),\ n^{-2\zeta}\sum_{t=1}^n (X_t - \bar{X})^2,\ n^{-2\zeta+1}\widehat{LM}(\rho)) \Rightarrow_{\mathcal{L}}$$

$$\begin{cases} (S, U, 0), & \text{if } 1/\alpha > (\beta+1)/2 \\ (V, 0, \mu^{2/\rho}C^{1/\rho}), & \text{if } 1/\alpha < (\beta+1)/2 \\ (S+V, U, \mu^{2/\rho}C^{1/\rho}), & \text{if } 1/\alpha = (\beta+1)/2. \end{cases}$$

Where $\mu = E\sigma_t$, S is $\alpha-$stable random variable, V is Gaussian with the mean zero, and U is $\alpha/2-$stable random variable.

Self-normalized statistic T_n will also converge to a non-degenerate distribution.

Theorem

Let

$$T_n = \frac{\sqrt{n}(\bar{X} - \eta)}{\sqrt{\frac{1}{n}\sum_{t=1}^n (X_t - \bar{X})^2 + \widehat{LM}(\rho)}}$$

and an absolutely continuous random variable is in the form:

$$Q = \begin{cases} S/\sqrt{U}, & \text{if } 1/\alpha > (\beta+1)/2 \\ V/\mu^{2/\rho}C^{1/\rho}, & \text{if } 1/\alpha < (\beta+1)/2 \\ (S+V)/\sqrt{U+\mu^{2/\rho}C^{1/\rho}}, & \text{if } 1/\alpha = (\beta+1)/2, \end{cases}$$

then

$$T_n \Rightarrow_{\mathcal{L}} Q.$$

Consistency of subsampling method for Politis, McElroy model

Assumptions:

- the statistic $\hat{\theta}_n$ is an estimator of unknown parameter θ, with normalization $\hat{\sigma}_n$
- $\frac{\tau_n[\hat{\theta}_n - \theta]}{\hat{\sigma}_n}$ converges weakly to the random variable with distribution function J
- the empirical distribution function L_n, is computed from the sample of length n drawn from strictly stationary, $\theta-$weakly dependent time series with the weak dependence parameters $\theta_r = \mathbf{O}(r^{-a})$ for $a \ge \frac{1}{2}$
- $L_n(x) \to J(x)$ if $n \to \infty$
- $\alpha_n(\hat{\theta}_n - \theta)$ converges weakly to Z, where Z is random variable
- $\delta_n\hat{\sigma}_n$ converges weakly to W, where W is random variable and is positive with the probability 1
- $\{\alpha_n\}$ and $\{\delta_n\}$ are positive sequences: $\tau_n = \frac{\alpha_n}{\delta_n}$
- $\frac{\alpha_b}{\alpha_n} \to 0$, $\frac{\tau_b}{\tau_n} \to 0$, $\frac{b}{n} \to 0$, and $b \to 0$ if $n \to \infty$.

Theorem

(Politis, Jach, McElroy, 2011) If assumption 1-8 hold, then:

- If x is the point of the continuity J, then $L_{n,b}(x) \to J(x)$.
- If J is continuous then $\sup_x |L_{n,b}(x) - J(x)| \to 0$.
- If J is continuous in $c(1-\rho)$, then if $n \to \infty$

$$P(\tau_n(\hat{\theta}_n - \theta)/\hat{\tau}_n \le c_{n,b}(1-\rho)) \to 1-\rho.$$

A real data example (Politis, Jach, McElroy, 2011)

Data source: http://ita.ee.lbl.gov/html/traces.html.

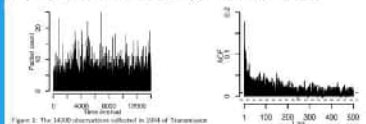

Figure 1: The 1400 observations collected in 1994 of Transmission Control Protocol (TCP) traffic between the Lawrence Berkeley Laboratory and the rest of the world.

Figure 2: ACF (TCP)

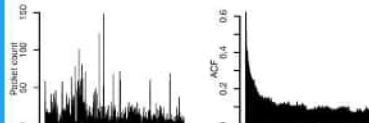

Figure 3: The OctExt Ethernet trace collected in 1994 at the Bellcore Morristown Research and Engineering (20000 observations)

Figure 4: ACF (OctExt)

The time series of packet-counts were obtained by counting the number of packets arriving in next intervals of a fixed length: 0.5s for TCP and 1s for OctExt.
Figures 2 and 4 suggest the presence of long-memory. Both series have heavy tails - Bates and McLaughilin (2000).

For those two time series the 95% confidence intervals for the mean were counted (Politis, Jach, McElroy, 2011).
For above time series nonoverlapping blocks of length 50 (28 subseries for TCP and 40 for OctExt) were created. In each chunk a 95% confidence interval for the mean was counted. The question was if the mean changes over time.

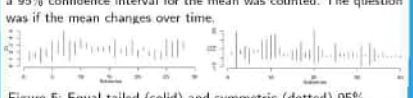

Figure 5: Equal-tailed (solid) and symmetric (dotted) 95% confidence intervals for the mean ($\rho = 0.8$).

We can observe that, the midpoint of the intervals changes, the interval width fluctuates.

Questions for the future

Questions

- For independent data and stationary time series resampling procedures are well investigated, but not for non stationary models. Many real-life phenomena is characterized by seasonal behavior.
  Should we improve the statistical tools for PC models? Can we substitute assumption about stationarity of the model from the poster for assumption about PC?
- Consistency of resampling methods for the $\alpha-$mixing PC models are well investigated.
  Do we obtain consistency of resampling methods if we substitute $\alpha-$mixing assumptions for weak dependence?

Speaker: Emmanuel Candès (Stanford University)

Title: “Robust principal component analysis?”

ABSTRACT: This talk is about a curious phenomenon. Suppose we have a data matrix, which is the superposition of a low-rank component and a sparse component. Can we recover each component individually? We prove that under some suitable assumptions, it is possible to recover both the low-rank and the sparse components exactly by solving a very convenient convex program. This suggests the possibility of a principled approach to robust principal component analysis since our methodology and results assert that one can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted. This extends to the situation where a fraction of the entries are missing as well. In the second part of the talk, we present applications in computer vision. In video surveillance, for example, our methodology allows for the detection of objects in a cluttered background. We show how the methodology can be adapted to simultaneously align a batch of images and correct serious defects/corruptions in each image, opening new perspectives.

Two stage threshold instrumental variables estimation of linear regression models

Emmanuel Guerre (QMUL)
Joint with Andrea Carriero and George Kapetanios (QMUL)

This paper proposes a choice of a suitable subset of instruments when no a priori ranking of these variables is available. The proposed estimator uses a simple thresholding two stage Jackknife procedure which is first-order efficient. It achieves a second-order rate which is very close to the ones of the second-order optimal IV estimators of Donald and Newey (2001) which assumes that the instruments are ranked. The threshold IV estimator is also robut to heteroscedasticity and weak instruments. A simulation study and sensitivity analysis reveal the good practical performances of the thresholding IV estimator.

Semiparametric Regression with Nonparametrically Generated Covariates

Enno Mammen*, Christoph Rothe and Melanie Schienle

University of Mannheim, Germany. *emammen@rumms.uni-mannheim.de*
Toulouse School of Economics, France.
Humboldt University Berlin, Germany.

Abstract:

In this talk, we study a general class of semiparametric optimization estimators of a vector-valued parameter. The criterion function depends on two types of infinite-dimensional nuisance parameters: a conditional expectation function that has been estimated nonparametrically using generated covariates, and another estimated function that is used to compute the generated covariates in the first place. Such estimators appear in numerous econometric applications, including nonparametric estimation of simultaneous equation models, sample selection models, treatment effect models, and censored regression models, but so far there seems to be no unified theory to establish their statistical properties. We study the asymptotic properties of estimators in this class, which is a non-standard problem due to the presence of generated covariates. We give conditions under which estimators are root-n consistent and asymptotically normal, derive a general formula for the asymptotic variance, and show how to establish validity of the bootstrap.

Eric Matzner-Lober

Title : Adaptive Multivariate Nonparametric Smoothers in Action

Multivariate non-parametric smoothers are adversely impacted by the sparseness of data in higher dimension, also known as the curse of dimensionality.
Adaptive smoothers, that can exploit the underlying smoothness of the regression function, may partially mitigate this effect.
We will present an iterative procedure based on traditional kernel smoothers, thin plate spline smoothers or Duchon spline smoother that can be used when the number of covariates is important.
However the method is limited to small sample sizes ($n<2000$) and we will propose some thoughts to circumvent that problem using for example preclustering of the data or using stochastic algorithm for linear algebra calculations such as SVD.

This is a join work with P-A. Cornillon and N. Hengartner

Testing independence of consecutive truncated and censored data

Ewa Strzalkowska-Kominiak
Department of Mathematics, Universidade da Coruña, Spain

Abstract

Let (U_1, U_2) be a vector of consecutive random variables with distribution function F so that U_1 may be truncated and $U_1 + U_2$ censored by a common variable Z. More precisely, U_1 can be observed only if $U_1 \leq Z$. In the other case, we say that U_1 is truncated from the right by Z. Furthermore, U_2 can be observed only if $U_1 + U_2 \leq Z$. Otherwise, we observe Z so that U_2 is censored from the right.

Basing on the observations $(U_{1i}, \tilde{U}_{2i}, Z_i, \delta_i)_{i=1,\ldots,n}$, where $\tilde{U}_{2i} = \min(U_{2i}, Z_i - U_{1i})$ and $\delta_i = 1_{\{U_{1i}+U_{2i}\leq Z_i\}}$, we proposed in [3] the following estimator of $F(x, y)$

$$F_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_{1i}\leq x, \tilde{U}_{2i}\leq y\}} \frac{\delta_i}{A_n(U_{1i} + \tilde{U}_{2i})}$$

where

$$A_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1_{\{Z_i\geq x\}}}{F_{1n}(Z_i)} \text{ and } F_{1n}(x) = \prod_{U_{1i}>x} \left[1 - \frac{1}{\sum_{j=1}^{n} 1_{\{U_{1j}\leq U_{1i}\leq Z_j\}}}\right].$$

The goal of this work is to model dependence between the times U_1 and U_2 using the copula function. In particular, basing on $F_n(x, y)$, we define the empirical copula as follow

$$C_n(u, v) = F_n(F_{1n}^{-1}(u), F_{2n}^{-1}(v)),$$

where F_{1n}^{-1} denotes the quantile function of the well-known Lynden-Bell estimator $F_{1n}(x)$ for right truncated data (see,e.g., [4]) and F_{2n}^{-1} is the quantile function of the marginal $F_{2n}(y) = F_n(\infty, y)$. We show the asymptotic properties of C_n and develop tests for independence between U_1 and U_2 based on Spearman's rho ρ_C and Cramér-von Mises statistic.

Keywords: Truncation, censoring, empirical copula, testing independence

References:

[1] Genest, C., Rémillard, B., Beaudoin, D. (2009) *Goodness-of-fit tests for copulas: A review and a power study*. Insurance: Mathematics and Economics 44, 199-213

[2] Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

[3] Strzalkowska-Kominiak, E., Stute, W. (2010) *The statistical analysis of consecutive survival data under serial dependence.* Journal of Nonparametric Statistics 22, 585-597

[4] Stute, W., Wang, J.-L. (2008). *The central limit theorem under random truncation*. Bernoulli 14, 604-622.

Asymptotics of the discrete log-concave maximum likelihood estimator and related applications

Fadoua Balabdaoui

May 22, 2012

CEREMADE, Université Paris-Dauphine, Paris, France

Abstract

The assumption of log-concavity is a flexible and appealing nonparametric shape constraint in distribution modeling. In this work, we study the log-concave maximum likelihood estimator (MLE) of a probability mass function (pmf). We show that the MLE is strongly consistent and derive its pointwise asymptotic theory under both the well– and misspecified settings. Our asymptotic results are used to calculate confidence intervals for the true log-concave pmf. Both the MLE and the associated confidence intervals may be easily computed using the R package **logcondiscr**. We illustrate our theoretical results using recent data from the H1N1 pandemic in Ontario, Canada.

Abstract Submission

Title: Semiparametric Functional Linear Model with High-Dimensional Covariates

Authors: Dehan Kong[†], Fang Yao[‡] (presenting), and Hao Helen Zhang[*].

Affiliations: [†] Department of Statistics, North Carolina State University; [‡] Department of Statistics, University of Toronto.

Abstract: We propose and study a new class of semiparametric functional regression models motivated by the complex nature of data encountered in modern scientific experiments. With a scalar response, multiple covariates are collected, a large number of which are time-independent and directly observed and a few may be functional with underlying processes. The goal is to jointly model the functional and non-functional predictors, identifying important scalar covariates while taking into account the functional covariate. In particular we exploit a unified linear structure to incorporate the functional predictor as in classical functional linear models that is of nonparametric feature. Simultaneously we include a potentially large number of scalar predictors as the parametric part that may be reduced to a sparse representation. We propose a simultaneous procedure to perform variable selection and estimation, by naturally combining the functional principal component analysis (FPCA) and the smoothly clipped absolute deviation (SCAD) penalized regression under one framework. Theoretical and empirical investigation reveals that the efficient estimation regarding important scalar predictors can be obtained and enjoys the oracle property, despite contamination of the noise-prone functional covariate. The study also sheds light on the influence of the number of eigenfunctions for modeling the functional predictor on the correctness of model selection and accuracy of the scalar estimates.

Keywords: Functional data analysis, Functional linear model, Model selection, Principal components, SCAD, Semiparametric regression

Specification Test for Partially Identified Models defined by Moment Inequalities*

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Xiaoxia Shi
Department of Economics
University of Wisconsin, Madison
xshi@ssc.wisc.edu

May 14, 2012

Abstract

This paper studies the problem of specification testing in partially identified models with special focus on models defined by a finite number of moment equalities and inequalities (i.e. (in)equalities). Under the null hypothesis of the test, there is at least one parameter value that simultaneously satisfies all of the moment (in)equalities whereas under the alternative hypothesis of the test, there is no such parameter value. While this problem has not been directly addressed in the literature (except in particular cases), several papers have suggested implementing this inferential problem by checking whether confidence intervals for the parameters of interest are empty or not.

We propose two hypothesis tests that use the infimum of the sample criterion function over the parameter space as the test statistic. The difference between these tests lies in the method to compute the critical values. One is straightforward to compute using existing set inference procedures but potentially conservative, while the other one is less asymptotically conservative but require a separate procedure to compute. We show that both of these hypothesis tests are: (a) asymptotically size correct in a uniform sense and (b) more powerful than tests that check whether the confidence intervals for the parameters of interest are empty or not.

KEYWORDS: Partial Identification, Moment Inequalities, Inference, Hypothesis Test, Specification Test.

JEL CLASSIFICATION: C01, C12, C15.

*We thank the participants at the 2011 Triangle Econometrics Conference, the 2012 Junior Festival on New Developments in Microeconometrics Conference, and seminar participants at IUPUI, UC Berkeley, UC Davis, Yale University, and Ohio State University for helpful comments. Bugni and Canay thank the National Science Foundation for research support via grants SES-1123771 and SES-1123586, respectively.

Conditional Akaike Information for Proportional Hazards Mixed Effects Models

Random effects have been widely used to analyze clustered data. In this talk we discuss the concept of conditional Akaike information (cAI) for proportional hazards mixed effects models, and we contrast it with the standard Akaike information. The criterion focuses on the parametric part of the model, while profiling out the nonparametric baseline hazard. We show how, based on the concept of effective degrees of freedom for generalized linear mixed models, a criterion can be defined in a straightforward way, and we show that this criterion is asymptotically unbiased for the cAI. Simulation studies show that the conditional Akaike information criterion has good operating characteristics. We illustrate its use on an example derived from a cancer clinical trial.

Key words: Cox model, conditional AIC, frailty models.

Speaker: Francesca Chiaromonte (Penn State University)

Title: Statistical characterizations of genome dynamics

Abstract: Recent studies of sequenced genomes have shown that the rates of different types of mutations tend to vary and co-vary along the nuclear DNA. Here, we provide statistical characterizations of this variation structure and of its association with features of the local genomic landscape using a number of multivariate techniques on data from primate alignments. These include linear and non-linear Principal Components of the mutation rates, linear and non-linear Canonical Correlations of rates and genomic features, and multivariate Hidden Markov Models to segment the genome based on mutation behavior. Our results provide crucial insights into genome dynamics, and the mutagenic role of various molecular mechanisms such as replication, recombination, repair and local chromatin environment. Moreover, we identify regions with distinct mutational signatures whose locations, lengths and landscapes are biologically meaningful.

Francesco Bravo

"Smoothed generalised empirical likelihood estimation and inference for semiparametric moment conditions models with dependent data"

This is the abstract of the paper

This paper considers the problem of estimation and inference for smooth semiparametric models with weakly dependent data. Estimation and inference is based on a two-step (or plug-in) smoothed generalized method of moment, generalised empirical likelihood and exponentially tilted methods. Smoothing is necessary to achieve asymptotic efficiency. The paper shows that the resulting estimators are asymptotically normal with covariance matrix that depends on whether there is estimation effect from the first step estimation of the infinite dimensional parameter. The paper proposes a number of test statistics that can be used to test various hypotheses of interests and to construct semiparametric versions of M-test. The distribution of some of the resulting test statistics depends crucially on the presence of the estimation effect. The paper illustrates the finite sample properties of some of the proposed estimators and test statistics with simulations using an instrumental variable partially linear model.

Bootstrapping Realized Bipower Variation

Gang Feng
Technische Universität Braunschweig, Institut für Mathematische Stochastik, Germany
e-mail: g.feng@tu-braunschweig.de

Realized bipower variation is often used to measure the volatility in financial markets with high frequency intraday data. Considering a nonparametric volatility model, we propose a nonparametric i.i.d. bootstrap procedure by resampling the noise innovations based on the discrete time returns, and a nonparametric wild bootstrap procedure by generating pseudo-noise that imitates correctly the first and second order properties of the ordinary noise, in order to approximate the distribution of the realized bipower variation. Asymptotic validity of the proposed procedures is proved. Furthermore, the finite sample properties of the proposals are investigated in a simulation study and are also compared to other bootstrap methods.
This is joint work with Jens-Peter Kreiss.

References

[1] O.E. Barndorff-Nielsen and N. Shephard *Power and bipower variation with stochastic volatility and jumps*. Journal of Financial Econometrics, 2, 1-48, 2004.
[2] S. Gonçalves and N. Meddahi Bootstrapping Realized Volatility *Econometrica*, 77:283–306, 2009.
[3] M. Podolskij and D. Ziggel *Bootstrapping Bipower Variation*. Ruhr-University Bochum, 2007.

Estimation of Small Area Means under Semi-Parametric Measurement Error Models

G.S. Datta [1], P. Hall [2] and L. Wang [3]

Abstract

In recent years, demand for reliable estimates for characteristics of small domains (small areas) has considerably increased worldwide due to the growing use of such estimates in formulating policies and programs, allocating government funds, planning regional development, and making marketing decisions at the local level. However, due to cost and operational considerations, it is rarely possible to get a large enough sample at the small area level to support direct estimates with adequate precision for all domains of interest. Model-based inference has gained immense popularity in producing indirect but reliable small area estimates. These indirect estimates borrow strength from related areas and other data source by linking them through appropriate models. Existing methods in small area estimation are mostly parametric, and they usually treat the explanatory variables as if they are measured without error. However, explanatory variables are often subject to measurement error. A few authors have addressed the measurement error problem in small area estimation through parametric approach based on the normality assumption. Resulting estimates are usually sensitive to the distributional assumptions. In this talk, we consider structural measurement error models and a semi-parametric approach to produce reliable point estimates and prediction intervals for small area means. Specifically, we consider an adaptation of the Fay-Herriot model for the area-level data where one of the covariates is measured with error. We replace the normality assumption of the sampling error and the normality assumption of the measurement error of a covariate by heavy-tailed distributions. Estimating the unknown measurement error density nonparametrically, we develop both point estimates and prediction intervals of small area means. We have obtained an expansion of the coverage error of the proposed prediction intervals.

[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email:gauri@uga.edu

[2]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia. email:halpstat@ms.unimelb.edu.au

[3]Department of Statistics, University of Georgia, Athens, GA 30602, USA. email:lilywang@uga.edu

George Kapetanios

Adaptive forecasting in the presence of recent and ongoing structural change joint with Liudas Giraitis and Simon Price

We consider time series forecasting in the presence of ongoing structural change where both the time series dependence and the nature of the structural change are unknown. Methods that downweight older data, such as rolling regressions, forecast averaging over different windows and exponentially weighted moving averages, known to be robust to historical structural change, are found to be also useful in the presence of ongoing structural change in the forecast period. A crucial issue is how to select the degree of downweighting, usually defined by an arbitrary tuning parameter. We make this choice data dependent by minimizing forecast mean square error, and provide a detailed theoretical analysis of our proposal. Monte Carlo results illustrate the methods. We examine their performance on 191 UK and US macro series. Forecasts using data-based tuning of the data discount rate are shown to perform well.

Asymptotic properties of the estimation of the error distribution in right censored and selection biased regression models

C. Heuchenne and G. Laurent

C.Heuchenne@ulg.ac.be G.Laurent@student.ulg.ac.be
QuantOM, HEC-Management School-University of Liège,
Rue Louvrex, 14 (Bât N1), B-4000 Liège, Belgium

Abstract

Suppose the random vector (X, Y) satisfies the nonparametric regression model $Y = m(X) + \sigma(X)\varepsilon$ where $m(X) = E[Y|X]$ and $\sigma^2(X) = Var[Y|X]$ are unknown smooth functions and the error ε, with unknown distribution, is independent of the covariate X. The pair (X, Y) is subject to generalized bias selection and the response Y to right censoring. We define a new estimator for the cumulative distribution function of the error ε, where the estimators of $m(\cdot)$ and $\sigma^2(\cdot)$ are obtained by extending the conditional estimation methods introduced in de Uña-Alvarez and Iglesias-Perez (2010). The asymptotic properties of the proposed estimator are established. A bootstrap technique is proposed to select the smoothing parameter involved in the procedure. Finally, this method is studied via extended simulations and applied to real data.

Reference

de UNA-ALVAREZ, J., IGLESIAS-PEREZ, M.C. (2010): Nonparametric estimation of a conditional distribution from length-biased data. *Annals of the Institute of Statistical Mathematics, Vol. 62, 323-341.*

Nonparametrically consistent depth-based classifiers.

Germain Van Bever [1] & Davy Paindaveine [2]

[1] *Departement of Mathematics.*
Université libre de Bruxelles
Av. F.D. Roosevelt, 50, CP210, 1050 Brussels, Belgium.
Email: gvbever@ulb.ac.be
[2] *Ecares and Departement of Mathematics.*
Université libre de Bruxelles
Av. F.D. Roosevelt, 50, CP114/04, 1050 Bruxelles, Belgique.
Email: dpaindav@ulb.ac.be

Abstract. In this talk, we focus on one of the most natural applications of depth, namely classification, and present the methods introduced in Ghosh and Chaudhuri (2005) and Li et al. (2012). We propose a class of depth-based classification procedures that are of a nearest neighbor nature. The corresponding nearest neighbors are identified through a symmetrized depth construction. Our classifiers enjoy (i) the robustness and affine-invariance of depth-based procedures and (ii) the good asymptotic properties of nearest-neighbor classifiers. We investigate their finite-sample performances through simulations.

Keywords. Depth, Classification, Nearest Neighbor.

Bibliography

[1] G. Biau, L. Devroye, V. Dujmović, and A. Krzyzak, An affine-invariant k-Nearest Neighbor regression estimate. Soumis.
[2] A. K. Ghosh and P. Chaudhuri, On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32:327–350, 2005.
[3] J. Li, J. A. Cuesta-Albertos, and R. Y. Liu, DD-classifier: nonparametric classification procedure based on DD-plot. Submitted.
[4] R. Y. Liu, On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414, 1990.
[5] R. Liu, J. Parelius, and K. Singh, Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics*, 27:783–840, 1999.
[6] R. H. Oja and D. Paindaveine, Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323, 2005.
[7] J. W. Tukey, Mathematics and picturing data. *Proceedings of the International Congress on Mathematics* (R.D. James, ed.), Canadian Math. Congress, 2:523–531, 1975.

[8] Y. Zuo, and R. Serfling, General notions of statistical depth functions. *The Annals of Statistics*, 28:461–482, 2000.

Gilles Fay

Nonparametric anisotropy detection on the sphere

Ultra-high energetic cosmic rays are very rare directional events (points on the sphere) whose origin is still mysterious. We propose a nonparametric detection procedure for testing the isotropic distribution of the angles of incidence. This multiple test is based on multiple and linear wavelet estimations of the density at finer and finer scales.

On moment estimators for the parameters of the quadratic polynomial within the integrated Pearson family, with applications in testing hypothesis*

Giorgos Afendras

Department of Mathematics, Section of Statistics and O.R.
University of Athens
Panepistemiopolis, 157 84 Athens, Greece
e-mail: g_afendras@math.uoa.gr

Abstract

For a random variable X in the integrated Pearson family, the mean μ together with the quadratic polynomial q characterize the distribution. Using a Stein-type covariance identity we obtain moment estimators for the parameters of Pearson's quadratic polynomial and we calculate the variance matrix of their asymptotic distribution. Next, we provide hypothesis testing based on the above estimators and their limiting distribution. Among them, our interest focuses in testing normality. Some simulation studies are used to compare the performance of the proposed tests with other existing tests of normality.

*based on a common work with N. Papadatos, H. Papageorgiou and V. Papathanasiou

LOCALLY STATIONARY LATENT FACTORS*

Giovanni Motta†
Maastricht University

Abstract

Current approaches for fitting dynamic factor models to multivariate time series are based on the principal components of the spectral matrix. These approaches rely on the assumption that the underlying process is temporally stationary which appears to be restrictive because, over long time periods, the parameters are highly unlikely to remain constant.

The more general model introduced by Eichler, Motta and Sachs (2011) allows the spectral matrix to be smoothly time-varying, which imposes very little structure on the moments of the underlying process. However, the estimation becomes fully non-parametric and delivers time-varying filters that are high-dimensional and two-sided. Moreover, the estimation of the spectral matrix strongly depends on the chosen bandwidths for smoothing over frequency and time.

As an alternative, we propose a semi-parametric approach in which only part of the model is allowed to be time-varying. More precisely, the latent factors admit a dynamic representation with time-varying autoregressive coefficients while the loadings are constant over time.

Estimation of the model parameters is accomplished by application of the EM algorithm and the Kalman filter. The time-varying parameters are modeled locally by polynomials and estimated by maximizing the likelihood locally. Simulation results show that compared to estimation by principal components, our approach produces superior results in particular for small cross-sectional dimensions. We illustrate the performance of our approach through applications to real data.

This is a joint work with Michael Eichler (Maastricht University).

*This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme.

†*E-mail:* g.motta@maastrichtuniversity.nl

Cross-Validation as an Alternative to Out-of-Sample Inference Under Instability

Helle Bunzel
Iowa State University
CREATES

Gray Calhoun
Iowa State University

May 3, 2012

Abstract

This paper studies the behavior of pseudo out-of-sample statistics for testing hypotheses about models' forecasting performance, such as those developed by Diebold and Mariano (1995, *J. Bus. Econ. Statist.*) and West (1996, *Econometrica*), when the Data Generating Process is subject to a one-time structural break. We find that these methods are unreliable unless the out-of-sample period is very short, which is at odds with current empirical practice and severely limits their use. Specifically, the prebreak observations are weighted too heavily unless $P = O(\sqrt{T})$, where P is the number of out-of-sample observations and T the total sample size. We also propose a new cross-validation method that does not have this limitation. Monte Carlo simulations demonstrate that our method has better finite-sample size and higher power than existing out-of-sample statistics.
JEL Classification: C12, C22, C52, C53
Keywords: Forecasting, Model Selection, Structural Breaks

Cross-Validation as an Alternative to Out-of-Sample Inference Under Instability

Helle Bunzel
Iowa State University
CREATES

Gray Calhoun
Iowa State University

May 3, 2012

Abstract

This paper studies the behavior of pseudo out-of-sample statistics for testing hypotheses about models' forecasting performance, such as those developed by Diebold and Mariano (1995, *J. Bus. Econ. Statist.*) and West (1996, *Econometrica*), when the Data Generating Process is subject to a one-time structural break. We find that these methods are unreliable unless the out-of-sample period is very short, which is at odds with current empirical practice and severely limits their use. Specifically, the prebreak observations are weighted too heavily unless $P = O(\sqrt{T})$, where P is the number of out-of-sample observations and T the total sample size. We also propose a new cross-validation method that does not have this limitation. Monte Carlo simulations demonstrate that our method has better finite-sample size and higher power than existing out-of-sample statistics.
JEL Classification: C12, C22, C52, C53
Keywords: Forecasting, Model Selection, Structural Breaks

Speaker: Gregory Ryslik (Yale University)

Talk Title: Identification of Non-Random Somatic Mutation Clustering in Proteins.

Abstract: Human cancer is a genetic disease that is caused by the accumulation of somatic mutations in tumor suppressors and oncogenes. In the case of oncogenes, recent theory suggests that there are only a few key "driver" mutations responsible for tumorigenesis. As there have been significant pharmacological successes in treating these driver mutations several methods that rely on mutational clustering have been developed to identify them. We will discuss the application of these methods to a recent study on prostate cancer as well as explore extensions that would utilize the 3D structure of the protein in order to find both novel clusters and combine previously disparate clusters together.

Nonparametric Variable Selection in High Dimensional Data

Haiyan Wang
Kansas State University

The selection of variables for regression and classification in high dimensional data is of great importance with vast data collection techniques available due to high performance computing. An example is the classification of phenotypes for diseases with gene expression data. In this talk I will present two nonparametric methods for variable selection, one that allows both continuous and discrete response while the other one is specific for categorical response. The first method, motivated by Least Angle Regression, is a multi-step nonparametric model selection algorithm to select variables in sparse ultra-high dimensional additive models. The variables go through a series of nonlinear dependence evaluation following a Most Significant Regression algorithm. The predictors are linearly or nonlinearly related to the response. Some properties of the algorithm will be discussed. Simulation results demonstrate that this algorithm has comparable true positive rate and much lower false positive rate than a few recently developed other methods such as penGAM, INIS, g-INIS (see Fan, Feng, and Song, JASA 2011). The second method considers variable selection when the response variable has multiple categories. The variables are selected with a multi-step ranking of top-scoring variables with leave-one-out classification accuracy within the training data as the objective function. Application of the methods in 10 classic cancer data sets achieved excellent leave-one-out classification accuracy.

Multiscale Estimation for Simultaneous Change-point Detection in Exponential Families

Hannes Sieling
University of Goettingen

In change-point detection the regression function is assumed to be piecewise constant with an arbitrary number of change-points. This model hence provides for discontinuous jumps at the change-points. The statistical problem consists in estimating the number of change-points and their locations. We provide a locally adaptive approach which is built on a multiscale likelihood statistic. By a dynamic programing algorithm the method becomes feasible for sequences of considerable length and, as a by-product, we obtain confidence bands for the regression function. We further give asymptotic results and illustrate the capability of the proposed method by various examples and simulations.

On multivariate signs and ranks

Hannu Oja

University of Tampere, Finland

Keywords: affine equivariance; affine invariance; efficiency; L_1 objective function; multivariate Hodges-Lehmann estimate; multivariate median; robustness.

The univariate concepts of sign and rank are based on the natural ordering of the univariate data. In the multivariate case there is no natural coordinate-free ordering of the data points. In this talk multivariate extensions of sign and rank which are based on different multivariate L_1 objective or criterion functions are considered. Different multivariate sign and rank based tests and the companion estimates, multivariate extensions of the median and the Hodges-Lehmann estimators, are discussed and compared. The equivariance and invariance properties, efficiencies, and robustness of the tests and estimates are also briefly discussed. Presentations of multivariate methods based on different multivariate signs and ranks can be found in Puri and Sen (1971), Oja (1999), Oja and Randles (2004), Oja (2010), and Hettmansperger and McKean (2010).

References

Hettmansperger, T.P. and McKean, J.W. (2010), *Robust Nonparametric Statistical Methods*, Second Edition, Crc Press, New York.

Oja, H. (1999), Affine invariant multivariate sign and rank tests and corresponding estimates: a review, *Scandinavian Journal of Statistics*, **26**, 319–343.

Oja, H. (2010), *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks*. Springer, New York.

Oja, H. and Randles, R. (2004), Multivariate nonparametric tests. *Statistical Science*, **19**, 598–605.

Puri, M.L. and Sen, P.K. (1971), *Nonparametric Methods in Multivariate Analysis*, Wiley, New York.

Two sample problem for mean location

Harrie Hendriks
Radboud University Nijmegen

Abstract

We will report on work with Zinoviy Landsman. Suppose given a compact submanifold M embedded in some Euclidean space. Associated with a random variable X on M is a mean location μ satisfying the condition that the expected squared distance $\mathbb{E}(\|X-\mu\|^2)$ is minimal. Given an other random variable Y on M we will construct an asymptotic test of the hypothesis that the mean locations of X and Y coincide. We consider the mean locations μ and ν of X and Y respectively, to be unknown. The test refers to data consisting of independent random samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ of X and Y respectively, and asymptotics is understood as $\min(m,n)\to\infty$.
First issue is that the limit behavior with respect to μ of the empirical mean location $\overline{X}$ of the sample of X is described in the tangent space to M at $\overline{X}$. On the other hand the behavior of the sample of Y is described in the tangent space to the empirical sample mean $\overline{Y}$. These are different tangent spaces.
Second issue is that the usual reference to Slutsky's Theorem or the Continuous Mapping Theorem refers to the existence of a limit distribution of the terms involved in the test statistic. In our situation the limit distribution need not exist, especially if m and n grow in such a way that m/n is not converging. This issue is already relevant in the asympotical approach to the most basic two sample problem for univariate distributions with unequal variance.

New estimators of the Pickands dependence function and a test for extreme-value dependence

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@ruhr-uni-bochum.de

April 4, 2012

Abstract

Pickands dependence function A is convex and satisfies the boundary conditions

$$\max\{t, 1-t\} \leq A(t) \leq 1$$

for $t \in [0,1]$. We propose a new class of estimators for Pickands dependence function which is based on the best L^2-approximation of the logarithm of the copula by logarithms of extreme-value copulas. The estimators $\hat{A}(t)$ are obtained by replacing the unknown copula by its empirical counterpart and weak convergence of the process $\sqrt{n}\{\hat{A}(t) - A(t)\}_{t\in[0,1]}$ is shown. A comparison with the commonly used estimators is performed from a theoretical point of view and by means of a simulation study. Our asymptotic and numerical results indicate that some of the new estimators outperform the rank-based versions of Pickands estimator and an estimator which was recently proposed by Genest and Seegers (2009). As a by-product of our results we obtain a simple test for the hypothesis of an extreme-value copula, which is consistent against all alternatives with continuous partial derivatives of first order satisfying $C(u,v) \geq uv$.

Empirical likelihood for time series

Hugo Harari-Kermadec & Jacek Leśkow

Hugo Harari-Kermadec, ENS Cachan et Laboratoire SAMM, Université Paris 1, 90, rue de Tolbiac, 75634 PARIS CEDEX 13

Key words: semi-parametric statistics; PARMA; non-stationarity.

Abstract:

In this talk, we show how Empirical Likelihood can be used on times series, for different kinds of dependence. In particular, we introduce an algotithm to deal with non-stationary times series, when the non-stationarity is due to a periodicity. PARMA are classical models with this kind of non-stationnarity. A key paper is played by the construction of data blocs adapted to the dependence structure.

Estimation of fatigue behaviour based on a parametric model for the inverse relation

Ida Hertel*, Michael Kohler and Markus Kontny

Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: hertel@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de, kontny.markus@gmail.com

May 9, 2012

Abstract

In experimental fatigue life assesment the relation between strain-amplitude and number of cycles till failure is often investigated based on a parametric model for the inverse relation. In this paper a standard least squares estimate based on the inverse functions of the parametric model is defined, and it is shown how this estimate can be computed approximately using algorithms from optimization. Furthermore, a simulation approach is proposed for choosing the strain-amplitudes used in the experiments. The estimate is evaluated on real and simulated data.

*Corresponding author. Tel: +49-6151-16-2294, Fax: +49-6151-16-6822

Running title: *Estimation of fatigue behaviour*

Aliasing and locally stationary time series

Idris A. Eckley
Lancaster University, U.K.

Aliasing occurs when power exists in a signal at frequencies higher than the Nyquist rate (which is determined by the sampling rate). When it occurs, aliasing causes high frequency information to wrap round and mimic power at lower frequencies. It is all too easy to overlook aliasing when conducting an analysis of a time series. Indeed it is rarely tested for, even though a bispectrum-based test of aliasing for (stationary) time series was proposed by Hinich and Wolinsky in 1988. In this talk we consider recent work on aliasing in locally stationary wavelet time series.

My Journal
arXiv: math.PR/0000000

SPECIFICATION TESTS FOR NONLINEAR TIME SERIES MODELS*

By Igor Kheifets†

New Economic School

We propose a new model adequacy test for parametric conditional distributions in nonlinear time series models. Uniformity and independence of pseudo-residual series obtained by applying the conditional probability integral transform are simultaneously checked by means of continuous functionals of a bi-parameter empirical process of contemporaneous and lagged transforms. We establish weak convergence of the empirical process under parameter uncertainty. The tests have power against local alternatives converging under the null with a parametric rate. We justify a parametric bootstrap approximation that accounts for parameter estimation effects. Monte Carlo experiments show that the test has good size and power properties. We check adequacy of various heteroscedastic models for stock exchange index data.

1. Introduction. Suppose that we have a real-valued time series Y_t, $t = 0, \pm 1, \pm 2, \ldots$. Let $\Omega_t = \sigma(Y_{t-1}, Y_{t-2}, \ldots)$ be the σ-field generated by the observations obtained up to time t (information set at time t, not including Y_t). We consider the family of conditional distribution functions $F_t(y|\Omega_t, \theta)$, parameterized by $\theta \in \Theta$, where $\Theta \subseteq R^L$ is a finite dimensional parameter space. We permit changes over time in the functional form of the distribution using subscript t in F_t. We want to test the correct specification of the conditional distribution:

H_0 : The conditional distribution of Y_t conditional on Ω_t is in the parametric family $F_t(y|\Omega_t, \theta)$ for some $\theta_0 \in \Theta$.

Testing the specification of nonlinear time series models is crucial in applied statistics, macroeconomics and finance to make relevant analysis. It is often not enough to check only conditional moments, while specification of the *conditional*

*Footnote to the title with the thankstext command.

†I would like to thank Carlos Velasco for his excellent supervision. I also thank Miguel A. Delgado, Abderrahim Taamouti, Manuel A. Dominguez, Oliver Linton, Stefan Sperlich and Vanessa Berenguer for their comments and Yuichi Kitamura for organizing my visit to Yale in 2008. I acknowledge financial support from the Spanish Ministerio de Educación y Ciencia, Ref. no. BES-2006-12932 and 2007/04329/001, SEJ2007-62908.

AMS 2000 subject classifications: Primary 62M10, 62G10; secondary 62G30

Keywords and phrases: Goodness-of-fit test, conditional distribution, ARCH processes, bootstrap, Rosenblatt transforms

Boundary estimation in the presence of measurement error with unknown variance

Ingrid Van Keilegom

Institut de statistique
Université catholique de Louvain
Voie du Roman Pays 20
1348 Louvain-la-Neuve
Belgium

ingrid.vankeilegom@uclouvain.be

Boundary estimation appears naturally in economics in the context of productivity analysis. The performance of a firm is measured by the distance between its achieved output level (quantity of goods produced) and an optimal production frontier which is the locus of the maximal achievable output given the level of the inputs (labor, energy, capital, etc.). Frontier estimation becomes difficult if the outputs are measured with noise and most approaches rely on restrictive parametric assumptions. This paper contributes to the direction of nonparametric approaches.

We consider a general setup with unknown frontier and unknown variance of a normally distributed error term, and we propose a nonparametric method which allows to identify and estimate both quantities simultaneously. The asymptotic consistency and the rate of convergence of our estimators are established, and simulations are carried out to verify the performance of the estimators for small samples. We also apply our method on a dataset concerning the production output of American electricity utility companies.

This is joint work with Alois Kneip and Léopold Simar.

Key words: deconvolution, stochastic frontier estimation, nonparametric estimation, penalized likelihood

Non-Parametric Kernel Regression: A Simple and Robust Approach for Inference in Predictive Regression

Elena Andreou
University of Cyprus

and

Ioannis Kasparis
University of Cyprus

and

Peter C. B. Phillips
Yale University, University of Auckland
University of Southampton & Singapore Management University

Version: May 14, 2012

A unifying framework for inference is developed in predictive regressions where the predictor's integration properties are unknown. Two simple nonparametric F-tests, similar to those considered by Kasparis and Phillips (JoE, 2012) are proposed for testing predictability in predictive regressions. The test statistics are based on functionals of the Nadaraya-Watson kernel regression estimator. The limit distribution of the predictive tests is invariant with respect to the exact integration order of the predictor. In fact, limit distribution is standard for a wide range of predictors including stationary as well as non-stationary fractional and near unit root process. In this sense the proposed tests provide a unifying framework for inference. Moreover, this approach allows studying and testing a possibly nonlinear relationship, of unknown form. Therefore, tests are robust to integration order and functional form. The limit distribution of the tests, under the null hypothesis (predictability), is determined by functionals of independent χ^2 variates. Under the alternative hypothesis (predictability), the tests are consistent. Some theoretical and simulation results provided show that the proposed nonparametric tests are more powerful than existing parametric predictability tests, when deviations from unity are large or the predictive regression is nonlinear. The tests are applied to monthly S&P500 stock returns data and a comprehensive set of predictors used in the literature (e.g. Welch and Goyal, 2008) over the period 1926-2010 as well as various subsamples. We find significant and robust stock market predictability evidence for the following predictors: the Dividend Price ratio, the smoothed Earnings Price ratio, the Default Yield spread, the CPI Inflation rate and the monthly stock market Realised Volatility.

Functional data and conditional dependencies and association measures

Irène Gijbels, Marek Omelka and Noël Veraverbeke*

*Department of Mathematics and Leuven Statistics Research Centre, Katholieke Universiteit Leuven, Belgium
E-mail: irene.gijbels@wis.kuleuven.be

The dependence structure between two continuous random variables can be modelled via the copula function. In this talk we consider the situation in which this dependence structure is influenced by a functional covariate. We discuss nonparametric estimation of the conditional copula function and associated conditional association measures, such as a conditional Kendall's tau. Asymptotic properties, including bias, variance and weak convergence, are established. The finite sample performance of the estimators is demonstrated in a simulation study and the use of the methodology is illustrated in the analysis of a real data example.

Speaker: Ismael Castillo,

Affiliation: University of Paris, France

Title: On nonparametric adaptive Bayesian estimation

Abstract : In a recent article by A. van der Vaart and H. van Zanten, the authors show that adaptive convergence rates for estimating a function defined on $X=[0,1]$ in terms of "testing"-distances can be achieved using smooth Gaussian priors with well-chosen random rescaling (or path-shrinkage). In this work, we explore properties of families of rescaled priors for some other state-spaces X.

This is joint work with GØrard Kerkyacharian and Dominique Picard (LPMA, Paris)

L_1-regression for multivariate clustered data

Jaakko Nevalainen[1], Klaus Nordhausen[2] and Hannu Oja[2]

[1]University of Turku and [2]University of Tampere

Abstract

We consider the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

with a p-variate response variable, q-variate vector of explanatory variables and a p-variate random error. The goal is to make inference on the unknown qxp regression coefficient matrix $\boldsymbol{\beta}$. Commonly, the estimation of the parameters is based on L_2 (optimal in the the case of iid normal errors) or L_1 (a robust alternative) objective functions. However, the standard assumption of the independence of the random errors does not hold if the data are clustered; they are correlated. In order to make valid inference, this feature of the data must be accounted for.

In this talk we review the multivariate L_1 regression theory in the case of iid error variables, and then we extend the asymptotic theory to the cluster-correlated errors case. Transformation-retransformation (TR) estimates of the regression coefficients, with carefully selected transformation matrices, result in improved L_1 regression estimates in the sense that: (i) the estimates are affine equivariant and (ii) they are more efficient than ordinary L_1 regression estimates. We discuss and compare different approaches for constructing the TR estimates with clustered data. The theory is illustrated with a simulation study and some examples.

Subsampling-based frequency identification for nonstationary time series

Jacek Leśkow
Department of Econometrics,
The Nowy Sacz Graduate School of Business, Nowy Sącz, Poland.

March 11, 2012

Recently, the cyclostationary and almost periodically correlated time series and signals have attracted attention in engineering sciences like telecommunication and vibromechanics. From the inferential point of view, in such models it is important to identify arguments in the bispectrum of such time series that are significant for modelling. This leads to the problem of establishing tests of significance for frequencies in the bispectral domain analysis for cyclostationary. We will provide tests based on subsampling approximation of the limiting distribution of such statistics. The asymptotic results will be accompanied with applications to wheel bearing diagnostics.

Keywords: bispectrum analysis, frequency identification, subsampling approximation.

Author: Jan Beran
Title: Rapid change points under long memory

Abstract:

This is joint work with Sucharita Ghosh and Patricia Menendez.

Motivated by questions in paleoclimatology, we consider inference for rapid change points. The underlying continuous time model is observed at rregularly spaced discrete time points. The model is locally stationary in a generalized sense, and includes changing residual distributions together with Gaussian (or more general) subordination and long memory.
Rapid change points are estimated using a Priestley-Chao approach for derivatives. Asymptotic results are derived together with optimal bandwidth selection. The method is illustrated by an application to measurements of oxygen isotopes trapped in the Greenland ice sheets in the last 20 000 years (Greenland Ice Core Project).

Another look at bootstrap confidence intervals

Jan W.H. Swanepoel
North-West University, Potchefstroom, South Africa

It is well known that the m/n bootstrap is often useful, not only in cases where the traditional n/n bootstrap fails, but also in situations where it is valid. In this talk we apply the former approach to correct for coverage error in the construction of bootstrap confidence intervals for a paramater in a smooth function model setting. We show, for example, that the coverage error of a classical bootstrap percentile confidence bound, which is typically of order $O\left(n^{-1/2}\right)$, can be reduced to $O\left(n^{-1}\right)$ by applying this new approach. Rigorous arguments suggest that the optimal value of m is often substantially smaller than n and hence the construction of the new intervals will require less computational effort than that of traditional intervals. A simulation study is conducted to illustrate our findings.

Tests for continuity of regression functions

Jaromir Antoch and Marie Hušková

Department of Statistics, Faculty of Mathematics and Physics, Charles University of Prague, Sokolovská 83, CZ-186 75 Praha 8, Czech Republic

antoch@karlin.mff.cuni.cz, huskova@karlin.mff.cuni.cz

Keywords: Non-smooth regression, local linear estimators, limit properties

We consider the regression model:

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $(X_1, Y_1), \dots, (X_n, Y_n)$, $1 \le i \le n$, are i.i.d. random vectors with $\varepsilon_1, \dots, \varepsilon_n$ being i.i.d. random errors with

$$E\varepsilon_i = 0,\ var\varepsilon_i = 1 \quad \text{and} \quad E|\varepsilon_i|^{2+\Delta} < \infty \tag{2}$$

with some $\Delta > 1/2$.

The regression function $m(\cdot)$, the variance function $\sigma^2(\cdot)$ and the density $f(\cdot)$ of X_i are unknown functions that are supposed to be smooth except possibly a finite number of points. The density $f(\cdot)$ is bounded away from 0 on $[0, 1]$ and is equal to 0 outside of $[0, 1]$. We are interested in the testing problem concerning the smoothness of the regression function $m(\cdot)$, i.e., we want to test

$$H_0 : m(\cdot) \text{ is smooth function on } [0, 1] \tag{3}$$

against

$$H_1 : m(\cdot) \text{ has at least one jump in } (0, 1). \tag{4}$$

The aim of the lecture is to develop procedures for testing problem H_0 vs H_1 based on nonparametric estimators of $m(\cdot)$ and $\sigma^2(\cdot)$, particularly, based on their locally linear estimators.

Most of the papers in the area of nonparametric regression deals with efficient estimation of smooth regression curves. However, as soon as one suspects that the regression function $m(\cdot)$ can have some discontinuity points, one should start with testing continuity versus discontinuity of the regression function, eventually to estimate location(s) of jump(s) and, finally, to estimate the regression function. There are only few papers concerning this testing problem, see, e.g., Müller and Stadtmüller (1999) or Horváth and Kokoszka (2002) among others. Our test procedure is related to the procedure proposed by Horváth and Kokoszka (2002), however, we admit a more general model. Moreover, we provide also estimation of location of jumps. Particularly, our procedure is based on the supremum of properly standardized absolute values of the difference of the one-sided local linear estimators of the regression function $m(\cdot)$. We derive the limit distribution of the proposed test statistic under the null hypothesis which appears to belong to extreme value type ones. This provides approximations for the critical values. Results of a small simulation study will be discussed as well.

References

[1] Antoch J., Gregoire G, and Hušková M. (2007) Tests for continuity of regression functions. JSPI **137**, 753 – 777.

[2] Horváth L. and Kokoszka P. (2002) Change-point detection in non-parametric regression. Statistics **36**, 9 – 31.

[3] Maciak M. (2011) Flexibility, Robustness and Discontinuities in Nonparametric Regression Approaches. PhD thesis, Charles University of Prague

[4] Müller H.G. and Stadtmüller U. (1999) Discontinuous versus smooth regression. Ann. Statist. **27**, 299 – 337.

Testing for equality of an increasing number of spectral density functions

J. Hidalgo
Economics Department
London School of Economics, U.K.
f.j.hidalgo@lse.ac.uk
7 June 2012}}
keywords{Spectral density function, structural break.}

Abstract
The aim of this paper is to examine a nonparametric test for the equality among an increasing number of spectral density functions. One example of interest is when we wish to test the constancy of the covariance structure of a time series sequence. A second example of relevance is with spatio-temporal data \ or panel data models and we wish to decide if the dynamic structure along time does not depend on the individual or the location in space. Another situation is if we wish to test for separability.

Two features of the test are that $\left(a\right) $ even when the curve is not parametric under the null hypothesis, there is no need to choose any bandwidth parameter. The reason being that we can make use of the fact that the dimension increases to infinity, and $\left(b\right) $ the asymptotic distribution of the test is pivotal and under some mild conditions it converges to a normal random variable. Finally, we present a Monte-Carlo experiment to illustrate the finite sample performance of the test as well as a valid bootstrap procedure.

Jean Pierre Florens

Title - "Nonparametric Instrumental Derivatives"

Abstract - The focus of this paper is the nonparametric estimation of the marginal effects (i.e. first partial derivatives) of an instrumental regression function φ defined by conditional moment restrictions that stem from a structural econometric model $E[Y - \varphi(Z)|W] = 0$, and involve endogenous variables Y and Z and instruments W . The derivative function φ' is the solution of an ill-posed inverse problem and we propose an estimation procedure based on Landweber-Fridman regularization. The paper presents theoretical properties of the estimated nonparametric instrumental marginal effects, examines finite-sample performance, and considers an illustrative application.

The maxiset approach in nonparametric function estimation

Jean-Marc Freyermuth[*]

[*] ORSTAT, K.U.Leuven, Belgium, E-mail: Jean-Marc.Freyermuth@econ.kuleuven.be

In this talk we present some recent theoretical results about wavelet thresholding estimators. More precisely, for some selected estimators, we provide the maximal function space (maxiset) for which their quadratic risk reach a given rate of convergence. This recent approach, introduced by Cohen et al. (2001), has been proved useful to differentiate between estimators with equivalent minimax properties and to explain well how they behave in a pratical setting. We particularly emphasize the importance of pooling information from geometric structures in the coefficient domain to decide whether to keep/kill coefficients to obtain 'large' maxisets (see e.g. Autin et al. (2011a,b,c)). Finally, in the spirit of Schneider and von Sachs (1996), Neumann and von Sachs (1997), we emphasize the potential of such methods to adaptively estimate an evolutionary spectrum of a locally stationary process by 2-D wavelet smoothing over time and frequency.

References

Cohen, A., De Vore, R., Kerkyacharian, G., and Picard, D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, **11**, 167-191.

Autin, F., Freyermuth, J-M., von Sachs, R. (2011a). Ideal denoising within a family of Tree Structured Wavelets. *Electron. Journ. of Statist.*, vol. 5, 829-855.

Autin, F., Freyermuth, J-M., von Sachs, R. (2011b). Block-threshold-adapted estimators via a maxiset approach. *Submitted.*

Autin, F., Freyermuth, J-M., von Sachs, R. (2011c) Combining thresholding rules: a new way to improve the performance of wavelet estimators. *Submitted*.

Schneider, K., von Sachs, R. (1996). Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Appl. Comput. Harmon. Anal.*, **3**, 268-282.

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

Jeffrey Racine

Title - "Shape Constrained Nonparametric Instrumental Regression"

Abstract - We consider estimation and inference of nonparametric nstrumental regression functions in the presence of shape constraints using the approach of Du, Parmeter and Racine (2012).

GAUSSIAN ORACLE INEQUALITIES FOR NON-PARAMETRIC COX'S HAZARD MODEL

JELENA BRADIC AND RUI SONG

Abstract. In this paper we define nonparametric hazard model and propose general class of group penalties suitable for sparse structured variable selection and estimation with specified intergroup structure. We develop novel non-asymptotic sandwich bounds for the partial-log-likelihood and show how they extend notion of local asymptotic normality (LAN) of Le'Cam. Proposed non-asymptotic extension of LAN enables us to obtain new finite sample general and sparse oracle inequalities for penalized estimators of non-parametric partial likelihood when $p \gg n$. Moreover, we are able to extend LAN principles in high dimensional spaces by showing that finite sample prediction properties of penalized estimator in non-parametric Cox's proportional hazards model, under suitable censoring conditions, match those of penalized estimator in linear Gaussian model.

Department of Mathematics, University of California, San Diego
Department of Statistics, Colorado State University
E-mail addresses: jbradic@math.ucsd.edu;song@stat.colostate.edu .

THE HYBRID WILD BOOTSTRAP FOR TIME SERIES

JENS-PETER KREISS AND EFSTATHIOS PAPARODITIS

Abstract. We introduce a new and simple bootstrap procedure for general linear processes, called the hybrid wild bootstrap. The hybrid wild bootstrap generates frequency domain replicates of the periodogram that imitate asymptotically correct the first and second order properties of the ordinary periodogram including its weak dependence structure at different frequencies. As a consequence, the hybrid wild bootstrapped periodogram succeeds in approximating consistently the distribution of statistics that can be expressed as functionals of the periodogram, including the important class of spectral means for which all so far existing frequency domain bootstrap methods generally fail. Moreover, by inverting the hybrid wild bootstrapped discrete Fourier transform, pseudo-observations in the time domain are obtained. The generated time domain pseudo-observations can be used to approximate correctly the random behavior of statistics the distribution of which depends on the first, second and (to some extend) on the fourth order structure of the underlying linear process. Thus, the proposed hybrid wild bootstrap procedure applied to general time series overcomes several of the limitations of standard linear time domain bootstrap methods.

Title - Minimax properties of Fréchet means in deformable models for curve registration and image warping

Author's name - Jérémie Bigot

Affiliation
Institut de Mathématiques de Toulouse
Université de Toulouse et CNRS (UMR 5219)
31062 Toulouse, Cedex 9, France

Abstract - In this talk, we propose to study the problem of estimating a mean pattern from a set of similar curves or images. In the setting where the variability in the data is due to random deformations in space and additive noise, this problem requires to define non-Euclidean distances by using the action of a Lie group on an infinite dimensional space of curves or images. This approach leads to the construction of estimators based on the notion of Fréchet mean which is a generalization of the standard notion of barycenter to non-Euclidean spaces. A recent research direction in nonparametric statistics is the study of the properties of the Fréchet mean in deformable models, and the development of consistent estimators of a mean shape. Using such models, we propose to show the links that exist between minimax theory in nonparametric statistics, inverse problems and the analysis of high-dimensional data for the purpose of estimating a mean pattern from a sequence of curves or images. We will also highlight the connection between our approach and the well know problems of curve registration and image warping.

PARTICLE-BASED LIKELIHOOD INFERENCE IN PARTIALLY OBSERVED DIFFUSION PROCESSES USING GENERALISED POISSON ESTIMATORS

JIMMY OLSSON

Abstract. We discuss the use of the expectation-maximisation (EM) algorithm for inference in partially observed diffusion processes. In this context, a well known problem is that all except a few diffusion processes lack closed-form expressions of the transition densities. Thus, in order to estimate efficiently the EM intermediate quantity we construct, using novel techniques for unbiased estimation of diffusion transition densities, a random weight fixed-lag auxiliary particle smoother, which avoids the well known problem of particle trajectory degeneracy in the smoothing mode.

Centre for Mathematical Sciences, Lund University, Lund, Sweden

IDENTIFICATION AND SHAPE RESTRICTIONS IN NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

by

Joachim Freyberger

and

Joel L. Horowitz

Department of Economics
Northwestern University
Evanston, IL 60201
U.S.A.

Abstract

This paper is concerned with inference about an unidentified linear functional, $T(g)$, where the function g satisfies the relation $Y = g(X) + U;\ E(U \mid W) = 0$. In this relation, Y is the dependent variable, X is a possibly endogenous explanatory variable, W is an instrument for X, and U is an unobserved random variable. The data are an independent random sample of (Y, X, W). In much applied research, X and W are discrete, and W has fewer points of support than X. Consequently, neither g nor $T(g)$ is nonparametrically identified. Indeed, $T(g)$ can have any value in $(-\infty, \infty)$, so the data are uninformative about $T(g)$. In applied research, this problem is typically overcome and point identification is achieved by assuming that g is a linear function of X. However, the assumption of linearity is arbitrary and not testable. There are infinitely many other specifications of g that are observationally equivalent to linearity but give substantive results that are very different from those obtained with linearity. This paper explores the use of shape restrictions, such as monotonicity or convexity, for achieving interval identification of $T(g)$. Economic theory often provides such shape restrictions. This paper shows that they restrict $T(g)$ to an interval whose upper and lower bounds can be obtained by solving linear programming problems. Inference about the identified interval can be carried out by using by using the bootstrap. An empirical application illustrates the usefulness of shape restrictions for carrying out nonparametric inference about $T(g)$.

Nonseparable triangular models: testing instrument validity

Joeri Smits[1], Jeffrey S. Racine[2]

[1] Department of Economics and Resource Management, Norwegian University of Life Sciences, N-1432 s, Norway. E-mail: smits.joeri@ymail.com.
[2] Department of Economics, McMaster University, Hamilton, Ontario, Canada, L8S 4M4.

We provide a new criterion for instrument validity in the nonseparable triangular model and discuss its testability. Nonparametric conditional independence testing plays a key role. An important area of application is randomized trials with partial compliance.

1. Introduction

In recent years, estimators for nonseparable recursive (triangular) models have been developed relying on (an) instrumental variable(s) for identification. Suppose the data generating process is

$$Y = h_0(T, X, U)$$
$$T = h_1(Z, X, V)$$

where Y is the outcome, Z the instrument (set), T the endogenous variable of interest, which we call the treatment, X a set of covariates and U and V unobservables; h_0 and h_1 and g are unknown general functions. In this model, Z is a valid instrument if and only if $Y^T \perp Z \mid T, X$. This condition is untestabe since it contains potential outcomes. We therefore provide an alternative criterion due to the settable systems extension of the Pearl Causal Model framework due to [1].

2. Results

The new criterion states that Z is a valid instrument with respect to the effect of T on Y if and only if Z and Y are causally isolated given (T, X). Since $Y \perp Z \mid T, X$ implies that Y and Z are not causally isolated given (T, X), the exclusion restriction is statistically refutable. It is not statistically confirmable however, since a failure to reject conditional independence does not imply Z and Y to be causally isolated given (T, X). This problematic case of $Y \perp Z \mid T, X$ with an invalid instrument can be due (a) Z and Y have a hidden common cause and/or (b) Z affecting Y exclusive of (T, X). For certain instruments, possibility (a) can be ruled out a priori (f.i. when randomization instruments treatment status in clinical trials with partial compliance) or deemed highly implausible, of which we provide examples. Possibility (b) can occur only due to numerical coincidences in the observed distributions, when different pathways from Z to Y cancel out. Such numerical coincidences are less likely to occur when testing conditional independence repeatedly on subsamples defined by different values of the covariates. Exploiting the fact that $Y \perp Z \mid T, X \Leftrightarrow f_0(Y|T, X, Z) = f_1(Y|T, X)$, the nonparametric test of equality of conditional densities with mixed categorical and continuous data of [2] can be applied when Z is multinomial and (T, X) contains a mixture of categorical and continuous variables. Monte Carlo simulations show that the test has the correct size and has power that increases monotonically with a mean shift of the conditional outcome and the sample size.

3. References

[1] Chalak, K. and H. White (2012), Causality, Conditional Independence and graphical separation, *Neural Computation*, 24(6): 1-55.
[2] Li, Q., E. Maasoumi and J.S. Racine (2009), A Nonparametric Test for Equality of Distributions with Mixed Categorical and Continuous Data, *Journal of Econometrics*, 148(2): 186-200.

Title: Inglorious Count Copulas

Author: Johanna Neslehova, McGill University

Abstract:

The talk will review various facts about copula models linking discrete distributions. It will be shown that the presence of atoms in the marginal probability distributions invalidates various familiar relations that lie at the root of copula theory in the continuous case. It will be highlighted that undiscriminating transposition of modeling and inference practices from the continuous setting to the discrete one may produce misleading and invalid results. As a promising alternative avenue for inference, the so-called Maltese empirical process will be introduced and its asymptotic distribution studied.

Bridging centrality and extremity: refining empirical data depth using extreme value theory

John H.J. Einmahl
Tilburg University

Jun Li
University of California, Riverside

Regina Y. Liu
Rutgers University

Abstract. A data depth is a measure of centrality of a point with respect to a given distribution or data set. It provides a natural center-outward ordering of multivariate data points and yields novel systematic nonparametric multivariate analyses. In particular, the approaches derived from geometric depths (e.g. halfspace depth and simplicial depth) are especially useful since they generally reflect accurately the true probabilistic geometry underlying the data. However, the empirical geometric depths are defined to be zero outside the convex hull of the data set. This property has restricted much the utility of depth approaches in applications where the extreme outlying probability mass may be the focal point, such as in problems of classification or quality control charts with small false alarm rates.

To overcome this shortcoming, we propose to apply extreme value theory to refine the empirical estimator of half-space depth for points in the tails of the data set. This proposal provides an important linkage between data depth, which is useful for inference on centrality, and extreme value theory, which is useful for inference on extremity. The proposed refined estimator of the half-space depth can thus extend depth utilities beyond its data hull to the entire sample space, and broaden greatly the applicability of data depth. We show that the proposed refined depth is uniformly "ratio-consistent" on a very large region and that it significantly improves upon the original empirical estimator. This can be immediately translated into improvement for inference approaches based on depth. In a detailed simulation study we show the improvement of the estimator and how it leads to better performance of depth applications, especially in the directions of multivariate classification and construction of nonparametric control charts.

On and Off Semiparametric Models

Plenary Talk to be presented at the
International Conference on Nonparametric Statistics
Halkidiki, Greece
15-19 June, 2012

by

Jon A. Wellner
Department of Statistics, University of Washington, Seattle WA visiting Heidelberg

Abstract: The information lower bound theory of semiparametric models gives a fairly systematic treatment concerning efficient estimation when a semiparametric models holds. In some cases, especially problems involving missing data, efficient estimators (i.e. estimators achieving the bounds) are difficult to construct and evaluate, while various ad hoc estimators are available which are much easier to compute and study, but which are inefficient "on the model". What happens "off the model" when the model fails to hold, perhaps just by a small amount? Is it possible for the inefficient estimators to be more efficient on neighborhoods of the semiparametric model than the "efficient estimators (on the model)".

In this talk I will review some of the (old and new) literature on this problem and present several recent results concerning estimation on neighborhoods of semiparametric models. The main concern will be stability of efficiency properties on local neighborhoods of semiparametric models.

A NONPARAMETRIC TEST FOR RISK-RETURN RELATIONSHIPS

Juan Carlos Escanciano*
Indiana University

Juan Carlos Pardo-Fernández
Universidade de Vigo

Ingrid Van Keilegom
Université catholique de Louvain

April 9, 2012

Abstract

This article proposes nonparametric tests for risk-return relationships, i.e. restrictions between the conditional mean and the conditional variance of excess returns given a set of unobservable parametric factors. A distinctive feature of our tests is that they do not require a parametric model for the conditional mean and variance, while allowing for possibly time-varying parametric risk-return relationships. We propose new semiparametric estimators for parameters in the risk-return relationship and use those estimates to construct tests based on the difference between the estimated restricted and unrestricted errors distributions. A suitable transformation of this difference renders the tests asymptotically distribution-free, with limits that are transformation of a standard normal variable. Hence, the tests are straightforward to implement. A simulation study compares the finite sample performance of the proposed tests. Finally, an application to the CRSP value-weighted excess returns highlights the merits of our approach.

*Corresponding address: Indiana University, Department of Economics, 100 S. Woodlawn, Wylie Hall, Bloomington, IN 47405-7104, USA, e-mail: jescanci@indiana.edu.

On empirical Bayes posterior distributions

Judith Rousseau

CEREMADE - Université Paris Dauphine and ENSAE- CREST
rousseau@ceremade.dauphine.fr

Abstract: In this talk I will discuss some asymptotic of empirical Bayes procedure. This is a joint work with S. Petrone and C. Scriocciolo. In this work we have first established sufficient conditions ensuring consistency of empirical Bayes procedures which can be applied to both parametric and non parametric models. One of the consequences of consistency of the empirical posterior distribution is that it merges weakly with any other procedure. We then refine the analysis to understand how the empirical Bayes posterior distribution behaves asymptotically in the special case of marginal likelihood empirical Bayes procaracterization also gives some insight of what happens in some nonparametric models .

Julien Worms

Title : "About some empirical likelihood based confidence regions in extreme values statistics"

Abstract : "This talk will be devoted to the use of the empirical likelihood methodology for computation of confidence intervals/regions for classical parameters in the extreme values statistical field. The main parameter of interest will be the extreme value index γ, but joint estimation with the scale parameter in the Peaks Over Threshold framework will also be considered (in the heavy tail - {\it i.e.} Fr\'echet - case). The calibration problem, {\it i.e.} accuracy of the method in terms of coverage probabilities, will be adressed by simulations and comparison with other methods (some of them relying on asymptotic normality and estimation of the asymptotic variance). "

Sudden Changes in Nonparametric Time Series

Jürgen Franke
Department of Mathematics, University of Kaiserslautern, Germany

Co-authors:
C. Kirch
Department of Mathematics, Karlsruhe Institute of Technology
W.K. Li
Department of Statistics, University of Hongkong
J. P. Stockis, J. Tadjuidje-Kamgaing
Department of Mathematics, University of Kaiserslautern

The data generating mechanism of time series sometimes changes from one state to another, e.g. in finance from a low volatile market to a state with higher risk. We first present tests for detecting such changepoints which are based on nonparametric sieve estimates corresponding to neural networks and to not require a specific parametric structure.

Changepoints are related to regime switching if a time series rarely, but repeatedly switches between a finite number of data generating mechanism. We consider Markov switching between nonparametric time series models and algorithms for fitting them to data as well as filtering algorithms for reconstructing the sequence of hidden states driving the observed process.

K.M. Abadir, W. Distaso, F. Zikes

Design-free estimation of large variance matrices.

This talk introduces a new method for estimating variance matrices. Starting from the orthogonal decomposition of the sample variance matrix, we exploit the fact that orthogonal matrices are never ill-conditioned and therefore focus on improving the estimation of the eigenvalues. We estimate the eigenvectors from just a fraction of the data, then use them to transform the data into approximately orthogonal series that deliver a well-conditioned estimator (by construction), even when there are fewer observations than dimensions. We also show that our estimator has lower error norms than the traditional one. Our estimator is design-free: we make no assumptions on the distribution of the random sample or on any parametric structure the variance matrix may have. Simulations confirm our theoretical results and they also show that our simple estimator does very well in comparison with other existing methods, especially when the data are generated from fat-tailed densities.

General notions of depth for functional data - a projection approach

Professor Dr. Karl Mosler
Seminar fuer Wirtschafts- und Sozialstatistik
Universitaet zu Koeln

Data depth has become an important nonparametric tool of multivariate analysis. It measures the centrality of a point with respect to a (data or probability) distribution in Euclidean space and is closely related to multivariate quantiles, central ranks and outlyingness. The upper level sets of a depth function form central regions that reflect the location, scale and shape of the given distribution.

Recently, several proposals have been made to extend the notion of depth to functional data.

Here, a general approach to constructing data depths in functional spaces is proposed; it is based on depth infima over proper sets of finite dimensional projections. For these data depths a set of desirable properties regarding invariance and monotonicity is established.

The general definition includes known functional data depths and many others as special cases; in particular the new notions of location-scale depth and principal component depth are introduced

Efficient Markov Chain Monte Carlo schemes for time-varying parameters of epidemic dynamical systems

Konstantinos Kalogeropoulos[1], Joseph Dureau[1], Marc Baguelin[2]
[1]Department of Statistics, London School of Economics, London, UK
[2]Centre for the Mathematical Modeling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK

Epidemics are often modelled using non-linear state-space models such as the "Susceptible-Infected-Retired" (SIR) system of ODEs, and extensions thereof. Inference is typically based on various observation regimes that include partial and noisy observations. As a consequence, the likelihood function is generally intractable, posing a challenging estimation and computational problem. In our work we consider stochastic extensions to the popular SEIR model with parameters evolving in time, to capture unknown influences of changing behaviours, public interventions, seasonal effects etc. Our models impose little restriction to the trajectories of the time-varying parameters, using Brownian motion or integrals thereof.

Our inferential procedure is based on Markov Chain Monte Carlo methods (MCMC), and in particular on the particle MCMC algorithm, suitably adjusted to accommodate the features of this challenging non-linear stochastic model. We also implement and compare various recently developed inference methodologies such as the Maximum likelihood via Iterated Filtering algorithm, as well as alternative MCMC schemes. We discuss the pros and cons of the different algorithms and illustrate their performance on simulated data. Moreover, we elaborate on the robustness and flexibility of the adopted model and discuss the benefits of its application in real time, using data from the 2009 A/H1N1 pandemic in England.

Keywords: Infectious disease modelling; Bayesian Inference; Markov Chain Monte Carlo; Sequential Monte Carlo, Latent diffusion processes

Multivariate stochastic volatility modelling by sequential Monte Carlo methods

Kostas Triantafyllopoulos
University of Sheffield, UK

This talk considers a modelling framework for multivariate volatility in financial time series. The talk will briefly review the advantages of setting up a model, which inference is achieved by employing sequential Monte Carlo (SMC) methods. We will briefly review the long literature on multivariate volatility modelling and will then proceed to the model definition. The model of the returns consists of a multivariate skew-t distribution, while the dynamics of the volatility are driven by a Wishart autoregressive process. The skew-t distribution is a suitable distribution for financial returns being a more pragmatic model than Gaussian and t distributions. For inference, we adopt SMC methods, for which we discuss some of the challenges we face, within the context of high dimensional time series. We consider low and high dimensional data, consisting of share prices of the stock market, for which we provide minimum portfolio analyses.

Semi-parametric efficiency bounds for the Maximum Partial Likelihood Estimator in Nested-Case Control Sampling

Larry Goldstein, Department of Mathematics, University of Southern California, Los Angeles, CA

Sampling is inevitable in epidemiological studies involving large cohorts. Designs such as case-cohort sampling and nested-case control sampling were originally developed in semi-parametric settings where individuals with a common base line hazard are followed over time in order to detect relationships between disease and exposure.
Typically some version of the Cox model is assumed, and relative risk parameters are estimated by the maximum partial likelihood estimator, or MPLE. As the partial likelihood is not a true likelihood, questions arise about the absolute efficiency of such designs, even when their performance is favorable when measured by their asymptotic relative efficiency against situations where full cohort data is available.

In particular, the MPLE is known not to be efficient for the nested case control sampling design when covariates are fixed over time, and various estimators have been proposed that gain some advantage over the MPLE under a variety of conditions. For purposes of robustness against overmodeling, it is of interest to understand under which situations, if any, the MPLE is efficient under the nested case control sampling design.
By use of the convolution theorem in the framework of semi parametric models we show that the MPLE is asymptotically efficient for ‘highly stratified’ cohorts, also corresponding to a situation where covariates are highly variable in time.

Joint work with Haimeng Zhang

Sequential Prediction of Stationary Time Series

László Györfi
Budapest University of Technology and Economics

Abstract. We present sequential procedures for the prediction of a real valued time series with side information. For squared loss, the prediction algorithms are based on a machine learning combination of several simple nonparametric predictors. We show that if the sequence is a realization of a stationary and ergodic random process then the average of squared errors converges, almost surely, to that of the optimum, given by the Bayes predictor. We offer an analog result for the prediction of stationary gaussian processes, and show an open problem. These prediction strategies have some consequences for 0-1 loss (pattern recognition problem for time series).

References

[1] K. Bleakley, G. Biau, L. Györfi and G. Ottucsák. Nonparametric sequential prediction of time series, *Journal of Nonparametric Statistics*, 22, pp. 297–317, 2010.

[2] L. Györfi, L. and G. Ottucsák. Sequential prediction of unbounded time series, *IEEE Trans. Inform. Theory*, Vol. 53, pp. 1866–1872, 2007.

[3] L. Györfi and Gy. Ottucsák Nonparametric sequential prediction of stationary time series, in *Machine Learning for Financial Engineering*, eds. L. Györfi, G. Ottucsák, H. Walk, pp. 177-230, Imperial College Press, 2012.

Selecting informative BAGIDIS coefficients in nonparametric functional regression

L. Delsol[a] and C. Timmermans[b]
[a]MAPMO, Université d'Orléans, and [b]ISBA, Université Catholique de Louvain

May 24, 2012

The curse of dimensionality is a well known issue in nonparametric multivariate statistics. And several dimension reduction regression models (e.g. additive or single index) have been introduced to find a compromise between relevant convergence rates and model flexibility. Nowadays, statistical studies often involve functional data, usually corresponding to curves (but more generally images or surfaces), which may come from the observation of the evolution of a phenomenon over time, spectrometric studies or sound records Parametric regression models involving functional variables have been widely considered (see for instance [5], and [1]) and for many years the idea of nonparametric functional statistics has been considered as irrealistic.

A first attempt in that direction was to reduce the dimension of the functional data using only its first components and come back to well-known multivariate methods. Few years later, a more general approach has been introduced by [3]. Their main idea was to give more flexibility in the way curves are compared through the use of a suitable semi-metric (selected among a given set of potentialy relevant ones). Since this seminal paper, a lot of work has been done to construct kernel estimators adapted to functional data (see for instance [4], [2]). Semi-metrics are usually based on derivatives, PCA, PLS, ... A new family of semi-metrics based on an adaptative wavelet decomposition, called BAGIDIS, has been recently introduced in [6]. The aim of the present work is to explain how a cross validation procedure may be used to select an optimal BAGIDIS semi-metric (selecting informative patters of the explanatory curve) in regression on functional data.

References

[1] Bosq, D. (2000) *Linear Processes in Function Spaces : Theory and Applications*, Lecture Notes in Statistics, **149**, Springer-Verlag, New York.

[2] F. Ferraty et Y. Romain, (2011), Preface, *The Oxford Handbook of Functional Data Analysis.* [Ed. F. Ferraty et Y. Romain], Oxford Press.

[3] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés, *Compte Rendus de l'Académie des Sciences Paris*, **330**, 403-406.

[4] Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis : Theory and Practice.*, Springer Series in Statistics, Springer-Verlag, New York.

[5] Ramsay, J. and Dalzell, C. (1991) Some tools for functional data analysis, *J. R. Statist. Soc. B.*, **53**, 539-572.

[6] Timmermans, C. and von Sachs, R. (2010). BAGIDIS, a new method for statistical analysis of differences between curves with sharp discontinuities. *ISBA Discussion Paper*, **1030**, 28p.

Lian Heng

Title: Simultaneous variable selection and constant coefficient identification in varying-coefficient models

Abstract: We consider the problem of simultaneous variable selection and constant coefficient identification in high-dimensional varying coefficient models. Both objectives can be considered as some type of model selection problems and we show that they can be achieved by a double shrinkage strategy. Under suitable conditions, we show that consistency in terms of both variable selection and constant coefficient identification can be achieved, as well as the oracle property of the constant coefficients.

On a nonparametric resampling scheme for Markov random fields

Lionel Truquet

UMR 6625 CNRS Institut de Recherche Mathématique de Rennes (IRMAR) Université de Rennes I, Campus de Beaulieu, F-35042 Rennes Cedex, France, and
CREST - ENSAI. Timbre J380. Campus de Ker Lann. rue Blaise Pascal. 35170 BRUZ.
Email : lionel.truquet@univ-rennes1.fr

A nonparametric resampling scheme for stationary time series using kernel regression has been introduced by Paparoditis & Politis (2002). The goal of this procedure is to generate bootstrap replicates of a sample using the Markov property of the underlying stochastic process. This approach has been generalized by Bickel & Levina (2006) for some Markov random fields satisfying suitable conditional independence properties ; applications to the texture synthesis problem was the main motivation. In this talk, we study this resampling method for a general Gibbs-Markov random field using a Gibbs sampler. In particular we investigate conditions under which the distribution of the boostrap replicate has consistency properties. Several simulation examples will be presented.

Références

[1] Bickel, P., Levina, E. (2006) *Texture synthesis and nonparametric resampling of random fields.* The Annals of Statistics, Vol. 34, 4, 1751-1773.

[2] Paparoditis, E., Politis, D. (2002) *The local bootstrap for Markov processes.* J. Statist. Plann. Inference, 108, 301-328.

[3] Truquet, L. (2011) *On a nonparametric resampling scheme for Markov random fields.* Electronic Journal of Statistics, Vol. 5, No. 0, 1503-1536 (2011)

V. Dalla, L. Giraitis, H. Koul

Automatic studentization in nonparametric regression.

This paper presents a general result for studentizing a weighted sum of a linear process where weights are arrays of real numbers. This result is then used to propose a data based studentization method for automatic construction of pointwise confidence bands for the time varying mean function in a heteroscedastic non-parametric regression model with non-random uniform design and dependent errors. Estimation of the finite sample variance (standard error) does not involve the estimation of the parameters controlling the dependence form of the errors, and is easy to apply. The method is robust against unknown type of dependence in the errors, including short range, long range and negative dependence. A finite sample Monte-Carlo simulation study assesses the superiority of the proposed methodology, especially in the short memory case.

Nonparametric Permutation-Based Control Charts for Ordinal Data

Livio Corain[1], Luigi Salmaso[1]
[1]: Department of Management and Engineering, University of Padova, Italy
E-mail: livio.corain@unipd.it; luigi.salmaso@unipd.it

Abstract

In the literature of statistical process control (SPC), design and implementation of traditional Shewart-based control charts requires the assumption that the process response distribution follows a parametric form (e.g., normal). However, since in practice, ordinal observations may not follow the pre-specified parametric distribution these charts may not be reliable, as demonstrated for example by Qiu and Li (2011). To develop appropriate control charts that do not require specifying the parametric form of the response distribution beforehand, a number of distribution-free or nonparametric control charts have been proposed in literature (Chakraborti et al., 2001, gives a thorough overview on existing research).

In this connection, this work aims at providing a contribution to the nonparametric SPC literature, proposing univariate and multivariate nonparametric permutation-based control charts for ordinal response variables which are not only interesting as methodological solution but they have a very practical value particularly within the context of monitoring some measure of user's satisfaction, loyalty, etc. related to use of a given service.

In order to validate our proposal we performed a comparative simulation study where NPC chart for ordered categorical response variables has been compared with the most effective parametric and nonparametric counterparts proposed by literature. As confirmed by the simulation study and by the application to a real case study in the field of monitoring of customer satisfaction in services, we can state that the proposed NPC chart for ordered categorical response variables is certainly a good alternative with respect to the literature counterparts.

Keywords: PERMUTATION TESTS, NONPARAMETRIC COMBINATION, NPC CHART.

References

CHAKRABORTI, S., VAN DER LAAN, P., and BAKIR, S.T. (2001): Nonparametric control charts: an overview and some results. Journal of Quality Technology, 33, 304 315.

QIU, P. and LI, Z. (2011): On Nonparametric Statistical Process Control of Univariate Processes. Technometrics, 53(4), 390-405.

Poisson approximation to the number of near-in-time records

by

López Blázquez, Fernando and Salamanca-Miño, Begoña

lopez@us.es; bsm@us.es

Universidad de Sevilla, Spain

Abstract

Records are outstanding observations whose values are larger than all the previous ones. In many systems, large inputs may cause severe damages. These damages can be catastrophic if the largest inputs (records) occur close in time. For a given threshold d, we say that a record is catastrophic if it occurs within an interval of time of d units from the previous record.

In this talk, we study the distribution of the number of d-catastrophic records in a sample of size n, $N_{n,d}$. We show that this distribution is related to the number of cycles of length $\leq d$ in random permutations. We investigate its basic properties and propose a Poisson approximation for this statistics.

Luigi Salmaso

Title: Some Advances in Combination-based multivariate permutation tests

Abstract: In recent years permutation testing methods have increased both in number of applications and in solving complex multivariate problems.
Permutation tests are essentially of an exact nonparametric nature in a conditional context where conditioning is on the pooled observed data set which is generally a set of sufficient statistics in the null hypothesis.
There are many complex multivariate problems which are difficult to solve outside the conditional framework and in particular outside the nonparametric combination (NPC) of dependent permutation tests method. We discuss this method along with some application in different experimental or observational situations (e.g. in biostatistics).
For a given number of subjects, when the number of variables diverges and the noncentrality parameter of the combined test diverges, then the power of permutation combination-based tests converges to one.

Key words: finte-sample consistency, missing data, directional alternatives, mixed data.

Optimal Sample Fraction Selection in Reduced-Bias EVI-Estimation

M. Ivette Gomes
CEAUL, University of Lisbon, and DEIO, Faculty of Science of Lisbon, Portugal
Email: ivette.gomes@fc.l.pt
Co-authors: F. Figueiredo and M. Manuela Neves

Abstract

In the field of *statistics of extremes*, resampling methodologies, like the *jackknife* and the *bootstrap*, have recently revealed to be of high relevance in the adequate estimation of any parameter of extreme events, like a *high quantile*, the *expected shortfall*, the *return period* of a high level or the primary parameter in the field, the *extreme value index* (EVI). We shall revisit the role of the *bootstrap* methodology (Efron, 1979) in the obtention of reliable semi-parametric adaptive EVI-estimates. We shall consider underlying parents, F, with a heavy right-tail, $\overline{F}(t) := 1 - F(t) = t^{-1/\gamma} L(t)$, with $L(\cdot)$ a slowly varying function (Bingham *et al.*, 1987) and $\gamma > 0$ denoting a positive EVI. For these heavy-tailed parents, given the sample $\underline{\mathbf{X}}_n = (X_1, X_2, \ldots, X_n)$ and the associated sample of ascending order statistics (o.s.'s), $(X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n})$, all semi-parametric EVI-estimators, like the classical Hill estimator (Hill, 1975), are functionals of a usually small number of top o.s.'s in the sample, i.e. $\hat{\gamma} = \hat{\gamma}(k) = \hat{\gamma}(X_{n:n}, X_{n-1:n}, \ldots, X_{n-k:n})$. These estimators, $\hat{\gamma}(k)$, are consistent only if the level $X_{n-k:n}$ is an *intermediate* o.s., i.e., if $k \equiv k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$. The use of the *bootstrap* methodology has revealed to be particularly promising in the estimation of such a *tuning* parameter k. When we ask how to choose k, in the estimation of γ, done through $\hat{\gamma}(k)$, we usually want to estimate the value $k_{0|\hat{\gamma}} := \arg\min_k \mathrm{MSE}(\hat{\gamma}(k))$. In order to estimate $k_{0|\hat{\gamma}}$ one can then use a *double-bootstrap* method applied to an adequate *auxiliary statistic* which tends to *zero* and has an asymptotic behaviour similar to the one of $\hat{\gamma}(k)$. Among others, see Gomes and Oliveira (2001), for a classical adaptive EVI-estimation, and Gomes *et al.* (2012), for an adaptive reduced-bias EVI-estimation. But at such optimal levels, we have a non-null asymptotic bias. Next, if we still want to remove such a bias, we can then make use of the *generalized jackknife* (GJ) methodology (Gray and Schucany, 1972). Indeed, the GJ methodology has played a big role in the reduction of bias of semi-parametric estimators of any parameter of extreme events. It is then enough to consider an adequate pair of estimators of the parameter of extreme events under consideration, and to built a reduced-bias affine combination of them (see Gomes *et al.*, 2000, 2011, also among others). In order to illustrate the use of the bootstrap methodology, we shall consider the *minimum-variance reduced-bias* (MVRB) EVI-estimators in Caeiro *et al.* (2005), as well as two GJ EVI-estimators, studied in Gomes *et al.* (2011). In the lines of Gomes *et al.* (2012), we further provide information on the use of the bootstrap methodology for the choice of optimal sample fractions for the different EVI-estimators under consideration, making an appeal for the need of a calibrated bootstrap. Results associated with a small-scale Monte-Carlo simulation of the finite-sample properties of the adaptive estimators as well as applications to data in the fields of insurance, finance and environment, as well as to simulated data, will be provided.

References

[1] Bingham, N., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*. Cambridge Univ. Press, Cambridge.

[2] Caeiro, F., Gomes, M.I. and Pestana, D.D. (2005). Direct reduction of bias of the classical Hill estimator. *Revstat* **3**:2, 111-136.

[3] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **7**: 1, 1-26.

[4] Gomes, M.I. and Oliveira, O. (2001). The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction. *Extremes* **4**:4, 331-358, 2002.

[5] Gomes, M.I., Martins, M.J. and Neves, M.M. (2000). Alternatives to a semi-parametric estimator of parameters of rare events – the Jackknife methodology. *Extremes* **3**:3, 207-229.

[6] Gomes, M.I., Martins, M.J. and Neves, M.M. (2011). *Revisiting the Role of the Generalized Jackknife Methodology in the Field of Extremes*. Notas e Comunicações CEAUL 20/2011.

[7] Gomes, M.I., Figueiredo, F. and Neves, M.M. (2012). Adaptive estimation of heavy right tails: the bootstrap methodology in action. *Extremes*, DOI: 10.1007/s10687-011-0146-6

[8] Gray, H.L., and Schucany, W.R. (1972). *The Generalized Jackknife Statistic*. Marcel Dekker.

[9] Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

Multiscale local polynomials models for estimation and testing

Maarten Jansen
Université Libre de Bruxelles (ULB), Brussels, Belgium

The construction of multiscale analyses (wavelet decompositions) on nonequispaced observations has to reconcile two somehow contradictory objectives. The first objective is smoothness of the output when using the analysis for noise reduction: the decomposition has to take full account of the irregular structure of the observational grid, so that its regularity is not reflected in the reconstruction. The second objective is numerical condition: the decomposition has to be as close as possible to being an orthogonal transform. In order to understand why the two objectives are contradictory, the wavelet transform can be decomposed into basic steps, called lifting steps. In these steps, wavelet coefficients are defined as offsets of a subsample of observed values from a prediction based on adjacent observations. The prediction is often an interpolation. Interpolation may suffer from heavy oscillations on nonequispaced grids. From this we can conclude that many of the wavelet transforms on equispaced data do not extend to nonequispaced data without loss of the delicate balance between smoothness and stability.

This talk investigates the possibilities of replacing interpolation by kernel smoothing as basic building block for new wavelet decompositions. We first explain that smoothing poses a new problem w.r.t. the smoothness of the basis functions, even on equispaced grids. This problem is solved by the use of a slightly redundant decomposition, which can be seen as a nonequispaced version of Burt and Adelson's Laplacian pyramid. The redundancy follows by postponing the decimation in a multiscale decomposition by one step. The redundancy factor is 2, which is smaller than the $\log(n)$ factor in a fully nondecimated wavelet transform. A second observation is that, except in a Haar transform, we have to impose that a linear function can be decomposed with all detail (wavelet) coefficients equal to zero. Indeed, without this property, offsets are necessary to represent the function $y = x$, which corresponds to the observational grid. As a consequence, the grid would be reflected in the reconstruction, leading to non-smooth basis functions. The condition implies that simple kernel estimation (local constant smoothing) cannot be used in a successful multiscale decomposition on irregular point sets. We need at least local linear smoothing. As a third subject of this study, we investigate choices of the four main transform parameters. Three of them are related to the scaling functions: the first is choice of the kernel, which has an impact on the smoothness of the scaling function. The second is the degree of the local polynomial, which controls the degree of sparsity. The third is the kernel bandwidth, which does not act as a smoothing parameter here. It defines the scales of the transform at each level, which are note necessarily dyadic. The bandwidth should not be too large, for the sake of sparsity, and not too small either, for reasons of numerical stability.

Current research concentrates on estimation of the derivatives for use in ad-

vanced hypothesis testing.

Wavelet-based clustering for mixed-effects functional models

Madison Giacofci[1], Sophie Lambert-Lacroix[2], Guillemette Marot[3], Franck Picard[3]

[1] Laboratoire LJK, Université de Grenoble et CNRS,
UMR 5224, 38041 Grenoble cedex 9, France
[2] UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG
UMR 5525, Grenoble, F-38041, France
[3] Laboratoire Biométrie et Biologie Evolutive,
UMR CNRS 5558 - Univ. Lyon 1, F-69622, Villeurbanne, France

More and more scientific studies yield to the collection of large amounts of data that consist of sets of curves recorded on individuals. These data can be seen as an extension of longitudinal data in high dimension and are often modeled as functional data. Our purpose is to perform unsupervised clustering of these curves in the presence of inter-individual variability. Curve clustering is a widely studied subject and splines have already been proposed to account for inter-individual variability in this context [2]. However splines are known to be computationally inefficient and they can not be used to model irregular curves such as peak-like data.
We develop a new procedure to perform clustering of functional data in the presence of inter-individual variability. We use a wavelet decomposition of the data for both fixed and random-effects. This ensures that both fixed and random effects lie in the same functional space even when dealing with irregular functions that belong to Besov spaces [1]. In the wavelet basis the model resumes to a linear mixed-effects model that can be used for a model-based clustering algorithm.
Our approach follows two steps. First an efficient dimension reduction step based on wavelet thresholding and multiple testing is performed. This first step is necessary due to the high dimensionality of the data. Our aim is to select the wavelet coefficients that are the most informative with respect to the clustering objective of the procedure. Then a clustering step is applied on the selected coefficients. An EM-algorithm is used for maximum likelihood estimation and to predict jointly label variables and random effects.
The properties of the overall procedure are validated by an extensive simulation study. Then we illustrate our method on high throughput molecular data (omics-data) like microarray CGH or mass spectrometry data. Our procedure is available through the R package `curvclust`.

Key-words : functional data, wavelets, linear mixed model, clustering, EM-algorithm

References

[1] Antoniadis A., Sapatinas T.
Estimation and inference in functional mixed-effects models, *Computational Statistics & Data Analysis, Volume 51, Issue 10, 15 June 2007, Pages 4793-4813*

[2] James, G. and Sugar, C.,
Clustering for sparsely sampled functional data, *Journal of the American Statistical Association, Volume 98, Number 462, June 2003*

The two-sample problem for Poisson processes: adaptive tests with a non-asymptotic wild bootstrap approach

Magalie Fromont (1), Béatrice Laurent (2), Patricia Reynaud-Bouret (3)

(1) CREST Ensai - IRMAR, Rennes (FRANCE)
(2) IMT, INSA Toulouse (FRANCE)
(3) CNRS, Université de Nice Sophia-Antipolis (FRANCE)

Considering two independent Poisson processes, we address the question of testing equality of their respective intensities. We construct multiple testing procedures from the aggregation of single tests whose testing statistics come from model selection, thresholding and/or kernel estimation methods. The corresponding critical values are computed through a non-asymptotic wild bootstrap approach. The obtained tests are proved to be exactly of level at most α, and to satisfy non-asymptotic oracle type inequalities. From these oracle type inequalities, we deduce that our tests are adaptive in the minimax sense over a large variety of classes of alternatives based on classical and weak Besov bodies in the univariate case, but also Sobolev and anisotropic Nikol'skii-Besov balls in the multivariate case. A simulation study furthermore shows that they strongly perform in practice.

Keywords : two-sample problem, Poisson process, bootstrap, adaptive tests, minimax separation rates, kernel methods

Likelihood inference in spatio-temporal modeling of small area rare events

Mahmoud Torabi
University of Manitoba, Canada
torabi@cc.umanitoba.ca

In this talk, we use generalized Poisson mixed models for the analysis of geographical and temporal variability of small area rare events. In this class of models, spatially correlated random effects and temporal components are adopted. Spatio-temporal models that use conditionally autoregressive smoothing across the spatial dimension and B-spline smoothing over the temporal dimension are considered. Our main focus is to make inference for smooth estimates of spatio-temporal small areas. The frequentist analysis of these complex models is computationally difficult. On the other hand, the advent of the Markov chain Monte Carlo algorithm has made the Bayesian analysis of complex models computationally convenient. Recent introduction of the method of data cloning has made frequentist analysis of mixed models also equally computationally convenient. We use data cloning to conduct frequentist analysis of spatio-temporal modeling of small area rare events. The performance of the proposed approach is studied through a simulation study and an application to a real dataset.

Adaptive Nonparametric Instrumental Regression

Jan Johannes Maik Schwarz

Université catholique de Louvain

We consider the problem of estimating the structural function in nonparametric instrumental regression, where a response Y is modeled in dependence of an endogenous explanatory variable Z in the presence of an instrument W.

The proposed estimator is based on dimension reduction and additional thresholding. The minimax optimal rate of convergence of the estimator is derived assuming that the structural function belongs to some ellipsoids which are linked to the conditional expectation operator of Z given W. We illustrate these results by considering classical smoothness assumptions. However, the proposed estimator requires an optimal choice of a dimension parameter depending on certain characteristics of the unknown structural function and the conditional expectation operator of Z given W, which are not known in practice.

The main issue addressed in our work is a fully adaptive choice of this dimension parameter supposing that the conditional expectation operator of Z given W is smoothing in a certain sense. In this situation we develop a data-driven estimator which can attain the lower risk bound up to a constant over a wide range of smoothness classes for the structural function.

A goodness–of–fit test for the functional linear model with scalar response

Eduardo García–Portugués, Wenceslao González–Manteiga and Manuel Febrero–Bande

Abstract

In this work, a goodness–of–fit test for the null hypothesis of a functional linear model with scalar response is proposed. The test is based on a generalization to the functional framework of a previous one (see Escanciano, 2006) designed for the goodness of fit of regression models with multivariate covariates using random projections. A simulation study illustrates the finite sample properties of the test for several types of basis, basis dimensions and under different alternatives. Finally, the test is applied to two datasets to test for a functional linear model.

References

J. C. Escanciano. A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051, 2006.

Of Copulas, Quantiles, Ranks and Spectra an L_1-approach to spectral analysis

Holger Dette[a*], Marc Hallin[b*†], Tobias Kley[a*‡] and Stanislav Volgushev[a*]

[a] Ruhr-Universität Bochum

[b] ECARES, Université Libre de Bruxelles, and ORFE, Princeton University

Abstract

In this paper we present an alternative method for the spectral analysis of a strictly stationary time series $\{Y_t\}_{t\in\mathbb{Z}}$. We define a "new" spectrum as the Fourier transform of the differences between copulas of the pairs (Y_t, Y_{t-k}) and the independence copula. This object is called *copula spectral density kernel* and allows to separate marginal and serial aspects of a time series. We show that it is intrinsically related to the concept of quantile regression. Like in quantile regression, which provides more information about the conditional distribution than the classical location-scale model, the copula spectral density kernel is more informative than the spectral density obtained from the autocovariances. In particular the approach provides a complete description of the distributions of all pairs (Y_t, Y_{t-k}). Moreover, it inherits the robustness properties of classical quantile regression, because it does not require any distributional assumptions such as the existence of finite moments. In order to estimate the copula spectral density kernel we introduce rank-based Laplace periodograms which are calculated as bilinear forms of weighted L_1-projections of the ranks of the observed time series onto a harmonic regression model. We establish the asymptotic distribution of those periodograms, and the consistency of adequately smoothed versions. The finite-sample properties of the new methodology, and its potential for applications are briefly investigated by simulations and a short empirical example.

AMS 1980 subject classification : 62M15, 62G35.

Key words and phrases : Time series, Spectral analysis, Periodogram, Quantile regression, Copulas, Ranks, Time reversibility .

*Supported by the Sonderforschungsbereich "Statistical modelling of nonlinear dynamic processes" (SFB 823) of the Deutsche Forschungsgemeinschaft.

†Académie Royale de Belgique, CentER, Tilburg University, and ECORE. Supported by a Discovery Grant of the Australian Research Council.

‡Supported by a PhD Grant of the Ruhr-Universität Bochum and by the Ruhr-University Research School funded by Germany's Excellence Initiative [DFG GSC 98/1].

On the Range of Validity of the Vector Autoregressive Sieve Bootstrap

Presenter: **Marco Meyer (joint work with Jens-Peter Kreiss)**
TU Braunschweig, Germany

Extending the results of Kreiss, Paparoditis and Politis (2011), the limits of the vector autoregressive (VAR) sieve bootstrap are explored. This procedure is designed to be applied to multivariate stationary stochastic processes that are linear and invertible since these processes possess an autoregressive representation with independent white noise. However, there is a much wider class of non-linear processes with infinite-order AR representations which have uncorrelated innovations instead of independent ones, only. It will be explored what the VAR sieve bootstrap really does in this more general situation. A so-called companion process, which has a slightly different structure than the original VAR representation of the process, will be defined and it will be shown that the VAR sieve bootstrap asymptotically works if and only if the limiting distribution of the statistic of interest for the original process and the one for the companion process are identical. This yields a general check-criterion for the validity of the VAR sieve bootstrap which will be applied to some important statistics like the sample mean, sample autocovariances and sample autocorrelations.

References

[1] Kreiss, J.-P., Paparoditis, E., Politis, D.N.: On the Range of Validity of the Autoregressive Sieve Bootstrap. *The Annals of Statistics*, Vol. 39, No. 4, pp. 2103-2130, 2011.

Local polynomials for nonparametric dimension reduction

Francesco Giordano*& Maria Lucia Parrella†
Department of Economics and Statistics, University of Salerno,
Via Ponte Don Melillo, 84084 Fisciano, Italy.

Abstract

Estimating a high-dimensional regression function is notoriously difficult, due to the *curse of dimensionality*. Some nonparametric estimators, such as the local polynomials, are particularly affected by this problem, since they have slow rates of convergence when the number of covariates is high. However, for some applications a sparse condition can be formulated, which assumes that the true multivariate regression function depends only on a small number of the covariates. In such cases, an estimation procedure which is capable of isolating the relevant variables can be used for dimension reduction. Contextually, satisfactory rates of convergence can be reached.

In this work, we begin describing the different solutions recently proposed in the literature to deal with the *curse of dimensionality* problem in nonparametric regression, among which the RODEO method of Lafferty and Wasserman (2008). Inspired by the last paper, we propose an iterative estimation procedure based on a modified version of the sparseness condition. In particular, we focus on the number of covariates for which the gradient is not constant. These covariates are defined "nonlinear covariates", meaning that the marginal relation between the dependent variable and each of such covariates is nonlinear. We show that the new estimation procedure has a good rate of convergence even when the effective dimension of the regression function (= total number of relevant covariates) is high, provided that the number of "nonlinear covariates" is low. The proposed procedure works along the following three main steps:

a) identifying the "nonlinear covariates", for which we estimate the marginal bias and the optimal bandwidth of the local polynomial estimator; to this aim, we propose new estimators of the marginal bandwidth and the marginal bias which are easily scalable to high dimension;

b) identifying the "linear covariates", for which oversmoothing can be done thanks to the conditional unbiasedness of the local polynomial estimator;

c) removing the "irrelevant covariates" from the set of regressors.

In order to identify the "nonlinear covariates" and the "linear covariates", in steps *a*-*b*, new tests are proposed based on nonparametric inferential methods. This let to extend the applicability of the proposed estimation procedure to a more general nonparametric framework, relaxing some conditions of the RODEO tecnique. Moreover, the proposed estimation procedure can be used not only for dimension reduction, but also to explore the multivariate structure of the regression function. As a consequence, further interesting applications of the new procedure arise: sensitivity analisys, nonparametric multivariate godness-of-fit test; nonparametric multivariate nonlinearity test.

Some of the theoretical results of the proposed estimation method are discussed, and a simulation study is presented to show the empirical performance of the procedure.

Keywords: nonparametric regression; local polynomials; dimension reduction; curse of dimensionality, multivariate bandwidth selection.

*email: giordano@unisa.it
†email: mparrella@unisa.it

CONFESS: Feature Selection for Classification With a Large Number of Classes

Marianna Pensky
Department of Mathematics, University of Central Florida

In what follows, we introduce two Bayesian models for feature selection in high dimensional data, specifically designed for the purpose of classification. When the number of variables is large, model selection is of uttermost importance. Without it, as Fan and Fan (2008) show, classification is as good as pure guessing. We consider the situation when the number of components is much larger than the number of samples, the number of classes is also large and the number of samples per class is small. This set up is motivated by classification of animal communication signals, in particular, by a data set comprising electric signals of 473 individual fishes recorded at a site in the Amazon Basin from 21 geographically co-occurring species of tropical South American electric knife fishes.

Many model selection algorithms have been developed but majority of them are not specifically designed for classification and usually select the features which exhibit high variablity even though this variability can be meaningless for classification purposes. We suggest the new model selection technique CONFESS, CONstant Feature Selection Strategy, which identify the features which exhibit high between class variability in comparison with within class variability and is similar to ANOVA. When one has only two classes, the method reduces to the Feature Annealed Independence Rule (FAIR) introduced by Fan and Fan (2008) and can be viewed as a natural generalization of FAIR to the case of more than two classes.

We study performance of CONFESS and derive conditions under which it allows separation between "meaningful" and "meaningless" features. It turns out that when the number of classes is relatively large, one can select the "meaningful" features with probability tending to one when the number of features grow. The unexpected result is that precision of model selection improves when the number of classes grows and, hence, under certain conditions, enables conclusive classification.

(This is joint work with Justin Davis, University of Central Florida)

Sequential Nonparametric Procedures

Marie Hušková
Charles University in Prague, Faculty of Mathematics and Physics,
Department of Statistics, Prague, Czech Republic

Various sequential procedures for detection of instability, changes, or unexpected fluctuations play an increasingly important role in applications as data sets are often collected automatically or without significant costs. Such situations arise for example in finance (risk management, exchange rates monitoring), in econometrics, in medicine (monitoring intensive care patients, sequential clinical trials), and meteorology (global warming issues, severe weather warning systems). In these cases, it is important to monitor the current situation and to react as soon as the monitored process indicates some instability.

The talk concerns nonparametric sequential procedures for detection of such instabilities in probability distribution in a series of observations. Particularly, procedures based on either ranks, U-statistics, empirical distribution functions, or empirical characteristic functions. Results on their limit behavior both under the null hypothesis as well as alternative ones.

Most of the procedures assume that a training (historical) data set is available at the beginning of the monitoring. Both independent observations as well as extensions to time series are discussed.

Results of simulation study will be presented.

Possibility an extension to more general models will be discussed.

The presented results are mostly based on the joint work with Hlávka, Kirch, Meintanis.

Keywords: Change in distribution; Empirical characteristic function; Empirical distribution function; Ranks; Sequential procedures; U-statistics.
Subject Classifications: 62G20; 62E20; 60F17.

APPLICATIONS OF FUNCTIONAL DATA ANALYSIS ON THERMAL ANALYSIS

Mario Francisco-Fernández*, Javier Tarrío-Saavedra† and Salvador Naya†

*University of A Coruña, Faculty of Computer Science, Campus de Elviña s/n 15071, A Coruña (Spain).

†University of A Coruña, Higher Polytecnic School, Campus de Esteiro s/n 15403, Ferrol (Spain).

ABSTRACT

In this research, some functional data analysis approaches are applied in the field of thermal analysis. Two specific problems are studied in this work, considering in both cases the so-called thermogravimetric (TG) curves as sample data. They explain the mass loss when temperature is increased. In the first problem, a functional ANOVA test for a one way treatment is applied to measure the influence of adding fumed silica on the thermal degradation of an epoxy resin. The procedure is applied to rescaled TG curves and also to their derivatives. Previously to the application of the test, statistical techniques like penalized b-splines or the concept of depth are used to prepare the data sets to be analyzed. All tests performed have resulted highly significant, that is, the addition of fumed silica affects the way of degradation of epoxy resin involved in the sample. Moreover, by applying pairwise comparisons using the functional ANOVA method and a bootstrap distance based test, it can be observed that the way of degradation for each group considered in this research is significantly different from the others. In the second problem, functional and multivariate statistical techniques are used to classify seven commercial wood species from their TG curves. A functional nonparametric classification method based on the Nadaraya-Watson regression estimator, and classical and machine learning multivariate classification procedures are compared using real wood samples, through leave-one-out cross-validation and external validation. To apply the multivariate methods, two curve feature extraction techniques are applied, using PCA and fitting a novel logistic regression model from where parameter vectors used as features are obtained. Additionally, a comprehensive simulation study to validate the performance of the different approaches in a variety of scenarios is presented. Synthetic curves imitating the TG curves are generated. Then, apart from obtaining important conclusions related with wood discrimination through this study, many different classification techniques are compared from a statistical point of view in this framework.

Zero crossings for time series

Mariola Molenda
Faculty of Mathematics and Information Science, Warsaw University of Technology

Kedem [1] has found a direct relationship between the first order autocorrelation and the expected zero crossing rate. The number of zero crossings, like the first order autocorrelation, is a measure of the oscillation in time series. There is a possibility of reconstructing the first order autocorrelation of time series via zero crossings. Explicit formula (*cosine formula*), connecting first order autocorrelation ρ_1 and the expected number of zero crossings $E[D]$, exists, among other things, in the case of time series which is zero mean stationary Gaussian time series. The relationship looks as follows

$$\rho_1 = \cos(\frac{\pi E[D]}{n-1}).$$

The zero crossings method provides an alternative estimator of the first order autocorrelation coefficient that is reducing the data storage requirements and is more robust with respect to outliers when compared to the classical estimator.

Having a simple and effective tool as the number of zero crossings, we can estimate the first order autocorrelation. The reverse is also true. However, computing the zero crossing counts is faster. Using real time series we ilustrate how accurate the cosine formula is. We also answer the question how far precisely we can compute the first order autocorrelation using zero crossings.

References

[1] B.Kedem, *Time Series Analysis by Higher Order Crossings* IEEE Press New York 1993.

[2] J.Leskow, M.Molenda, *Resampling methods for time series level crossings* Communications in Statistics - Theory and Methods (accepted).

Shape-constrained nonparametric density estimation in three high-dimensional convex models

Marios G. Pavlides

Centre for Statistical Science and Operational Research,
Queen's University Belfast, Belfast BT7 1NN,
Northern Ireland, United Kingdom.

Abstract

Shape-constrained nonparametric density estimation in high-dimensional settings has attracted the attention of the statistical community, for a couple of decades now, particularly with regards to establishing asymptotic limit theory of the proposed density estimators. Many such high-dimensional models find an abundance of applicability in prominent fields such as in survival analysis, reliability theory, and econometrics. In this talk, I will emphasize attention on the properties of the "Nonparametric Maximum Likelihood Estimator" (NPMLE) in three convex density models, as well as discuss results in terms of "local asymptotic minimax theory."

I will start my talk by considering the NPMLE in the family of all unimodal and L_p–spherical, Lebesgue densities on $\mathbb{R}^d$, that assumes a Grenander-type form and, thus, converges both pointwise as well as globally, in the L_1–metric, at a rate of $n^{1/3}$ to a known, non-degenerate limit. It should be noted that, when $p = 2$, this family includes the standard multivariate normal distribution.

I will then talk about a large family of Lebesgue densities on $(0, \infty)^d$ which finds applicability in survival analysis and reliability theory. The family of all *block-decreasing densities*, $\mathfrak{F}_{\mathrm{SMU}}(d)$, includes all Lebesgue densities on $(0, \infty)^d$ that are non-increasing in each coordinate, while keeping all other coordinates fixed. Such densities are *isotonic* and the properties of the NPMLE in such families has been studied in several published works. In this talk, I will present a "local asymptotic minimax lower bound," at the optimal rate of $n^{1/(d+2)}$, consistent with the optimal L_1 and Hellinger rates of the same order established in a recent publication. I will then discuss conjectures on rates of convergence of the unique NPMLE in this family which, if true, would render the NPMLE "rate sub-optimal."

I will then present a subclass of $\mathfrak{F}_{\mathrm{BDD}}(d)$, that family of all *multivariate scale mixtures of uniform densities,* $\mathfrak{F}_{\mathrm{SMU}}(d)$, that includes all densities of vectors $(X_1U_1 \ldots, X_dU_d)$, where $(U_1, \ldots, U_d)$ follows the uniform distribution on $[0, 1]^d$ and is independent of the vector $(X_1, \ldots, X_d)$ that has an *arbitrary* distribution function, G, on $(0, \infty)^d$. For this family, we establish an almost-sure unique NPMLE, derive its Fenchel characterizations, establish strong consistency (both locally and globally) as well as derive a "local asymptotic minimax lower bound," of the optimal order of $n^{1/3}$. Conjectures on the pointwise asymptotic limit theory of the NPMLE in this setting will also be discussed.

In closing, I will present a window-kernel density estimator that, under local d–times differentiability of the true density at a fixed point in the interior of its support, has an explicit local Gaussian limit at the exact rate of $n^{1/3}$.

Part of the work in this talk was done in collaboration with Jon A. Wellner.

Asymptotic normal estimation of jump location curves in noisy images

Matthias Eulert

Philipps University of Marburg

There is an intrinsic relation between the estimation of change-points in univariate jump regression analysis and the discovery of edges in statistical image reconstruction: Both these objects can be modelled as discontinuities of the regression function under study (which is the image itself in the latter case). An edge typically is understood as a boundary curve which separates the image into different parts according to their color values or brightness levels. In applications one is often interested in the determination of these boundaries since they contain valuable information about the structure of the whole image and about the outline of objects which are depicted in the image, respectively.

For simplicity we assume that the image to be analyzed consists of two colors (black and white) only and that the edge in the image can be described by a mathematical function ϕ with respect to the canonical two-dimensional coordinate system. This means that all discontinuities in the image are located on the graph of ϕ which is called the jump location curve (JLC). We discuss a modified version of the pointwise estimator for the JLC as proposed by Qiu (1997). In our approach we use a locally dilated variant of Qiu's criterion function which enables us to determine the (Gaussian) limit distribution of the modified estimator at a particular boundary point. The asymptotic variance term depends in a quite intuitive manner on the noise level of the observations, a kernel constant and the slope of the JLC at the point where the estimation is done. By this, it is immediate to develop (pointwise) asymptotic confidence bands for the JLC.

The idea behind this method is to use the difference of certain compactly supported kernel estimators of the regression function to detect boundary points on the JLC. If the mechanism is appropriately defined, the JLC can be supposed to be next to points where this local difference is maximized. The procedure by Qiu (1997) involves in a first step the estimation of a specific rotation angle θ of the supports of the kernel functions which accounts for an unknown non-constant derivative of ϕ. We treat this angle θ as a nuisance parameter and identify the boundary point estimator as a suitably transformed bivariate M-estimator (with respect to the locally dilated criterion function) which itself has a non-standard rate of convergence because of the localization. The proofs are inspired by methods used in Müller (1992) for the derivation of the asymptotic distribution of a nonparametric change-point estimator and by general results on M-estimators as presented in van der Vaart and Wellner (1996).

References

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, **20** 737–761. http://projecteuclid.org/euclid.aos/1176348654

Qiu, P. (1997). Nonparametric estimation of the jump regression surface. *Sankhyā Ser. A*, **59** 268–294. http://sankhya.isical.ac.in/search/59a2/59a2020.html

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics, Springer: New York.

Robust Change-point Estimation with Dependent Random Errors

Matúš Maciak - maciak@karlin.mff.cuni.cz
Department of Probability and Mathematical Statistics, Charles University in Prague

Abstract:

In our work we will focus on three different aspects of modern statistical modelling approaches: specifically, we will address flexibility of the model (with respect to assumptions as well as with respect to an unknown regression function which is of interest), robustness (with respect to outliers and heavy-tailed distributions of random errors) and finally, we will also mention possible discontinuity issues in the model.

We will present a nonparametric regression M-smoothers estimation with possible change-points (jumps or structural-breaks respectively) under the assumption that the random error terms are generated by some α-mixing random process.
The main point of our interest is to statistically decide if some given point from the domain of interest is statistically significant for a jump to occur or not (we assume a possible jump in the regression function itself and moreover, in its derivatives up to a specific order as well).

Given some technical limits (unknown quantities) and computational difficulties (plug-in techniques) in obtaining the limit distribution under the null hypothesis we will apply a moving block bootstrap (MBB) algorithm to mimic the unknown distribution of interest.

Mathematical proofs for the consistency and asymptotic normality of M-smoothers estimates themselves as well as a consistency proof for the MBB procedure are briefly presented.

Keywords:

Local polynomial M-smoothers, robustness, discontinuity in nonparametric regression, moving block-bootstrap, α-mixing dependence.

Submitted for the First Conference of the International Society for NonParametric Statistics (ISNPS'2012)
June 15-19, 2012, Chalkidiki, Greece

Semiparametric Estimation with Generated Covariates

Enno Mammen, Christoph Rothe, and Melanie Schienle*

University of Mannheim, Toulouse School of Economics, and Humboldt University Berlin

Abstract

In this paper, we study a general class of semiparametric optimization estimators of a vector-valued parameter. The criterion function depends on two types of infinite-dimensional nuisance parameters: a conditional expectation function that has been estimated nonparametrically using generated covariates, and another estimated function that is used to compute the generated covariates in the first place. We study the asymptotic properties of estimators in this class, which is a non-standard problem due to the presence of generated covariates. We give conditions under which estimators are root-n consistent and asymptotically normal, derive a general formula for the asymptotic variance, and show how to establish validity of the bootstrap.

JEL Classification: C14, C31

Keywords: *Semiparametric estimation, generated covariates, profiling, propensity score*

*Enno Mammen, Department of Economics, University of Mannheim, D-68131 Mannheim, Germany. E-mail: emammen@rumms.uni-mannheim.de. Christoph Rothe, Toulouse School of Economics, 21 Allée de Brienne, F-31000 Toulouse, France. E-mail: rothe@cict.fr. Melanie Schienle, School of Business and Economics, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany. E-mail: melanie.schienle@wiwi.hu-berlin.de.

Kernel Density Outlier Detector

Abstract

Based on the widely known kernel density estimator of the probability function, a new algorithm is proposed in order to detect outliers. With the help of the Gaussian Transform , a weighted kernel density estimation of the density probability function is calculated, referring to the whole of the data including the outliers. In the next step , the data points having the smallest values are removed as the least probable to belong to the pdf of the clear data.
The program based on this algorithm is more accurate even on greatly correlated outliers with the clear data , and even with ouliers with small Euclidean distance from clear data. It is also faster than the methods presented and also provides with the same accuracy without restriction on the dimension of the data , as seen for example in MCD.

Model Checking and Test-based Variable Selection

Michael Akritas

The Pennsylvania State University

May 15, 2012

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 Nadaraya-Watson type kernel estimator and $d \leq 4$, is established under the null hypothesis and local alternatives. Simulations suggest that under a sparse model, the applicability of the test extends to arbitrary d through sufficient dimension reduction. Using p-values from this test, a variable selection method based on multiple testing ideas is proposed. The proposed test outperforms existing procedures, while additional simulations reveal that the proposed variable selection method performs competitively against well established procedures. A real data set is analyzed.

Estimating Long-term Multivariate Progression from Short-term Data

Michael C. Donohue
Department of Family and Preventive Medicine
Division of Biostatistics and Bioinformatics
University of California, San Diego
9500 Gilman Dr., La Jolla, CA, 92093-0949.

May 9, 2012

Abstract

Motivation: Diseases that progress over long periods of time are often studied by observing cohorts at different stages of disease for short periods of time. The Alzheimer's Disease Neuroimaging Initiative (ADNI) follows cohorts of elders with various degrees of cognitive impairment, from normal to impaired. The study includes a rich panel of novel cognitive tests, biomarkers, and brain images collected every six months for up to six years. The relative timing of the observations with respect to disease pathology is unknown. We apply an Alternating Conditional Expectation algorithm to estimate pathologic timing and long-term growth curves.

Results: Simulations demonstrate that the method can recover long-term disease trends from short-term observations. The method also recovers temporal ordering of individuals with respect to disease pathology, providing subject-specific prognostic estimates of the time until onset of symptoms. When applied to ADNI data, we see the estimated growth curves support prevailing theories of the Alzheimer's disease cascade. Other datasets with common outcome measures could be combined using the proposed algorithm.

Availability: Software to fit the proposed model and reproduce results with R is available as the *grace* package (https://bitbucket.org/mdonohue/grace). ADNI data can be downloaded from the Laboratory of NeuroImaging (http://www.loni.ucla.edu).

Presenter: Michael H. Neumann (Friedrich Schiller University Jena, Germany)

Title: Asymptotics and bootstrap consistency for degenerate von Mises-statistics of ergodic processes

Abstract:
Many important test statistics can be rewritten as or approximated by degenerate von Mises- (V-) statistics. In the first part, I review new results on the asymptotic behavior of degenerate V- and related U-statistics under ergodicity. To set critical values for tests, bootstrap methods are an important tool whenever the distribution of a given test statistic cannot be determined analytically. I present consistency results for bootstrap approximations under easily verifiable conditions. The talk is based on joint work with Anne Leucht (University of Hamburg).

Nonparametric Regression For Locally Stationary Time Series

Michael Vogt

Department of Economics, University of Cambridge, UK
(e-mail: mv346@cam.ac.uk)

Abstract:

We study nonparametric models allowing for locally stationary regressors and a regression function that changes smoothly over time. These models are a natural extension of time series models with time-varying coefficients. We introduce a kernel-based method to estimate the time-varying regression function and provide asymptotic theory for our estimates. Moreover, we show that the main conditions of the theory are satisfied for a large class of nonlinear autoregressive processes with a time-varying regression function. Finally, we examine structured models where the regression function splits up into time-varying additive components. As will be seen, estimation in these models does not suffer from the curse of dimensionality. We complement the technical analysis by an application to financial data.

Keywords:

Local stationarity, nonparametric regression, smooth backfitting.

References:

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* **25** 1-37.

Fryzlewicz, P. & Subba Rao, S. (2010). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli* **17** 320-346.

Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726-748.

Subba Rao, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Advances in Applied Probability* **38** 1155-1172.

Mammen, E., Linton, O. & Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27** 1443-1490.

Michael Wolf

Title: Nonlinear shrinkage estimation for large-dimensional covariance matrices.

Abstract:
Many statistical applications require an estimate of a covariance matrix and/or its inverse. When the matrix dimension is large compared to the sample size, which happens frequently, the sample covariance matrix is known to perform poorly and may suffer from ill-conditioning. There already exists an extensive literature concerning improved estimators in such situations. In the absence of further knowledge about the structure of the true covariance matrix, the most successful approach so far, arguably, has been shrinkage estimation. Shrinking the sample covariance matrix to a multiple of the identity, by taking a weighted average of the two, turns out to be equivalent to linearly shrinking the sample eigenvalues to their grand mean, while retaining the sample eigenvectors. Our paper extends this approach by considering nonlinear transformations of the sample eigenvalues. We show how to construct an estimator that is asymptotically equivalent to an oracle estimator suggested in previous work. As demonstrated in extensive Monte Carlo simulations, the resulting bona fide estimator can result in sizeable improvements over the sample covariance matrix and also over linear shrinkage.

Nonparametric hypothesis testing for means of compositional vectors

Michail Tsagris, Simon P. Preston, and Andrew T.A. Wood
School of Mathematical Sciences, University of Nottingham, UK

Abstract

Compositional data are a special type of multivariate data in which the elements of each observation vector are non-negative and sum to one. Data of this type arise in many areas, such as geology, archaeology and biology amongst others. Analysis of such data may be carried out in various ways, e.g. by transforming the data using logs of ratios formed by the components (Aitchison, 2003) or by neglecting the compositional constraint and treating the observations as standard multivariate data. Recently, Tsagris, Preston and Wood (2011) developed a more general approach which includes both approaches indicated above as particular cases. It involves a Box-Cox type transformation, known as the α-transformation, with a free parameter, α.

Our focus will be on nonparametric hypothesis tests for means based on two samples of compositional data. We shall consider two types of nonparametric likelihood: Owen's (2001) empirical likelihood, a nonparametric likelihood which shares some of the desirable properties of parametric likelihood such as Bartlett correctability; and exponential empirical likelihood, also known as nonparametric tilted likelihood, due to Efron (1981), which is similar to empirical likelihood in some respects but does not admit a Bartlett correction. These nonparametric likelihoods will be compared with bootstrapped versions of Hotelling-type and James-type statistics (the latter is similar to the former but allows for the possibility that population covariance matrices are different in different populations). In this and other contexts, both nonparametric likelihoods tend to lead to tests whose type I error is substantially higher than the nominal level. For this reason bootstrap calibration of the two nonparametric likelihoods is also considered.

The performance of the various test statistics was investigated through simulation and through asymptotic analysis. We simulated compositional data from different types of population including the Dirichlet distribution, mixtures of Dirichlet distributions, and the logistic normal distribution. The size of each of the tests was estimated along with the actual time required. Under all circumstances the α-transformation was applied to the data beforehand at a fixed value of the parameter. The conclusion was that bootstrap calibration is necessary in all cases to achieve the correct test size. Furthermore, the Hotelling and James statistics tended in practice to be much faster to compute than the bootstrap-calibrated versions of the nonparametric likelihoods but at the same time achieved similar levels of accuracy.

Keywords: empirical likelihood, James statistic, Hotelling statistic, bootstrap.

References

Aitchison J. (2003). *The Statistical Analysis of Compositional Data.* The Blackburn Press: New Jersey.

Efron B.(1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics 9(2): 139–172.*

Owen A.B. (2001). *Empirical Likelihood.* Chapman & Hall/CRC Press: New York.

Tsagris M., Preston S.P. and Wood A.T.A. (2011). A data-based power transformation for compositional data. *4th International Workshop on Compositional Data Analysis, Spain.*

Michailidis, George
The University of Michigan
Abstract Title: Adaptive sampling procedures for estimating function thresholds

In this talk, we discuss multi (two) - stage sampling procedures for estimating a threshold for a regression function, such as a point where the function crosses some critical value. It will be shown how the proposed procedures, which involve sampling a pre-fixed budget (number) of points (covariate-response pairs) at two stages -the first, a learning stage using an agnostic sampling design, and the second a "zoom-in" stage with sampling in the vicinity of the initial estimate obtained from the learning stage - lead to accelerated convergence rates over onestage procedures, in certain cases even allowing the parametric square-root-n rate to be achieved or exceeded for a nonparametric problem. The proposed procedure is illustrated on synthetic and real data.

Input variable selection for interpretable neural network models

Michele La Rocca and Cira Perna *
Department of Economics and Statistics, University of Salerno (Italy)

Artificial neural networks are widely accepted as useful tools for modeling non linear structures. They are especially helpful when the underlying data generating process is not fully understood or when the nature of the relationships being modeled may display complex nonparametric structure. A neural network can be considered as a parallel distributed model made up of simple data processing units. This parallel structure gives reason to its well known approximation capability: given a sufficiently large number of non linear terms and a suitable choice of the parameters, they are able to approximate arbitrary functions of variables arbitrarily well. However, even if several model selection strategy have been proposed in the literature, the selection of a proper topology for neural networks still appear an open issue.

In this paper a two-step procedure for model selection in neural network modeling is presented and discussed. The basic idea of the novel approach is that hidden neurons and input neurons play a different role in neural network modeling and they should be selected by using different criteria. Particularly, the hidden layer size models the nonlinearity degree of the functional relationship and it acts as a smoothing parameter, taking into account the trade-off between bias and variability. As a consequence, the number of hidden neurons are selected in order to maximize the predictive accuracy of the network, avoiding overfitting, On the contrary, input neurons are related to the explanatory variables, and so they have a very clear interpretation in terms of relevance of a given set of variables to the model. As a consequence, the number and the type of input neurons are selected by means of a test procedure, based on appropriate measures of relevance of a given input variable, to avoid omission of relevant variables or inclusion of redundant ones. In this latter step, in order to bypass the data snooping problem, which can be even be more dangerous in a neural network framework due to the lack of theory supporting the model selection strategy, familywise error rate is controlled by using a multiple testing scheme. Clearly, analytical derivation of the sampling distribution of the test statistic involved is difficult but it can be approximated by using resampling techniques such as the subsampling, which is able to deliver consistent results under very weak assumptions.

*[larocca,perna]@unisa.it

Modified Hill estimation for heavy tails

Milan Stehlík,
Department of Applied Statistics and Econometrics,Johannes Kepler University Linz,
e-mail: Milan.Stehlik@jku.at

Abstract

If the model underlying the data, is a regularly varying function with index -1/alpha it is usually supposed that the top scaled order statistics are Pareto distributed. 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. 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, 2010 or Vandewalle et al. 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)). In Fabián and Stehlík (2010) we understand "The Hill estimator" as a specific procedure for studying of the tail of Pareto distribution. Instead of implementing "The Hill estimator" procedure, we implement the score moment procedure. For the case of Pareto distribution, the Hill estimator procedure with the score moment estimator has been investigated in Stehlík et al. (2012) for optimal testing for normality against Pareto tail. Since the score moment estimator is simple, it is easy to implement it to the Hill procedure. In literature, mainly the asymptotical properties of tail estimators are studied. However, in many situations, asymptotics is simplifying the underlying process too much (see e.g. Huisman, R. et al. (2001)). We may illustrate this fact by a severe bias of the Hill based estimators or by a distributional insensitivity of asymptotical estimators. In many practical situations, such as operational risk assessment, data are sparse (with sample size often bellow 50) and therefore robust estimator with good small sample properties is needed. We will show how to construct a "naturally" robust, distribution sensitive heavy tail estimator and prove its weak consistency together with its good small sample properties. In special cases, the t-score method matches with the other methods, e.g. those base by harmonic moments in Henry III (2009). We will illustrate its applications in robust testing for normality against heavy tails. Many open questions remains: these will be also introduced.

References

Beran J. and Schell D. (2010). On robust tail index estimation, *Computational Statistics and Data Analysis,* doi:10.1016/j.csda.2010.05.028

Fabián Z. (2001). Induced cores and their use in robust parametric estimation, *Communication in Statistics, Theory Methods,* 30 (2001), pp. 537-556.

Fabián Z. and Stehlík M. (2010). On robust and distribution sensitive Hill like method, Technical report

Henry III, J.B. (2009). A harmonic moment tail index estimator, *Journal of Statistical Theory and Applications,* Vol.8, No.2, pp. 141-162.

Hill, B. (1975), A Simple General Approach to Inference About the Tail of a Distribution, *The Annals of Statistics*, 3, 1163-1173.

Huisman, R. et al. (2001). Tail-Index Estimates in Small Samples, *Journal of Business & Economic Statistics,* American Statistical Association, vol. 19(2), pages 208-16.

Stehlík M., Potocký R., Waldl H. and Fabián Z. (2010) On the favourable estimation of fitting heavy tailed data, *Computational Statistics,* 25:485-503.

Stehlík M., Fabián Z. and Střelec L. (2012), "Small sample robust testing for Normality against Pareto tails", *Communications in Statistics - Simulation and Computation,* 41:7, 1167-1194

Vandewalle B., Beirlant J., Christmann A., Hubert M. (2007). A robust estimator for the tail index of Pareto-type distributions, *Computational Statistics & Data Analysis*, Volume 51, Issue 12: 6252-6268

Min-ge Xie

Talk Title: A NONPARAMETRIC META-ANALYSIS FRAMEWORK AND CONFIDENCE INTERVALS FOR ORDER PARAMETERS

Talk Abstract

Fixed-effects and random-effects models are the two most widely used model settings in meta-analysis. In practice, however, it is difficult and almost impossible to verify the fundamental assumptions of these two models --- i.e., the underlying unknown study parameters are the same in a fixed-effects model or they are samples from a single (often normal) distribution in a random-effects model. In this talk, we propose an alternative nonparametric meta-analysis framework without taking a leap of faith to simply adapt these conventional assumptions. The development leads to an inference problem of constructing confidence intervals or hypothesis tests for order parameters as well as the extrema of parameters, for example of $\max\{\theta 1,...,\theta K\}$. Such an inference problem is considered in the literature as one of the existing problems where standard bootstrap estimators are not consistent and where alternative approaches also face significant challenges. Based on recent developments on confidence distributions, we propose a new resampling method to deal with the inference problem for the extrema of the parameters and also, more generally, for any order parameters. This new resampling method can be viewed as an extension of the well-studied and widely-used bootstrap method, but it enjoys a more flexible interpretation and manipulation. We provide a large sample theoretical support for the proposed method. We also explore the theoretical performance of both the standard bootstrap and proposed method, especially in the presence of ties and near ties among the θi s. Empirical performance of the proposed method is studied in numerical examples using both simulations and a real data set. (Joint work with B. Claggett, L.J. Wei and L. Tian)

Nonparametic Inference for Image Symmetries

Mirosław Pawlak
Department of Electrical & Computer Eng., University of Manitoba, Canada
pawlak@ee.umanitoba.ca

AMS subject classifications:Primary 62G05; Secondary 62G10

Key words and phrases: symmetry detection, symmetry estimation, noisy images, radial polynomials, limit distributions, degree of symmetry

Abstract

Symmetry plays an important role in image understanding and recognition. This paper formulates the problem of assessing reflection and rotations symmetries of an image function observed in the presence of noise. Rigorous nonparametric statistical tests are developed for testing image invariance under reflections or under rotations through rational angles, as well as joint invariance under both reflections and rotations. The symmetry relations are expressed as restrictions 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 image symmetry, e.g., estimating the axis of reflectional symmetry and the degree of rotational symmetry. It is shown that the obtained estimates converge at the parametric rate for all images of bounded variation. The asymptotic normality for image functions being Lipschitz continuous is also established. Further, we introduce an index of the degree of symmetry that takes value one for fully symmetric images and is smaller than one otherwise. The index possesses the desired invariance properties with respect to image scaling and shifting. An estimate of this symmetry functional with the $\sqrt{n}$ convergence is proposed.

Title: Regression Estimation of the Index of Regular Variation

Field: Specification tests for non-parametric and semi-parametric models and related methods

Authors: Mofei Jia Email: murphy.410@hotmail.com

Emanuele Taufer Email: emanuele.taufer@unitn.it

Affiliation: University of Trento, Italy

Abstract:

Exploiting the relationship between the characteristic function at the origin and the distribution function of the tail, we propose a regression estimator of the index of regular variation. We expect this estimator performs better given that it utilises the entire sample, and not just a few extreme order statistics. We assume independent observations and a partially known functional form of a regularly varying tail. We derive expressions for the estimator's asymptotic bias and variance. Consistency of the estimator is obtained in the domain of regular variation of index $0 < \alpha < 2$. We also assess small sample properties by means of simulations. This case is easily to be extended to the case $\alpha > 0$ after transforming and also to be used in dependent cases. Some tools employed in multifractal models can be considered as well.

On some aspects of order estimation and related results for univariate autoregressive processes.

Moritz Jirak
Austrian Financial Market Authority

For decades, model selection and order estimation has been an important issue in statistics. In case of an univariate autoregressive process $\left\{X_k\right\}_{k\in\mathbb{Z}}$, various estimators for the order q and the parameters $\mathbf{\Theta}_q = (\theta_1, ..., \theta_q)^T$ are known; the order is usually determined with Akaike's criterion or related modifications, whereas Yule–Walker, Burg or maximum likelihood estimators are used for the parameters Θ_q. In this talk, we present alternatives which are based on simultaneous confidence bands for the Yule–Walker estimators $\widehat{\theta}_i$, more precisely, we consider the quantities $\sqrt{n}\max_{1\leq i\leq d_n}\big|\widehat{\theta}_i - \theta_i\big|$ and $\max_{1\leq k\leq d_n}(2k)^{-1/2}\big|n\left(\widehat{\mathbf{\Theta}}_k - \mathbf{\Theta}_k\right)^T\widehat{\Gamma}_k^{-1}\left(\widehat{\mathbf{\Theta}}_k - \mathbf{\Theta}_k\right) - k\big|$ as a starting point. It is shown that - appropriately normalized - the corresponding limit distribution is the Gumbel-type distribution $e^{-e^{-z}}$, where $q \in \{0, ..., d_n\}$ and $d_n = \mathcal{O}\left(n^\delta\right)$, $\delta > 0$. On one hand, this allows to modify some of the currently used criteria (AIC, BIC, HQC, SIC), but also yields a new class of consistent estimators for the order q. These estimators seem to have some potential if the underlying process exhibits a sparsity in parameters. In addition, these results may be used to derive new consistency results for BIC, HQC and SIC if one allows an increase in the dimension d_n of the parameter space with $d_n = \mathcal{O}\left(n^\delta\right)$, $\delta > 0$.

Distributions of Clusters of Exceedances and Their Applications in Peer-to-Peer Overlay Networks

Natalia Markovich
Institute of Control Sciences Russian Academy of Sciences, Moscow, 117997 Russia

Keywords: cluster, extremes in time series, exceedance over threshold, extremal index, delay in clusters, lossless period

In many applications it is important to evaluate the impact of clusters of observations caused by the dependence and heaviness of tails of time series. We consider a stationary sequence of random variables (rvs) $\{R_n\}_{n\geq 1}$ with marginal cumulative distribution function $F(x)$ and the extremal index $\theta \in [0,1]$. The clusters contain consecutive exceedances of the time series over a threshold u separated by return intervals with consecutive non-exceedances.

We derive geometric forms of asymptotically equal distributions of the normalized cluster size and an inter-cluster size that depend on θ. The inter-cluster size determines the number $T_1(u)$ of inter-arrival times between events of interest arising between two consecutive clusters, i.e. between two consecutive exceedances of the process $\{R_n\}$ over u. The cluster size is equal to the number $T_2(u)$ of inter-arrival times within clusters, i.e. between events arising between two consecutive non-exceedances. The inferences are valid when u is taken as a sufficiently high quantile of the process $\{R_n\}$. The derived geometric models allow us to obtain the asymptotically equal means of $T_1(u)$ and $T_2(u)$.

It is shown that the limit tail distributions of the return intervals and the duration of clusters that are defined as sums of a random number of weakly dependent regularly varying inter-arrival times with tail index $0 < \alpha < 2$ are bounded by sums of stable and exponentially distributed components. The numbers of terms in these sums are determined by $T_1(u)$ and $T_2(u)$, respectively.

The main problem in the teletraffic theory concerns the transmission of information with minimal loss and delay during delivery. We consider the packet traffic in peer-to-peer (P2P) applications like Skype and IPTV where the packet lengths and inter-arrival times between packets are both random. The peers in a P2P overlay network may randomly join or leave the structure. Leaving peers cause the loss of all stored data at those peers and hence, downgrade the quality. We study the loss and delay at the packet layer. The main idea is that the packet loss is caused by exceedances of the rate of a transmission over a threshold u that can be interpreted as a channel capacity. Then the packets can be lost only in the clusters generated by exceedances of the rate above u. In this context, the rates are considered as underlying process $\{R_n\}$, the return intervals as lossless periods and cluster durations as delay between successfully delivered packets. Since active streams may share the capacity of a channel, we propose an algorithm for the management of the equivalent capacity (i.e., the part of capacity allocated for each stream) in such a way to minimize the missing probability of a packet and to meet the delay deadline. Using geometric distributions of $T_1(u)$ and $T_2(u)$ and Wald's equation we propose to estimate the means of the lossless time and the delay in the clusters in an on-line regime. The latter can be used to optimize the quality of packet transmissions in P2P overlay networks.

On maximal correlation coefficient and some counterexamples

N. Papadatos

Department of Mathematics
Section of Statistics and O.R.
University of Athens,
Panepistemiopolis, 157 84 Athens
Greece
e-mail:npapadat@math.uoa.gr

Abstract

We propose a simple method that enables the convenient calculation of the maximal correlation coefficient for bivariate distributions having diagonal structure. Applying this method, a large number of elementary counterexamples are constructed, showing that linear regression of each variable on the other does not imply that the maximal correlation equals to the absolute value of the correlation. We also extend a result of Castaño-Martínez, López-Blázquez and Salamanca-Miño (Maximal correlation between order statistics. In: *Recent Developments in Ordered Random Variables*, M. Ahsanullah and M. Raqab (eds.), Nova Science Publishers, 2007, pp. 55–68) related to maximal correlations of partial minima (or maxima) from an i.i.d. sequence. Our model considers partial minima based on two different branches of a splitting sequence. It is shown that the maximal correlation is not attained but, as in the ordinary case of partial minima, it is approximated by suitable power distributions. This result is based on a Lemma given by Yu (On the maximal correlation coefficient, *Statist. Probab. Lett.*, vol. 78, pp. 1072–1075).

Multilevel Spatially Correlated Binary Longitudinal Data

Nicoleta Serban[1]
H. Milton Stewart School of Industrial Systems and Engineering
Georgia Institute of Technology
nserban@isye.gatech.edu

Ana-Maria Staicu
Department of Statistics
North Carolina State University
ana-maria_staicu@ncsu.edu

Raymond J. Carroll
Department of Statistics
Texas A&M University
carroll@stat.tamu.edu

In this article, we provide insights into and introduce new methodology for the analysis of multilevel binary data observed longitudinally when the repeated longitudinal measurements are spatially correlated. The proposed model is logistic functional regression conditioned on three latent processes describing the within- and between-variability, and describing the spatial dependence of the repeated longitudinal measurements. We estimate the model components without employing mixed-effects modeling but assuming an approximation to the logistic link function. The primary objectives of this paper are to highlight the challenges in the estimation of the model components, to compare two approximations, linear and exponential, and to discuss their advantages and limitations. The linear approximation is computationally efficient whereas the exponential approximation applies for rare events functional data. Our methods are inspired by and applied to data obtained from a state-of-the-art colon carcinogenesis scientific experiment. However, our models are general and will be relevant to many new binary functional data sets, with or without cross-dependence between functions.

Key Words: Binary longitudinal data; Colon carcinogenesis; Covariogram estimation; Functional data analysis; Hierarchical modeling; Mixed models; Multilevel functional data; Principal component estimation; Spatial modeling.

Short title: Multilevel Binary Longitudinal Data

[1]Correspondent Author

Bayesian function analysis via locally constant priors: A unified approach

Various statistical models involve a certain function, say f, like the mean regression as a function of a covariate, the hazard rate as a function of time, the spectral density of a time series as a function of frequency, or an intensity as a function of geographical position, etc. Such functions are often modelled parametrically, whether for frequentist or Bayesian uses, and under weak conditions there are so-called Bernstein-von Mises theorems implying that these two approaches are large-sample equivalent. Results of this nature do not necessarily hold up in nonparametric and high-dimensional setups, however.

The aim of the present work is to exhibit a unified framework and methodology for both frequentist and Bayesian nonparametric analysis, involving priors that set f constant over windows, and where the number m of such windows grows with sample size n. We work out conditions on the number and sizes of the windows under which Bernstein-von Mises type theorems can be established, with the prior changing with sample size via the growing number of windows. Illustrations of the general methodology are given, including a setup for inference about frequency spectra for stationary time series.

Copulas and covariates

Noël Veraverbeke
Universiteit Hasselt, Belgium
noel.veraverbeke@uhasselt.be

Studying the relationship between two (or more) random variables in the presence of a covariate can be done based on a conditional version of Sklar's theorem: there exists a copula function expressing the joint conditional distribution as a function of the one dimensional conditional marginal distributions. We discuss recent results on several estimators for this unknown copula function. First of all there is the nonparametric method which uses empirical estimators with weights that smooth over the covariate space. An application is the asymptotic theory for association measures like the conditional Kendall's tau [2,4]. A second method is semi-parametric in nature: it starts from a parametric family of copulas in which the parameter depends on the covariate. This parameter function is estimated by local likelihood [1]. A third method provides a smooth estimator by the use of Bernstein polynomials [3].

References

[1] F. Abegaz, I. Gijbels and N. Veraverbeke, Semi-parametric estimation of conditional copulas. *J. Multivariate Analysis* (to appear) 2012.

[2] I. Gijbels, N. Veraverbeke and M. Omelka, Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis*, **55**, 1919-1932, 2011.

[3] P. Janssen, J. Swanepoel and N. Veraverbeke, Large sample behavior of the Bernstein copula estimator. *J. Statist. Planning and Inference*, **142**, 1189-1197, 2012.

[4] N. Veraverbeke, M. Omelka and I. Gijbels, Estimation of a conditional copula and association measures. *Scandinavian J. Statistics*, **38**, 766-780, 2011.

Goodness of Fit Tests for Probabilistic Index Models

Olivier Thas[1,2] and Jan De Neve[1]

[1] Department of Mathematical Modelling, Statistics and Bioinformatics,
Ghent University, Belgium
[2] School of Mathematics and Applied Statistics, University of Wollongong, Australia

Probabilistic Index Models (Thas et al., 2012) form a class of semiparametric models for the probabilistic index. In particular, for a couple of independent continuous response variables Y and Y^*, associated with covariate vectors X and X^*, respectively, the model imposes the restriction

$$\mathrm{P}\{Y < Y^* | X, X^*\} = m(X, X^*; \beta),$$

with m a function with range $[0, 1]$ and β the parameter vector. An important subclass resembles generalised linear models in the sense that $m(X, X^*; \beta) = g^{-1}(Z^t\beta)$ with g a proper link function and Z a vector with elements taken from X and X^* (e.g. $Z = X^* - X$).

We propose a test and a graphical tool for assessing the model adequacy (De Neve et al.). As the model shows similarities with logistic regression, we first suggest to assess the goodness of fit using a method which is based on the rationale of the test of Hosmer and Lemeshow (1980). Our second test makes more explicitly use of the structure of the probabilistic index models by recognising that X and X^* are two instances of the same covariate. Simulation results indicate that both methods succeed well in detecting lack-of-fit. The methods are also illustrated on a case study.

References

De Neve, J., Thas, O. and Ottoy, J.P. Goodness-of-fit methods for probabilistic index models. *Communications in Statistics – Theory and Methods*. To appear.
Hosmer, D. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model *Communications in Statistics - Theory and Methods*, **9**, 1043-1069.
Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012). Probabilistic Index Models (with discussion). *Journal of the Royal Statistical Society - Series B*, **74**(4), 1-29 (to appear).

Ori Davidov

Title: The linear stochastic order and directed inference for multivariate ordered distributions.

Abstract: Researchers are often interested in drawing inferences regarding the order between two experimental groups on the basis of multivariate response data. Since standard multivariate methods are designed for two sided alternatives they may not be ideal for testing for order between two groups. In this article we introduce the notion of the linear stochastic order and investigate its properties. Statistical theory and methodology are developed to both estimate the direction which best separates two arbitrary ordered distributions and to test for order between the two groups. The new methodology generalizes Roy's classical largest root test to the nonparametric setting and is applicable to random vectors with discrete and/or continuous components.

The proposed methodology is illustrated using data obtained from a a 90-day pre-chronic rodent cancer bioassay study conducted by the National Toxicology Program (NTP). Not only is the proposed methodology more sensitive in detecting a dose-related increase in response, but is also simple to use.

Title: The second-order frequency identification for cyclostationary time series using subsampling
Author: Oskar Knapik
Affiliation: Cracow University of Economics

Abstract
Recently, there is a growing interest in research in signal processing was dedicated to frequency identification and analysis of cyclostationarity. There are many important applications of cyclostationary signals in telecommunications, vibroacoustics, finance and economy. The aim of this poster is to show applicability of subsampling technique in second-order frequency identification for such signals. The theoretical results are accompanied with applications to frequency analysis of cyclostationary signals from vibroacoustics, neuroscience and economy. The results presented on this poster provide a new perspective on cyclostationary signal analysis and on frequency identification for such signals.

Selected references:

1. Cioch W., Knapik O., Leśkow J. (2012) „Finding a frequency signature for a cyclostationary signal with applications to wheel bearing diagnostics", (to be appeared in Mechanical Systems and Signal Processing)
2. Lenart Ł. , Leśkow, J. Synowiecki R. (2008), „Subsampling in estimation of autocovariance of PC time series", Journal of Time Series Analysis, Vol.29 , No. 6, pg. 995-1018.
3. Leśkow J. (2012) „Cyclostationarity and Resampling for Vibroacoustic Signals", Acta Physica Polonica A, Vol. 121, pg. 160-163

Non-Parametric Estimation of the Residue Function in a Dynamic PET Imaging Study.

Positron emission tomography (PET) provides the ability to non-invasively measure the time-course of radio-labeled molecules and their metabolites, in-vivo. This is a powerful tool with a range of pre-clinical and clinical applications - particularly in cancer medicine. Because the PET reconstructed time-course is a function of vascular delivery and tissue retention, an analysis is required to separate these and extract better information of the PET study. The time-course is a convolution of the tissue residue (a life-table) and tracer activity in the arterial supply. Although there are a range of parametric blood-tissue exchange model constructs available, non-parametric evaluation of the tissue residue offers both computational and theoretical advantages. Significantly, key kinetic variables including tracer flow, volume of distribution and flux are general quantities that do not require a model form for their definition or analysis. A non-parametric analysis of these variables is feasible. In the case that the arterial input function is not directly sampled - typically standard for clinical and small animal studies now - analysis of time-course data leads to a problem of blind deconvolution. Although there are theoretical challenges here the use of physiological constraints makes the problem tractable. Some current developments in this area will be described. The work is illustrated by examples taken from on-going PET imaging studies of cancer and its response to therapy.

[Supported in part by Science Foundation Ireland (MI-2007) and the National Cancer Institute (CA-65537)]

Sequential adaptive estimators in nonparametric autoregressive models

Ouerdia ARKOUN

Laboratoire de Mathématiques Raphaël Salem, UMR 6085 CNRS, Université de Rouen, Avenue de l'Université, BP.12, 76801 Saint Etienne du Rouvray (France).
email: ouerdia.arkoun@gmail.com

Abstract

We construct a sequential adaptive procedure for estimating the autoregressive function at a given point in nonparametric autoregression models with Gaussian noise. We make use of the sequential kernel estimators. The optimal adaptive convergence rate is given as well as the upper bound for the minimax risk.

Key words: Adaptive estimation, kernel estimator, minimax, nonparametric autoregression.

1 Introduction

Our problem is the following. Suppose we observe data from the model :

$$y_k = S(x_k)y_{k-1} + \xi_k\,, \quad 1 \le k \le n\,, \tag{1.1}$$

where $x_k = k/n$, y_0 is a constant and $(\xi_k)_{k\in\{1,\dots,n\}}$ are independent and standard Gaussian random variables.

In this paper, similarly to Galtchouk and Pergamenshchikov (2001), we apply the Lepskiĭ procedure to the model (1.1) based on the sequential kernel estimates. We construct the sequential kernel estimator using the method proposed in Borisov and Konev (1977) for the parametric case. It should be noted that to apply the Lepskiĭ procedure the kernel estimators must to have the tail distribution of the Gaussian type. To obtain this property one needs to use the sequential approach. To this end we show some modification of the Levy theorem for discrete time and then, using this result, we show that the sequential kernel estimators have the same form for the tail distribution as a Gaussian random variable. It should be noted that the non sequential kernel estimator does not have the above property in the case of the model (1.1). Thus, in this case, the adaptive pointwise estimation is possible only in the sequential framework.

Let us describe now the sequential kernel estimators. For a constant $H > 0$, we define α_H, such that

$$\sum_{j=1}^{\tau_H - 1} Q(u_j)\, y_{j-1}^2 + \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H - 1}^2 = H \quad \text{with} \quad u_j = \frac{x_j - z_0}{h}\,,$$

where the kernel $Q(\cdot)$ is the indicator function of the interval $[-1; 1]$, and τ_H is the stopping time defined as follows:

$$\tau_H = \inf\{1 \le k \le n : \sum_{j=1}^{k} Q(u_j)\, y_{j-1}^2 \ge H\}, \tag{1.2}$$

and $\tau_H = n$ when this set is empty.

Note that

$$A_k = \sum_{j=1}^{k} Q(u_j) y_{j-1}^2\,,$$

where h is a positive parameter that we will be define in the next section.

Thus the sequential kernel estimator is written as follows:

$$S_h^*(z_0) = \frac{1}{H}\left(\sum_{j=1}^{\tau_H - 1} Q(u_j)\, y_{j-1}\, y_j \;+\; \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H - 1}\, y_{\tau_H}\right) \mathbf{1}_{(A_n \ge H)}, \tag{1.3}$$

with $H = nh$. Note that, on the set $\{A_n \ge H\}$ the coefficient $0 \le \alpha_H \le 1$.

Such an estimator is very convenient to estimate the quantity $\mathbf{E}\,|S_h^*(z_0) - S(z_0)|$.

We describe in detail the statement of the problem in section 2 and in Section 3, we illustrate the obtained results by numerical examples.

2 Statement of the problem

The problem is to estimate the function S at a fixed point $z_0 \in]0, 1[$, i.e. the value $S(z_0)$. For any estimate $\tilde{S}_n = \tilde{S}_n(z_0)$, the risk is defined on the neighborhood $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ by

$$\mathcal{R}_n(\tilde{S}_n) = \sup_{\beta \in [\beta_*; \beta^*]} \sup_{S \in \mathcal{H}^{(\beta)}(z_0, K, \varepsilon)} N(\beta)\, \mathbf{E}_S |\tilde{S}_n(z_0) - S(z_0)|\,, \tag{2.1}$$

where $N(\beta) = \left(\dfrac{n}{\ln n}\right)^{\beta/(2\beta+1)}$ corresponds to the convergence rate of adaptive estimators on class $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ and $\mathbf{E}_S$ is the expectation taken with respect to the distribution $\mathbf{P}_S$ of the vector $(y_1, ..., y_n)$ in (1.1) corresponding to the function S.

We consider model (1.1) where $S \in \mathbf{C}_1([0,1],\mathbb{R})$ is the unknown function. To obtain the stable (uniformly with respect to the function S) model (1.1), we assume that for some fixed $0 < \varepsilon < 1$, the unknown function S belongs to the *stability set*

$$\Gamma_{\varepsilon} = \{S \in \mathbf{C}_1([0,1],\mathbb{R}) : \|S\| \leq 1-\varepsilon\},$$

where $\|S\| = \sup_{0<x\leq 1} |S(x)|$. Here $\mathbf{C}_1[0,1]$ is the Banach space of continuously differentiable $[0,1] \to \mathbb{R}$ functions. For fixed constants $K > 0$ and $0 < \beta \leq 1$, we define the corresponding *stable local Hölder class* at the point z_0 as

$$\mathcal{H}^{(\beta)}(z_0, K, \varepsilon) = \{S \in \Gamma_{\varepsilon} \, : \, \Omega^*(z_0, S) \leq K\}\,,$$

with

$$\Omega^*(z_0, S) = \sup_{x\in[0,1]} \frac{|S(x) - S(z_0)|}{|x - z_0|^{\beta}}\,.$$

The regularity $\beta \in [\beta_*; \beta^*]$, is supposed to be unknown, where the interval $[\beta_*; \beta^*]$ is known, $\beta_* > 0$ and $\beta^* \leq 1$.

First we give the lower bound for the minimax risk. We show that with the convergence rate $N(\beta)$ the lower bound for the minimax risk is strictly positive.

Theorem 2.1. *The risk* (2.1) *admits the following lower bound:*

$$\liminf_{n\to\infty} \inf_{\tilde{S}_n} \mathcal{R}_n(\tilde{S}_n) \geq \frac{1}{4}\,,$$

where the infimum is taken over all estimators $\tilde{S}_n$.

Now we give the upper bound for the minimax risk of the sequential kernel estimator defined in (1.3). Since β is unknown, one can not use this estimator because the bandwidth h depends on β. That is why we partition the interval $[\beta_*; \beta^*]$ to follow a procedure of Lepskiĭ. Let us set

$$d_n = n/\ln n \quad \text{and} \quad h(\beta) = \left(\frac{1}{d_n}\right)^{\frac{1}{2\beta+1}}\,.$$

We define the grid on the interval $[\beta_*; \beta^*]$ with the points :

$$\beta_k = \beta_* + \frac{k}{m}(\beta^* - \beta_*), \quad k = 0, \ldots, m \quad \text{with} \quad m = [\ln d_n] + 1\,.$$

We denote $N_k,\ h_k$ and $\omega(h_j)$ as

$$N_k = N(\beta_k),\ h_k = h(\beta_k)\,,$$

and

$$\omega(h_j) = \max_{0\le k\le j}\left(|S^*_{h_j} - S^*_{h_k}| - \frac{\lambda}{N_{k+1}}\right).$$

We also define the optimal index of the bandwidth as

$$\widehat{k} = \inf\left\{0 \le j \le m\ :\ \omega(h_j) \ge \frac{\lambda}{N_j}\right\} - 1\,. \tag{2.2}$$

We note that $\omega(h_0) = -\lambda/N_1$ and thus $\widehat{k} \ge 0$. The positive parameter, λ, is chosen as $\lambda > K + e\sqrt{4 + \dfrac{4}{2\beta_* + 1}}$.
The adaptive estimator is now defined as

$$\widehat{S}_n = S^*_{\widehat{h}} \quad \text{with} \quad \widehat{h} = h_{\widehat{k}}\,. \tag{2.3}$$

The following result gives the upper bound for the minimax risk of the sequential adaptive estimator defined above.

Theorem 2.2. *For all* $0 < \varepsilon < 1$, *we have*

$$\limsup_{n\to\infty} \mathcal{R}_n(\hat{S}_n) < \infty\,.$$

Remark 2.3. *Theorem 2.1 gives the lower bound for the adaptive risk, i.e. the convergence rate* $N(\beta)$ *is best for the adapted risk. Moreover, by Theorem 2.2 the adaptive estimates* (2.3) *possesses this convergence rate. In this case, this estimates is called optimal in sense of the adaptive risk* (2.1)

Lemma 2.4. *For all* $z \ge 2$ *and* $H > 0$, *one has*

$$\mathbf{P}_S(|\zeta_H(h)| > z) \le 2\, e^{-z^2/8}\,,$$

where

$$\zeta_H = \frac{1}{\sqrt{H}}\left(\sum_{j=1}^{\tau_H - 1} Q(u_j)\, y_{j-1}\, \xi_j + \alpha_H\, Q(u_{\tau_H})\, y_{\tau_H - 1}\, \xi_{\tau_H}\right) \mathbf{1}_{(A_n \ge H)}.$$

3 Numerical simulations

We illustrate the obtained results by the following simulation which is established using Scilab.

The purpose is to estimate, at a given point z_0, the function S defined over $[0;1]$ by $S(x) = |x - z_0|^\beta$. We check that such a function belongs to $\mathcal{H}^{(\beta)}(z_0, K, \varepsilon)$ when $K \geq 1$. The values of z_0 and β are arbitrary, which permit the user to name his choice. As an example, take $z_0 = 1/\sqrt{2}$. Then $\beta_* = 0.6$ is a lower regularity value and $\beta^* = 0.8$ is the higher regularity value.

We simulated n data for the function $S(x) = |x - z_0|^\beta$ for $\beta = 0.7$. We obtained an estimation in constructing the estimator $\hat{S}_n$ defined in (2.3) with the procedure of Lepskiĭ which gives us the optimal bandwidth for the index $\hat{k}$ defined in (2.2).

Numerical results approximate the asymptotic risk of a sequential estimator defined in (2.3) used due to the calculation of an expectation (it performs an average for $M = 15000$ simulations) and the finite number of observations n. Here we calculate for the sequential estimator the quantity $\mathbf{R}_n = \dfrac{1}{M} \displaystyle\sum_{k=1}^{M} |\hat{S}_n^{(k)}(z_0) - S(z_0)|$.

By varying the number of observations n, we obtain different risks listed in the following table:

n	100	1000	5000	10000
$\mathbf{R}_n$	0.284	0.154	0.101	0.087

When taking $\beta = \beta^* = 1$, we obtain

n	100	1000	5000	10000
$\mathbf{R}_n$	0.201	0.097	0.058	0.047

As one can see, the sequential adaptive estimator $\hat{S}_n$ is converging to its true value $S(z_0) = 0$, but this convergence is slow. This is expected since the optimal adaptive convergence rate is $N(\beta) = \left(\frac{n}{\ln n}\right)^{\beta/(2\beta+1)}$. For $\beta = 1$, the results are slightly better than those we got in the first table.

References

[1] Arkoun, O. and Pergamenchtchikov, S. (2008) : Nonparametric Estimation for an Autoregressive Model. Vestnik of Tomsk State University, Ser. *Mathematics and Mechanics* **2** (3), 20 - 30.

[2] Belitser, E. (2000a) : Local minimax pointwise estimation of a multivariate density, *Statisti. Nederletica* **54** (3), 351-365.

[3] Borisov, V.Z. and Konev, V.V. (1977) : Sequential Estimation of Parameters of Discrete Processes, *Automat. and Remote control* **10**, 58-64.

[4] Dahlhaus, R. (1996a) : On the Kullback-Leibler information divergence of locally stationary processes, *Stochastic Process. Appl.* **62** (1), 139–168.

[5] Fourdrinier, D., Konev, V.V. and Pergamenchtchikov, S. (2009) : Truncated Sequential Estimation of the Parameter of a First Order Autoregressive Process with Dependent Noises, *Mathematical Methods of Statistics* **18** (1), 43-58.

[6] Galtchouk, L. and Pergamenshchikov, S. (2001) : Sequential nonparametric adaptive estimation of the drift coefficient in diffusion processes, *Math. Methods Statist.* **10** (3), 316–330.

[7] Helland, I. S. (1981) : Central limit theorems for martingales with discrete or continuous time. *Scet. J. Statist.* **9** (2), 79–94.

[8] Lepskiĭ, O. V. (1990) : A problem of adaptive estimation in Gaussian white noise, *Theory Probab. Appl.* **35** (3), 454-466.

[9] Tsybakov, A. B. (1998) : Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes, *Ann. Statist.* **26** (6), 2420–2469.

Level set estimation

P. Saavedra Nieves, W. González Manteiga and A. Rodríguez Casal*

*Universidad de Santiago de Compostela

Level set estimation theory deals with the nonparametric problem of reconstructing an unknown set of type $\{f \geqslant f_\tau\}$ from a random sample of points $\mathcal{X}_n = \{X_1, ..., X_n\}$, where f stands for the density of $\mathcal{X}_n$, $\tau \in (0,1)$ is a probability fixed by the practitioner and $f_\tau > 0$ denotes the threshold such that the level set $\{f \geqslant f_\tau\}$ has a probability at least $1-\tau$ with respect to the distribution induced by f. This problem has been considered in three different ways in the literature.

The plug-in methods are based on replacing the unknown density f by a suitable nonparametric estimator f_n, usually the kernel density estimator. So, this group of methods proposes $\{f_n \geqslant \hat{f}_\tau\}$ as an estimator, where $\hat{f}_\tau$ denotes an estimator of the threshold. This is the most common approach but its performance is heavily dependent on the choice of the bandwidth parameter. Baíllo and Cuevas (2006) were interested in choosing the best smoothing parameter to reconstruct a level set in the context of quality control. It was obtained by minimizing a cross-validation estimate of the probability of a false alarm. Samworth and Wand (2010) proposed an automatic rule to select the window for dimension 1 by deriving a uniform-in-bandwidth asymptotic approximation of an specific set estimation risk function. Of course, it is also possible to consider classical methods such as Seather and Jones or cross validation to calculate the bandwidth parameter although they are not specific to estimate level sets.

Another posibility consists of assuming that the set of interest satisfies some geometric condition such as convexity. Excess mass approach estimates the level set as the set of greatest mass and minimum volume under the assumed shape restriction. For example, Müller and Sawitzki's method assumes that the number of connected components is known, see Müller and Sawitzki (1991).

We can also consider hybrid methods. As the name suggests, they assume geometric restrictions and they use a kernel density estimator to decide which sample points are in the level set. For example, a generalization of the convexity is the property considered by granulometric smoothing method, see Walther (1997), and we proposed a new hybrid method to estimate convex and $r-$convex sets.

We have studied these three groups of automatic methods to reconstruct level sets. We have compared them through a detailed simulation study.

References

[1] Baíllo, A. and Cuevas, A. (2006). Parametric versus nonparametric tolerance regions in detection problems. Computational Statistics 21, 527-536.

[2] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. Journal of the American Statistical Association 86, 738- 746.

[3] Samworth, R.J. and Wand , M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. The Annals of Statistics 38, 1767-1792.

[4] Walther, G. (1997). Granulometric smoothing. Annals of Statistics 25, 2273-2299.

Robust critical values for unit root tests for series with conditional heteroscedasticity errors: An application of the simple NoVaS transformation

Panagiotis Mantalos

Departments of Statistics,

Swedish Business School at Orebro University, Sweden.

ABSTRACT

In this paper, we introduce a set of critical values for unit root tests that are robust in the presence of conditional heteroscedasticity errors using the normalizing and variance-stabilizing transformation (NoVaS) in Politis (2007) and examine their properties using Monte Carlo methods. In terms of the size of the test, our analysis reveals that unit root tests with NoVaS-modified critical values have actual sizes close to the nominal size. For the power of the test, we find that unit root tests with NoVaS-modified critical values either have the same power as, or slightly better than, tests using conventional Dickey–Fuller critical values across the sample range considered.

Keywords: Critical values, normalizing and variance-stabilizing transformation, unit root tests

JEL Classification Codes: C01, C12, C15

Efficient small-area calibration estimators integrating data from different surveys

T. Merkouris
Athens University of Economics and Business

Abstract

Small area estimation can be improved by integration of comparable area-specific data from different surveys. We consider best linear unbiased estimation (BLUE) of a small area total using a combination of small area estimators from the different surveys, and show that this can be done through a special combining calibration procedure. The resulting calibration estimator is a composite of optimal regression estimators when the calibration procedure incorporates information on auxiliary variables with known totals. This design-based direct estimator involves only area-specific data on the variables of interest, and its efficiency depends on the level (population or area) at which the auxiliary information from the various surveys is incorporated. It will be shown how a convenient substitute of small area BLUE can be developed as a particular type of generalized regression estimation, and how these two modes of estimation are equivalent under certain sampling conditions. Departures from the assumed comparability of small area data from the different surveys, and how they could be dealt with, will also be discussed.

Non-Asymptotic Exponential Bounds for Self-Normalized Sums

Emmanuelle GAUTHERAT
University of Reims and Laboratory of Statistics of CREST
3 avenue Pierre Larousse, 92245 Malakoff, France
emmanuelle.gautherat@gmail.com

keywords : Empirical Likelihood, Self-normalized sums, exponential bound, Hoeffding and Pinelis inequalities, symmetrization.

Abstract

It has been recognized by many authors that the behavior of empirical likelihood and its generalizations can be controlled by the tail of self-normalized mutivariate sums (see Bertail, Gautherat and Harari, 2009, Jing and Wang, 1999). In this talk, we improve non-asymptotic explicit exponential bounds for self-normalized sums in the multivariate case, which extend some results by Hoeffding (1963) and Pinelis (1994). More precisely, $Z_1, \cdots, Z_n$ are n random centered independent and identical distributed vectors in $\mathbb{R}^q$, and we call self-normalized sum the quantity $\sqrt{n\overline{Z}' S_n^{-2} \overline{Z}}$ where $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})(Z_i - \overline{Z})'$. We focus on results like

$$\mathbb{P}\left(n\overline{Z}' S_n^{-2} \overline{Z} > u\right) \leq C(u),$$

with $C(u)$ totally explicit, with exponential tail up to some polynomial factors (depending on q). We consider both the symmetric and non symmetric cases using Pachenko(2003)'s symmetrization approach.

joint work with P. Bertail, E. Gautherat and H. Harari.

Bertail, P. , Gautherat E. et Harari-Kermadec, H., 2009, Exponential inequalities for self normalized sums, Electronic Communications in Probability, 13, paper 57, 628-640.

Jing, Bing-Yi and Wang, Qiying, 1999, *An Exponential nonuniform Berry-Esseen bound for self-normalized sums*, Annals of Probability, vol = 27, n 4, 2068-2088

Hoeffding, Wassily, 1963. Probability inequalities for sums of bounded variables. Journal of the American Statistical Association, 58:1330.

Panchenko, Dmitry, 2003, *Symmetrization approach to concentration inequalities for empirical processes*, Annals of Probability, vol = 31, n 4, 2068-2081

Pinelis, Iosif, 1994. Probabilistic problems and Hotellings t2 test under a symmetry condition. Annals of Statistics, 22(1):357368.

Empirical energy minimizers for general Hadamard differentiable parameters

Patrice Bertail
University of Paris-Ouest and Laboratory of Statistics of CREST
3 avenue Pierre Larousse, 92245 Malakoff, France
patrice.bertail@gmail.com

keywords : Empirical Likelihood, ϕ-divergence, quasi-empirical likelihood, Hadamard differentiable parameters, duality.

Abstract

In this talk, we study some extensions of the empirical likelihood method, when the Kullback distance is replaced by some general convex distance or I_{γ^*} divergence. We recall some duality results which allow to establish the validity of empirical likelihood or empirical energy minimizers for a large class of Hadamard differentiable functionals. We propose to use instead of empirical likelihood some regularized form or quasi-empirical likelihood method, corresponding to a convex combination of the Kullback and χ^2 distance. Such method is known as log-quadatric proximal method in the convex optimization literature and enjoy some interesting properties from a purely algorithmic point of view. We show that for some adequate choice of the weight in this combination, the corresponding quasi-empirical likelihood is Bartlett correctable. We also show how the behavior of these quantity is related to the behavior of self normalized sums in the case of M-estimation.

Joint work with P. Bertail, H. Harari and E. Gautherat

About some empirical likelihood based confidence regions for extreme values statistics

Julien Worms
UFR Sciences, Universit de Versailles
45 avenue des Etats Unis 78035 Versailles cedex
julien.worms@uvsq.fr

keywords : Empirical Likelihood, extreme values, tail index.

Abstract

This talk will be devoted to the use of the empirical likelihood methodology for computation of confidence intervals/regions for classical parameters in the extreme values statistical field. The main parameter of interest will be the extreme value index γ, but joint estimation with the scale parameter in the Peaks Over Threshold framework will also be considered (in the heavy tail - *i.e.* Fréchet - case). The calibration problem, *i.e.* accuracy of the method in terms of coverage probabilities, will be adressed by simulations and comparison with other methods (some of them relying on asymptotic normality and estimation of the asymptotic variance). ”

Joint work with Rym Worms-Ramdani

Empirical likelihood for time series

Hugo Harari-Kermadec
ENS Cachan et Laboratoire SAMM,
Universit Paris 1, 90, rue de Tolbiac, 75634 PARIS CEDEX 13
hugo.harari@ens-cachan.fr

Abstract

In this talk, we show how Empirical Likelihood can be used on times series, for different kinds of dependence. In particular, we introduce an algorithm to deal with non-stationary times series, when the non-stationarity is due to a periodicity. PARMA are classical models with this kind of non-stationnarity. A key role is played by the construction of data blocs adapted to the dependence structure.

Joint work with Jacek Leśkow

Patrick Perry

Title: A Parametric View of Some Nonparametric Network Algorithms

Methods for analyzing network data are not always grounded in rigorous statistical theory. They do not make explicit modeling assumptions, and they do not come with consistency guarantees. Here we analyze two popular network tools, the expected degree model and modularity-based community detection.

We show that the expected degree model is the maximum likelihood estimate under a certain log-linear model, and we show that modularity-based community detection is approximate maximum likelihood estimation under a different model. These results enable us to extend the analyzed ad hoc network algorithms by applying standard statistical model selection and estimation devices.

Patrik Guggenberger

Title: "The Correct Asymptotic Size of LM and CLR Tests for Moment Condition Models Without Any Identification Assumptions"

Abstract:
An influential paper by Kleibergen (2005) introduces Lagrange multiplier (LM) and conditional likelihood ratio-like (CLR) tests and confidence intervals for nonlinear moment condition models. These procedures are designed to have good size performance even when the parameters are unidentified or poorly identified. The asymptotic results in the literature for these procedures, however, are incomplete. Results are available only for certain types of weak identification. The correct asymptotic size and the asymptotic similarity of these procedures (in a uniform sense) have not been established. We do so in this paper. For the special case of a linear IV regression model with two or more right-hand side endogenous variables, the results also are new to the literature.

Signed Rank Tests in Symmetric IC Models

Pauliina Ilmonen, Davy Paindaveine
Université Libre de Bruxelles

Abstract

We consider semiparametric location-scatter models for which the p-variate observation is generated through

$$X = \Lambda Z + \mu,$$

where μ is a real p-vector, Λ is a real invertible $p \times p$ matrix, and the unobserved random p-vector Z has marginals that are centered and mutually independent but are otherwise unspecified. As in blind source separation and independent component analysis (ICA), the parameter of interest is Λ. On the basis of n independent copies of X, we consider, under symmetry assumption on Z, *signed-rank* one-sample testing procedures for standardized version L of the mixing matrix matrix Λ. We do not only consider the problem of testing the null $\mathcal{H}_0 : L = L_0$ against the alternative $\mathcal{H}_1 : L \neq L_0$, for some fixed L_0, but we also extend the procedures to testing an abitrary (fixed) linear hypothesis on L. Our testing procedures are based on general score functions, which has an important impact on asymptotic results.

References

[1] P. Ilmonen, and D. Paindaveine (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics*, **39(5)** 2448–2476.

[2] P. Ilmonen, and D. Paindaveine (2012). Signed Rank Tests in Symmetric IC Models. *manuscript*.

Analyzing functional data with lattice spatial dependence: A generalization of LISA map

Pedro Delicado[1] **and Sonia Broner**[2]

[1]Universitat Politècnica de Catalunya, Bercelona, Spain
[2]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

Abstract:
In this work we analyze functional data with spatial dependence of lattice type: The geographical unit under study (a region, a country, the world, ...) is divided in smaller sub-areas (mainly because administrative reasons) and a functional data is observed at each sub-area (lattice data are also referred to as areal data). Spatial dependence arises when the data observed in a given sub-area are more similar to those observed in the neighboring areas than those corresponding to remote areas. LISA map is one of the descriptive tools most extensively used in spatial univariate data analysis of lattice type. LISA map represents graphically the significant values of local dependence Moran's index.

We propose a LISA map generalization based on the distances between the features observed in different areas. Therefore it can be applied when the observations are functional data. In particular we analyze the whole population pyramids (considering them as density functions) of all countries in the world. Our first aim is to detect whether there is spatial dependence between population pyramids or not. In addition to that, a modification of the proposed algorithm allows us to detect spatial clusters of neighboring countries with extremely similar population pyramids. To define the distance between population pyramids we consider that the density functions are compositional data with infinite dimension.

NONPARAMETRIC REGRESSION WITH HOMOGENEOUS GROUP TESTING DATA

PETER HALL

UNIVERSITY OF MELBOURNE, DEPARTMENT OF MATHEMATICS & STATISTICS

In this talk we introduce new nonparametric predictors for homogeneous pooled data in the context of group testing for rare abnormalities, and show that they achieve optimal rates of convergence. In particular, when the level of pooling is moderate then, despite the cost savings, the method enjoys the same convergence rate as in the case of no pooling. In the setting of ‘‘over-pooling” the convergence rate differs from that of an optimal estimator by no more than a logarithmic factor. Our approach improves on the random-pooling nonparametric predictor, which is currently the only nonparametric method available, unless there is no pooling, in which case the two approaches are identical.

Nonparametric Trending Regression with Cross-Sectional Dependence

Peter M. Robinson

London School of Economics

Abstract

Panel data, whose series length T is large but whose cross-section size N need not be, are assumed to have common time trend, of unknown form. The model includes additive, unknown, individual-specific components and allows for spatial or other cross-sectional dependence and/or heteroscedasticity. A simple smoothed nonparametric trend estimate is shown to be dominated by an estimate which exploits availability of cross-sectional data. Asymptotically optimal bandwidth choices are justified for both estimates. Feasible optimal bandwidths, and feasible optimal trend estimates, are asymptotically justified, finite sample performance of the latter being examined in a Monte Carlo study. Potential extensions are discussed.

Random function priors for exchangeable graphs and arrays

Peter Orbanz
University of Cambridge

Random arrays represent graph-valued data, networks and relational data. What is an appropriate nonparametric Bayesian model for such data? The relevant notion of exchangeability to substitute for de Finetti's theorem has been identified over 30 years ago by Aldous and Hoover and is already used in statistics. I will first discuss recent work in discrete analysis which extends this probabilistic concept into a full-fledged analytical theory that permits a rigorous formalization of statistical models. I will then sketch the general form of nonparametric Bayesian models for such data, and present one specific model as an example.

EXTREME VALUE THEORY WITH OPERATOR NORMING

Peter Scheffler, University of Siegen

Abstract: A new approach to extreme value theory is presented for vector data with heavy tails. The tail index is allowed to vary with direction. Basic asymptotic theory is developed, using regular variation and extremal integrals.

Shedding Confusion on the Light: A Contrary Perspective on Infinite-Dimensional Inverse Problems

Philip B. Stark
Department of Statistics
University of California, Berkeley

19 May 2012

Abstract

Infinite-dimensional inverse problems are ubiquitous in physical science. What counts as a "solution" seems very different for scientists, mathematicians, and statisticians. Scientists seem concerned primarily with constructing a model that agrees adequately with the data, and generally proceed by discretizing the problem and using regularized nonlinear least squares: To solve an inverse problem is to make a picture. Statistical properties—if they are considered at all—are typically an afterthought. Mathematicians seem concerned primarily with stability, uniqueness, and construction; there are infinitely many observations and "observational error," if it is considered, is a vector with bounded ℓ_2 norm. To solve an inverse problem is to prove existence and uniqueness of a model that agrees with a given set of data and to understand the stability of the inverse. Statisticians seem concerned primarily with asymptotic behavior under various hypotheses about the unknown (e.g., bounded Sobolev norm, sparsity in some representation, ...) with the number of observations going to infinity and/or the noise (generally assumed to be zero-mean normal) going to zero. To solve an inverse problem is to give a method that is asymptotically most wonderful, under an untestable set of unrealistic assumptions about the model and the data.

I will argue that we should care about inference, not estimation; about finite-sample properties, not asymptotics; about constraints

that come from the physics, not hypothetical constraints that would make our methods work well; about reliability despite the tail behavior of the noise, rather than on assumed normality; and about systematic measurement error. I will sketch a few examples of rigorous finite-sample inference using physical constraints. I will suggest—heretically—that in many applications "can't say anything without information that is not available" is the real solution infinite-dimensional inverse problems.

Philippe vieu

Title: "Central Limit Theorems for Stratified Spaces",

Abstract: In many applications, data occur on non-manifolds that are unions of manifolds of varying dimensions. We provide examples and investigate the asymptotic distributions of sample means. These limiting distributions exhibit properties that are typical for non-Euclidean spaces such as "omitting, hitting and sticking" to singular strata.

Bayesian asymptotics with misspecified models

Pierpaolo De Blasi*
University of Torino and Collegio Carlo Alberto, Italy
Stephen G. Walker
School of Mathematics, Statistics, & Actuarial Science,
University of Kent, Canterbury, UK

Abstract

In this paper, we study the asymptotic properties of a sequence of posterior distributions based on an independent and identically distributed sample and when the Bayesian model is misspecified. We find a sufficient condition on the prior for the posterior to accumulate around the densities in the model closest in the Kullback–Leibler sense to the true density function. Examples are presented.

Key words and phrases: Asymptotics, Consistency, Misspecified model.

*address for correspondence: pierpaolo.deblasi@unito.it

Nonparametric (smoothed) likelihood and integral equations

Piet Groeneboom

Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: P.Groeneboom@tudelft.nl, *url:* http://dutiosc.twi.tudelft.nl/~pietg/

Abstract: We show that there is an intimate connection between the theory of nonparametric (smoothed) maximum likelihood estimators for certain inverse problems and integral equations. This is illustrated by estimators for interval censoring and deconvolution problems. We also discuss the asymptotic efficiency of the MLE for smooth functionals in these models.

AMS 2000 subject classifications: Primary 62G05, 62N01; secondary 62G20.
Keywords and phrases: interval censoring, deconvolution, maximum likelihood estimators, maximum smoothed likelihood estimators, integral equations, smooth functionals, efficient estimation, asymptotic distribution.

Martingale limit theorems revisited and non-linear cointegrating regression

Qiying Wang
The University of Sydney

April 2, 2012

Abstract

For a certain class of martingales, the convergence to mixture normal distribution is established under the convergence in distribution for the conditional variance. This is less restrictive in comparison with the classical martingale limit theorem where one generally requires the convergence in probability. The extension partially removes a barrier in the applications of the classical martingale limit theorem to non-parametric estimates and inferences with non-stationarity, and enhances the effectiveness of the classical martingale limit theorem as one of the main tools in the investigation of asymptotics in statistics, econometrics and other fields. The main result is applied to the investigations of asymptotics for the conventional kernel estimator in a nonlinear cointegrating regression, which improves the existing works in literature.

A copula mixture model for assessing reproducibility and combining information for high-throughput data

Abstract:
Reproducibility is essential to reliable scientific discovery in high-throughput experiments. Recently, we developed a unified approach to measure the reproducibility of findings identified from replicate experiments and select findings according to their reproducibility. Unlike traditional correlation-based measures of reproducibility, this approach quantitatively assesses when the findings are no longer consistent across replicates using a copula mixture model, and reports a reproducibility score, which we call the ”irreproducible discovery rate” (IDR) analogous to the FDR. This score can be computed at each set of replicate ranks and permits the principled setting of thresholds both for assessing reproducibility and combining replicates.

In this talk, I will present this method, and discuss its extension to grouped and truncated data, as well as a rank-based approach to aggregate information from multiple samples based on this framework. We demonstrate this method in some next-generation sequencing data.

Cointegration and Phase Synchronization: Bridging Two Theories

Rainer Dahlhaus, University of Heidelberg

In this talk we present with VEC-state oscillators a new multivariate time series model for oscillators with random phases. In particular the phases may be synchronized. The model is a nonlinear state space model where the phase processes follow a vector error correction model used in econometrics to model cointegration. We demonstrate the relevance of this model for phase synchronization. In that way we bridge the theories of cointegration and phase synchronization which have been important theories in econometrics and physics, respectively. The common ground of both theories is that they describe the fluctuation of some multivariate random process around an equilibrium. We demonstrate how the methods from cointegration can be applied to phase synchronization. In particular we consider an unidirectionally coupled R ssler-Lorenz system and identify the unidirectional coupling, the phase synchronization equilibrium and the phase shifts with cointegration tests.

Rank-based Testing for Semiparametric Cointegration Models

Marc Hallin[a,b,c], Ramon van den Akker[c], Bas J.M. Werker[c,d]

[a] *ECARES, Université Libre de Bruxelles*
[b] *ORFE, Princeton University*
[c] *Econometrics group, CentER, Tilburg University*
[d] *Finance group, CentER, Tilburg University*

Abstract

This paper discusses asymptotically efficient testing for hypotheses about the cointegrating rank or about the cointegrating vectors in a cointegration model with elliptically distributed innovations. The model is semiparametric in the sense that the radial density and scatter matrix of the innovations are both unknown. The tests developed use a multivariate notion of ranks and are asymptotically distribution-free. The tests are build on a reference density that can be chosen freely. Validity of the test in terms of asymptotic size is guaranteed irrespective of the reference density. The asymptotic power of the test improves when the chosen reference density happens to be closer to the actual innovation density. A suitably estimated reference density leads to fully semiparametrically efficient tests.

May 14, 2012

Random effects generalized linear models have played an important role in the analysis of longitudinal data. A common requirement of the existing estimation approaches to these models is the specifications of random effects distributions. The inference based on specific distributional assumption might be sensitive to the choice of random effects distributions under certain circumstance. In addition, the parametric distributional requirement beyond multivariate normality usually restricts the flexibility of modeling covariance structures for longitudinal data. In this paper, we incorporate serially dependent distribution-free random effects into Tweedie generalized linear models to accommodate a wide range of covariance structures for longitudinal data. An optimal estimation of our model has been developed using the orthodox best linear unbiased predictors of random effects. Our approach unifies population-averaged and subject-specific inferences. Our method is illustrated through the analyses of patient-controlled analgesia data and Framingham cholesterol data.

FUNCTIONAL PREDICTION FOR THE RESIDUAL DEMAND IN ELECTRICITY SPOT MARKETS

Germán Aneiros[(1)], Ricardo Cao*[(1)], Juan M. Vilar-Fernández[(1)]and Antonio Muñoz-San-Roque[(2)]

[(1)]Departament of Mathematics, Faculty of Computer Science, Universidade da Coruña, Campus de Elviña, 15071 A Coruña, Spain
[(2)]Escuela Técnica Superior de Ingeniería, ICAI, Universidad Pontificia de Comillas, C/ Alberto Aguilera, 25, 28015 Madrid, Spain

ABSTRACT

Nowadays, in many countries all over the world, the production and sale of electricity is traded under competitive rules in free markets. The agents involved in this market: system operators, market operators, regulatory agencies, producers, consumers and retailers have a great interest in the study of electricity load and price. Since electricity cannot be stored, the demand must be satisfied instantaneously and producers need to anticipate to future demands to avoid overproduction. Good forecasting of electricity demand is then very important from the system operator viewpoint. In the past, demand was predicted in centralized markets but competition has opened a new field of study. On the other hand prediction of residual demand of an agent is a valuable tool to establish good bidding strategies for the agent itself. Consequently, prediction of electricity residual demand is a significant problem in this sector.

Residual demand curves have been considered previously in the literature. In each hourly auction, the residual demand curve is defined as the difference of the combined effect of the demand at any possible price and the supply of the generation companies as a function of price. Consequently 24 hourly residual demand curves are obtained every day. These curves are useful tools to design optimal offers for companies operating in a day-ahead market. We focus on one day ahead forecasting of electricity residual demand curves. Therefore, for each day of the week, 24 curve forecasts need to be computed.

Hourly residual demand curves are predicted using nonparametric regression with functional explanatory and functional response variables. Semi-functional partial linear models are also used in this context. Forecasted wind energy as well as forecasted hourly price and demand are incorporated as explanatory linear variables in the model. Results from the electricity market of mainland Spain are reported. The new forecasting functional methods are compared with a naive approach.

An Empirical Characteristic Function Approach to Selecting a Transformation to Improve Normality or Symmetry

Richard A. Johnson
University of Wisconsin

In-Kwon Yeo
Sookmyung Women's University

Xinwei Deng
Virginia Tech

Abstract

We study the problem of transforming to normality or another symmetric distribution. Our approach is to minimize the integrated weighted squared difference between the empirical characteristic function of the transformed data and the characteristic function of the normal distribution or other symmetric target distribution. The choice of weight function is crucial because the behavior of a characteristic function near zero is of prime importance.

Asymptotic properties are established when a random sample is selected from an unknown distribution.

After describing our estimation procedure in terms of a U-statistic, we propose a method for calculating an influence function related to U-statistics.

The influence function calculation and Monte Carlo simulations, in the context of Box-Cox transformations, suggest that our estimates are less sensitive, than the maximum likelihood estimators, to a few outliers. They also compare favorably with the robust estimators in Carroll.

Independent component analysis via nonparametric maximum likelihood estimation

(May 14, 2012)

Abstract

Independent Component Analysis (ICA) models are very popular semiparametric models in which we observe independent copies of a random vector $X = AS$, where A is a non-singular matrix and S has independent components. We propose a new way of estimating the unmixing matrix $W = A^{-1}$ and the marginal distributions of the components of S using nonparametric maximum likelihood. Specifically, we study the projection of the empirical distribution onto the subset of ICA distributions having log-concave marginals. Remarkably, from the point of view of estimating the unmixing matrix, it turns out that it makes no difference whether or not the log-concavity is correctly specified. The approach is further justified by both theoretical results and a simulation study.

Small Sample LD50 Confidence Intervals Using Saddlepoint Approximations

Robert L. Paige (with Phillip Chapman and Ronald Butler)
Department of Mathematics and Statistics
Missouri University of Science and Technology
(formerly University of Missouri-Rolla)
Rolla, MO 65409
E-mail: paigero@mst.edu

April 3, 2012

Abstract

Confidence intervals for the median lethal dose (LD50) and other dose percentiles in logistic regression models are developed using a generalization of the Fieller theorem for exponential families and saddlepoint approximations. Simulation results show that, in terms of one-tailed and two-tailed coverage, the proposed methodology generally outperforms competing confidence intervals obtained from the classical Fieller, likelihood ratio, and score methods. In terms of two-tailed coverage, the proposed method is comparable to the Bartlett-corrected likelihood ratio method, but generally outperforms it in terms of one-tailed coverage. An extension to the competing risk setting is presented that allows binary response adjustments to be made using observed censoring times.

KEYWORDS: Bartlett correction, binary data, bootstrap, competing risks, Fieller's method, likelihood ratio, saddlepoint approximation, score statistic.

L1-Penalization Applied to Group Imaging Data Analysis

Rosanna Overholser[1] and Ronghui Xu[1,2*]

[1]Department of Mathematics

[2]Department of Family and Preventative Medicine

University of California, San Diego

*Presenting author

May 8, 2012

Abstract

The LASSO was initially introduced for the purposes of estimation and variable selection in linear regression. Most work on the LASSO has included the assumption of independent observations. Several papers have recently extended the LASSO to linear mixed models for clustered data. We will consider a further extension of the LASSO to general linear mixed models for functional data that contain serial correlation. Regression splines for correlated data can be formulated as a general linear mixed model so the problem of knot selection for splines is equivalent to variable selection of fixed effects. We can therefore use the LASSO to simultaneously select knots and estimate variance parameters. We apply our methods to functional MRI time course data from a group of subjects.

Sufficient Dimension Reduction for Longitudinally Measured Predictors

Ruth M. Pfeiffer

Biostatistics Branch, Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, MD 20892-7244, USA Email: pfeiffer@mail.nih.gov

We propose a method to combine several predictors (markers) that are measured repeatedly over time into a composite marker score without assuming a model and only requiring a mild condition on the predictor distribution. Assuming that the first and second moments of the predictors can be decomposed into a time and a marker component via a Kronecker product structure, that accommodates the longitudinal nature of the predictors, we develop first moment sufficient dimension reduction techniques to replace the original markers with linear transformations that contain sufficient information for the regression of the predictors on the outcome. These linear combinations can then be combined into a score that has better predictive performance than the score built under a general model that ignores the longitudinal structure of the data. Our methods can be applied to either continuous or categorical outcome measures. In simulations we focus on binary outcomes and present an algorithm that extends Sliced Inverse Regression (SIR) to that settingand. We show that our method outperforms existing alternatives using the AUC, the area under the receiver-operator characteristics (ROC) curve, as a summary measure of the discriminatory ability of a single continuous diagnostic marker for binary disease outcomes.

This is joint work with Liliana Forzani, Instituto de Matemática Aplicada del Litoral and Facultad de Ingenieri a Quimica, CONICET and UNL, Santa Fe, Argentina, and Efstathia Bura from the Department of Statistics, George Washington University, Washington, DC.

An application of generic chaining in high-dimensional statistics

Sara van de Geer
Seminar for Statistics, ETH Zürich

We review the concept of generic chaining as introduced by Talagrand [1996], and show how it can be used in non-linear ℓ_1-regularized estimation problems. The results lead to insights in the geometrical properties of ℓ_1 balls in high-dimensional spaces.

References

M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

Penalized maximum likelihood estimation of a sparse DAG

Sara van de Geer and Peter Bühlmann
Seminar for Statistics, ETH Zürich

Let X be an $n \times p$ matrix of observations. A directed acyclic graph (DAG) models the observations as

$$X = XB_0 + E,$$

where each row of E is $\mathcal{N}(0, \Omega_0)$-distributed, with Ω_0 a diagonal matrix. Moreover, writing the columns of X as X_k, $k = 1, \ldots, p$, and those of E as ϵ_j, $j = 1, \ldots, p$, it is assumed that ϵ_j and X_k are independent whenever X_k is a parent of X_j. There are several ways to represent the DAG (B_0, Ω_0). We consider one with the minimal number of edges. We estimate the DAG using maximum likelihood with a penalty proportional to the number of edges. We assume that any representation of the DAG has at least a given proportion of its non-zero coefficients above the noise level, and that the number of edges per node is sufficiently smaller than $n/\log p$. We prove convergence in Frobenius norm of the penalized maximum likelihood estimator, and show that it has about the same number of edges as the true DAG.

ON MIXING CONDITIONS FOR ASYMPTOTICS IN FUNCTIONAL TIME SERIES

SHAHIN TAVAKOLI
ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE (EPFL), LAUSANNE, SWITZERLAND

Motivated by DNA minicircle data from Molecular Dynamics, we investigate mixing conditions that enable us to draw statistical inferences for stationary functional data. We are interested in general stationary processes -as opposed to linear processes. We review existing functional mixing conditions, examples of processes that satisfy them, and asymptotic results they allow for. We then consider moment-based functional mixing conditions, and show how these can be used to recover or extend existing asymptotic results. We also consider examples of functional processes satisfying our mixing conditions, and probe the stability of our conclusions under discrete observation of the functional time series. (based on joint work with Victor M. Panaretos, EPFL)

A note on estimation in Hilbertian linear models

Siegfried Hörmann

Départment de Mathématique, Université Libre de Bruxelles, CP 215, Boulevard du Triomphe, B-1050 Bruxelles, Belgium

We study estimation of the operator Ψ in the linear model $Y = \Psi(X) + \varepsilon$, when X and Y take values in Hilbert spaces H_1 and H_2, respectively. Our main objective is to obtain consistency without imposing some rather inconvenient technical assumptions that have been used in the literature. We develop our theory in a time dependent setup which comprises as important special case the autoregressive Hilbertian model.

This talk is based on joint work with Łukasz Kidziński.

Bayesian Semiparametric SAE with Dirichlet Process priors

Silvia Polettini
Sapienza Università di Roma, P.le A. Moro 5 – 00185 Roma
silvia.polettini@uniroma1.it

Borrowing strength in small area estimation is most often achieved through mixed effects regression models. The default assumption of normality may fail to represent the distribution of the random effects for several reasons: missing covariates can lead to multimodal distributions; the distribution may otherwise be skewed. Any distributional assumption is difficult to check, being the random effects latent variables.

Accurate estimation of the random effects is crucial for predicting small area quantities, and the effect on model estimates of distributional assumptions on the random effects is shown to be important in the literature. Although point predictions of small area means are robust from deviations from normality, the latter may affect the precision of such predictions; also, estimation of nonlinear functionals may suffer from misrepresentation of the law of the random effects. The problem affects both frequentist and Bayesian analysis, although availability of MCMC techniques makes computational convenience less relevant in the latter framework.

For the reasons mentioned above, it would be important to rely on a model that has a flexible specification of the random components, so to achieve a greater flexibility and robustness against model misspecifications. For the Fay-Herriot model, Fabrizi and Trivisano (2010) develop two robustified versions by describing the random effects by either an exponential power (EP) or a skewed EP distribution and investigate robustness of such Fay-Herriot-type models under deviations from normality. Their aim is to understand whether estimates of linear and especially nonlinear functionals such as the c.d.f. are sensitive to deviations from normality of the random effects. Although the models proposed by Fabrizi and Trivisano are based on distributions that generalize the normal, yet these parametric models may fail to adequately describe the distribution of the random effects, and again the problem of checking the adequacy of these models arises.

In this paper we consider a different extension of the Fay-Herriot model, namely a semiparametric Bayesian area level linear mixed effects model, in which the random effects are modelled through a Dirichlet process with precision parameter M and base measure ϕ which is natural to assume here to be a normal distribution. The representation above not only relaxes the normal assumption, but also provides an enlarged model for describing the random effects.

In this context, the main aim is to assess improvements in precision of small area predictions.

The semiparametric setting described is reported in Kyung et al. (2009, 2010) to reduce the variability of the regression parameters estimates, producing uniformly shorter HPD intervals than the standard normal random effects models.

The model is applied to a pseudo-sample drawn from a known population using a standard sampling scheme. The target is here the estimation of the unemployment rate. A set of covariates was introduced and used without any model selection procedure. The true figures were known and therefore could be used to assess the estimators. For the characteristics of the sample, the random effects were not designed and therefore *a priori* there is no specific parametric family that can fully describe the random area effects.

The estimates obtained under the DPP model were compared with the EBLUP. For comparison, the standard hierarchical Bayesian (HB) model with "vague priors" is also estimated. The model formalization of the HB model coincides with the DPP except for the definition of the random effects, assumed to be normally distributed. With the "vague" prior choice HB predictions coincide with those obtained from the EBLUP (see e.g. Rao, 2003).

The EBLUP and DP prior point estimators of small area percentages perform similarly, as expected, and both agree quite well with the true figures.

To assess the model, it is important to compare the estimator above with the EBLUP with respect to measures of variability and coverage, being this one the feature where the effect of a more flexible specification of the random effects is expected. Since the model was not designed to achieve "calibration" between posterior variance and MSE (see Rao, 2003, p. 238), comparing posterior variances and MSE is not completely appropriate. To assess the variability of posterior quantities we refer to the standard hierarchical Bayesian (HB) model as a benchmark.

The application shows that the area level flexible model tends to result in more accurate estimation of small area quantities. Further advantages are expected for unit level models aimed at estimating nonlinear functionals such as quantities in the Foster-Greer-Thorbecke (FGT) class of poverty measures.

DIAGNOSTIC PROCEDURES FOR SPHERICALLY SYMMETRIC DISTRIBUTIONS

Simos G. Meintanis

Department of Economics, National and Kapodistrian University of Athens
Athens, Greece

Abstract. Goodness–of–fit tests are proposed for the null hypothesis that the distribution of an arbitrary random variable belongs to the general family of spherically symmetric distributions. The test statistics utilize a well known characterization of spherical symmetry which incorporates the characteristic function of the underlying distribution. An estimated version of this characterization is then employed both in the Kolmogorov–Smirnov sense and the Cramér–von Mises sense and yields corresponding test statistics. Both tests come in convenient forms which are straightforwardy applicable with the computer. Also the consistency of the tests is investigated under general conditions. Since results on the asymptotic null distribution are difficult to apply, an effective bootstrap procedure is used in order to actually carry out the tests. The behavior of the new methods is investigated with real as well as simulated data.

Keywords. Goodness–of–fit, Empirical characteristic function, Bootstrap test.

AMS 2000 classification numbers: 62G10, 62G20

INTERSECTION BOUNDS: ESTIMATION AND INFERENCE

VICTOR CHERNOZHUKOV, SOKBAE LEE, AND ADAM M. ROSEN

Abstract. We develop a practical and novel method for inference on intersection bounds, namely bounds defined by either the infimum or supremum of a parametric or nonparametric function, or equivalently, the value of a linear programming problem with a potentially infinite constraint set. Our approach is especially convenient for models comprised of a continuum of inequalities that are separable in parameters, and also applies to models with inequalities that are non-separable in parameters. Since analog estimators for intersection bounds can be severely biased in finite samples, routinely underestimating the size of the identified set, we also offer a median-bias-corrected estimator of such bounds as a natural by-product of our inferential procedures. We develop theory for large sample inference based on the strong approximation of a sequence of series or kernel-based empirical processes by a sequence of "penultimate" Gaussian processes. These penultimate processes are generally not weakly convergent, and thus non-Donsker. Our theoretical results establish that we can nonetheless perform asymptotically valid inference based on these processes. Our construction also provides new adaptive inequality/moment selection methods. We provide conditions for the use of nonparametric kernel and series estimators, including a novel result that establishes strong approximation for any general series estimator admitting linearization, which may be of independent interest.

Key words. Bound analysis, conditional moments, partial identification, strong approximation, infinite dimensional constraints, linear programming, concentration inequalities, anti-concentration inequalities, non-Donsker empirical process methods, moderate deviations, adaptive moment selection.

JEL Subject Classification. C12, C13, C14. AMS Subject Classification. 62G05, 62G15, 62G32.

Date: 1 November 2011.

We are especially grateful to Denis Chetverikov, Kengo Kato, Ye Luo, four anonymous referees, and the editor for making several extremely useful suggestions that have led to substantial improvements. We thank Richard Blundell, Andrew Chesher, Francesca Molinari, Whitney Newey, Nicolas Roys, Sami Stouli, and Jörg Stoye for detailed discussion and suggestions, and participants at numerous seminars and conferences for their comments. This paper is a revised version of "Inference on Intersection Bounds" initially presented and circulated at the University of Virginia and the Harvard/MIT econometrics seminars in December 2007, and presented at the March 2008 CEMMAP/Northwestern conference on "Inference in Partially Identified Models with Applications." We gratefully acknowledge financial support from the National Science Foundation, Economic and Social Research Council (RES-589-28-0001, RES-000-22-2761) and European Research Council (ERC-2009-StG-240910-ROMETA).

Victor Chernozhukov: Department of Economics, Massachusetts Institute of Technology, vchern@mit.edu.
Sokbae Lee: Department of Economics, Seoul National University and CeMMAP, sokbae@gmail.com.
Adam Rosen: Department of Economics, University College London and CeMMAP, adam.rosen@ucl.ac.uk.

Extending Rank Based Inference to Clustered Data when the Cluster Size is Potentially Informative

Somnath Datta, University of Louisville

We discuss how to extend rank based inference when the classical assumption of independence is violated due to clustering. Clustered data arise in a number of practical applications where observations belonging to different clusters are independent but observations within the same cluster are dependent. While making adjustments for possible cluster dependence, one should also be aware of the informative cluster size phenomenon which occurs when the size of the cluster is a random variable that is correlated to the outcome distribution within a cluster, often through a cluster specific latent factor. We demonstrate the correct rank based inference procedures in two examples: (i) testing for marginal symmetry and (ii) robust estimation of regression parameters.

Two Sample Tests for High Dimensional Covariance Matrices

Song Xi Chen
Department of Statistics, Iowa State University; and
Department of Business Statistics and Econometrics
and Center for Statistical Science, Peking University

We propose two tests for the equality of covariance matrices between two high-dimensional populations. One test is on the whole variance-covariance matrices, and the other is on off-diagonal sub-matrices which define the covariance between two non-overlapping segments of the high-dimensional random vectors. The tests are applicable (i) when the data dimension is much larger than the sample sizes, namely the ‘‘large p, small n" situations and (ii) without assuming parametric distributions for the two populations. These two aspects surpass the capability of the conventional likelihood ratio test. The proposed tests can be used to test on covariances associated with gene ontology terms. This is a joint work with Jun Li at Department of Statistics at Iowa State University.

Keywords: {High dimensional covariance; Large p small n; Likelihood ratio test; Testing for Gene-sets. }

Change points detection in functional data

Sophie Dabo-Niang ♯

♯ Université Charles De Gaulle, Lille 3, Laboratoire EQUIPPE, France.
sophie.dabo@univ-lille3.fr

Abstract

Functional data are becoming increasingly common in a variety of fields. Many studies underline the importance to consider the representation of data as functions. This has sparked a growing attention in the development of adapted statistical tools that allow to analyze such kind of data : functional data analysis (FDA). The aims of FDA are mainly the same as in the classical statistical analysis, e.g. representing and visualizing the data, studying variability and trends, comparing different data sets, as well as modeling and predicting,...

Recent advances in FDA allow to construct different change points detection methods, based on change in the mean or median curve. We review some procedures that have been used to detect functional time series change points in the mean curve. Theoretical advances on change points detection of dependent functional data, related to mean and median curves are provided. In addition, we show the good practical behaviors of these procedures on a sample of curves.

References

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

[2] Berkes, I., Gabrys, R., Horvath, L. and Kokoszka, P. (2009). *Detecting changes in the mean of functional observations*. J. R. Statist. Soc. B, **71** (5), 927-946.

[3] Bosq, D. (2000) Linear processes in function spaces. *Lecture Note in Statistics*, **149**, Springer-Verlag.

[4] Chebana, Fateh, Dabo-Niang, S and Ouarda, T. (2012). *Exploratory functional flood frequency analysis and outlier detection*. To appear in Water Resour. Res. doi :10.1029/2011WR011040

[5] Dabo-Niang, S and Yao, A-F. (2012). Spatial kernel density estimation for functional random variables. *METRIKA*. To appear. DOI 10.1007/s00184-011-0374-4.

[6] Ferraty, F. and Vieu, Ph., (2006). Nonparametric functional data analysis. *Springer-Verlag*.

[7] Ramsay, J.O. and Silverman, B.W. (2002). Applied functional data analysis; Methods and case studies. *Springer-Verlag, New York*.

A Penalized Empirical Likelihood Method in High Dimensions

S. N. Lahiri & Deep Mukhopadhyay
Texas A & M University

ABSTRACT

We formulate a penalized empirical likelihood (PEL) method for inference on the population mean when the dimension of the observations become unbounded with the sample size. We derive the asymptotic distribution of the PEL ratio statistic. We show that the limit distribution of the proposed PEL ratio statistic can vary widely depending on the correlation structure of the components of the observations. We consider all possible cases of serial dependence, namely, (i) non-Ergodic, (ii) long range dependent and (iii) short range dependent. We derive the limit laws in each case, all of which differ from the usual chi-squared limit of the empirical likelihood ratio statistic in the finite dimensional case. We propose a subsampling approximation for calibrating the PEL ratio test statistic and establish its validity. Finite sample properties of the method are investigated through a simulation study.

Soutir Bandyopadhyay

A Frequency Domain Empirical Likelihood Method for Irregularly Spaced Spatial Data

In this talk, we consider empirical likelihood methodology for irregularly spaced spatial data in the frequency domain. The main result of the paper shows that upto a suitable (and nonstandard) scaling, Wilk s phenomenon holds for the logarithm of the empirical likelihood ratio in the sense that it is asymptotically distribution free and has a chi-squared limit. As a result, the proposed spatial frequency domain empirical likelihood method can be used to build nonparametric, asymptotically correct confidence regions and tests for a class spectral parameters that are defined through spectral estimating equations. A major advantage of the method is that unlike the more common studentization approach, it does not require explicit estimation of the standard error, which itself is a difficult problem due to intricate interactions among several unknown quantities, including the spectral density of the spatial process, the spatial sampling density and the spatial asymptotic structure. Applications of the methodology to some important inference problems for spatial data are given.

Classification of Functions Using Data Depth

Stanislav Nagy

Charles University in Prague, Czech Republic, email: nagy@karlin.mff.cuni.cz

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

Nonparametric classification of data from certain subspaces of continuous functions $C([0,1])$ is discussed. Special attention is paid to depth–based classification rule and its possible generalisations. The decision is related to the concept of data depth, which is in this case a functional

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

It provides a measure of the centrality of an observation with respect to a data set or a distribution. Recently, several authors proposed their notions of depth for functional data (Fraiman and Muniz [3], López-Pintado and Romo [6], Cuevas and Fraiman [2]). Most of these depth functionals rely on the integration of a univariate depth measure over the domain $[0,1]$. Thus, none of them is able do deal with the shape of functions.

This problem is demonstrated in a functional classification task. A new class of depth functionals is utilized in order to handle it. The simplicial depth described by Liu [5] along with Fraiman-Muniz method are employed to involve derivatives into depth computation. The performance of the new approach is compared to similar results obtained by Cuveas et al. [1] in a simulation study of functional data supervised classification. We show that proper derivative using in combination with DD-plot (depth-depth plot) techniques proposed by Li et al. [4] is a powerful tool not only for the classification of functional observations.

At the end some minor issues concerning the consistency of depth for functional data are discussed.

References

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

[2] Cuevas, A. and Fraiman, R.: 2007, 'On depth measures and dual statistics. A methodology for dealing with general data'. *Journal of Multivariate Analysis* **100**(4), 753–766.

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

[4] Li, J., Cuesta-Albertos, J. A. and Liu, R. Y.: 'DD-Classifier: Nonparametric Classification Procedure Based on DD-plot'. *Journal of the American Statistical Association*, to appear.

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

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

Title: **Advances in permutation tests for covariates in a mixture model for preference data analysis**

Authors: Stefano Bonnini[a], Francesca Solmi [b]

[a] Department of Economics & Management, University of Ferrara, Italy

[b] Department of Statistical Sciences, University of Padova, Italy

ABSTRACT

The rating problem arises very often in statistical surveys, where respondents are asked to evaluate several topics of interest (products, services, treatments, etc.). In this framework, a new approach is represented by a class of mixture models (Covariates in the mixture of Uniform and shifted Binomial distributions, CUB models), proposed by Piccolo (2003), D'Elia and Piccolo (2005) and Piccolo (2006). Together with parametric inference, a permutation solution to test for covariates effects, when a univariate response is considered, has been discussed in Bonnini et al. (2012). In the present work a simulation study is presented to prove the good power behavior of the permutation solution in some specific situations, very common in real applications, where more than one covariate is present. Moreover a discussion on the minimum sample size needed to perform such permutation test and obtain a powerful solution is also given.

Keywords: Permutation test, CUB model, rating data, power study, covariate effect, sample size.

Speaker: Stefano Favaro

Affiliation: University of Torino, Italy

Title: On the stick-breaking representation for Gibbs-type priors

Abstract: Random probability measures are the main tool for Bayesian nonparametric inference, given their law acts as a prior distribution. Many well-known priors used in practice admit different, though (in distribution) equivalent, representations. Some of these are convenient if one wishes to thoroughly analyze the theoretical properties of the priors being used, others are more useful in terms of modeling and computation As for the latter purpose, the so-called stick-breaking constructions certainly stand out. In this talk we focus on the recently introduced class of Gibbs-type priors and provide a stick-breaking representation for it.

Heteroskedastic linear regression: steps towards adaptivity, efficiency, and robustness

Dimitris N. Politis*
Department of Mathematics -Univ. of California—San Diego
Stefanos Poulis
Department of Mathematics-Univ. of California—San Diego

Abstract

In Linear Regression with heteroscedastic errors, the OLS estimator is sub-optimal but still valid, i.e., unbiased and consistent. Halbert White, in his seminal paper (Econometrica, 1980) used the OLS residuals in order to get an estimate of the standard error of the OLS estimator that is valid under an unknown structure of the underlying heteroscedasticity. The optimal GLS estimator similarly depends on the unknown heteroscedasticity, and is thus intractable. In this paper, we introduce two different approximations to the optimal GLS estimator; the starting point for both approaches is in the spirit of White's correction, i.e., using the OLS residuals to get a rough estimate of the underlying heteroscedasticity. We show how the new estimators can benefit from the Wild Bootstrap both in terms of optimising them, but also in terms of providing valid standard errors for them despite their complicated construction. The performance of the new estimators is compared via simulations to the OLS and to (intractable) GLS.

*Research partially supported by NSF grant DMS-10-07513.

Title: Consensus Optimization via Alternating Direction Method of Multipliers

Stephen Boyd (joint work with Neal Parikh, Eric Chu, Borja Peleato, and Jon Eckstein)

Problems in areas such as machine learning and dynamic optimization on a large network lead to extremely large convex optimization problems, with problem data stored in a decentralized way, and processing elements distributed across a network.
We argue that the alternating direction method of multipliers is well suited to such problems. The method was developed in the 1970s, with roots in the 1950s, and is equivalent or closely related to many other algorithms, such as dual decomposition, the method of multipliers, Douglas-Rachford splitting, Spingarn's method of partial inverses, Dykstra's alternating projections, Bregman iterative algorithms for l_1 problems, proximal methods, and others.
After briefly surveying the theory and history of the algorithm, we discuss applications to statistical and machine learning problems such as the lasso and support vector machines, and to dynamic energy management problems arising in the smart grid.

Title: Extreme value theory with operator norming: Simulation and Statistics

Stilian Stoev, U of Michigan

We present some statistical aspects of a newly developed approach to extreme value theory with operator norming. We briefly introduce some limit theorems for the angular extremes of multivariate heavy tailed data. Operator norming allows us to handle in a unified way distributions with different tail exponents in different directions. We then present a method for simulating the limit process and a parametric bootstrap type procedure for testing for the need of operator norming. The statistical test is illustrated over simulated and real data sets.

Speaker: Subhashis Ghosal,

Affiliation: North Carolina State University

Title: Bayesian methods for clustering functional data

Abstract: Sophisticated instruments of today now allow taking essentially continuous measurements on subjects, leading to basic observations that are best viewed as functions. Clustering is about finding which functional observations are to be grouped together as similar in some appropriate sense. We expand functional observations in a wavelet basis function, and then use the coefficients in the expansion, to be called features, to represent the data in a multivariate setting. A latent variable approach using a Dirichlet process prior has become a standard technique for tieing observations a priori, but in the present context, such a tieing pattern may lead to unreasonable priors since two functional observations are unlikely to share all of their features. We design a prior to allow partial feature sharing which is analogous to the Indian buffet process. A more flexible species sampling prior that can represent the clustering process more realistically is also considered. We propose a similarity index to measure the extent of partial feature sharing between subjects and develop Markov chain Monte-Carlo methods for posterior computation.
A real data related to seizure activity is analyzed using the proposed method.
We also study some asymptotic properties related to our proposed procedure.
We further extend the proposed method to accommodate covariates and modify suitably to use in functional classification problems.

Variable selection and estimation for longitudinal survey data

Suojin Wang

Texas A&M University, USA

Abstract: There is wide interest in studying longitudinal surveys where sample subjects are observed successively over time. Longitudinal surveys have been used in many areas today, for example, in the health and social sciences, to explore relationships or to identify significant variables in regression settings. In this talk we discuss a general strategy for the model selection problem in longitudinal sample surveys. A survey weighted penalized estimating equation approach is proposed to select significant variables and estimate the coefficients simultaneously. The proposed estimators are design consistent and perform as well as the oracle procedure when the correct submodel were known. The estimating function bootstrap is applied to obtain the standard errors of the estimated parameters with good accuracy. A fast and efficient variable selection algorithm is developed to identify significant variables for complex longitudinal survey data. Simulated examples are illustrated to show the usefulness of the proposed methodology under various model settings and sampling designs. (This talk is based on joint work with Lily Wang of University of Georgia.)

Sylvain Sardy

Title: Combining information from several concomitant captors to detect gravitational wavebursts

Abstract: we consider wavelet-based nonparametric estimation of gravitational wavebursts from several time series with poor signal to noise ratio. We aggregate information by block thresholding wavelet coefficients levelwise between time series. The corresponding estimator can be viewed as a smooth generalization of the James-Stein estimator, which regularization parameters are selected by minimizing an unbiased estimate of the risk. We show with real data how block thresholding helps detecting wavebursts. We also show how the nonparametric estimator can be applied to the more complex setting of linear inverse problems.

Econometric Inference in the Vicinity of Unity.

Tassos Magdalinos
University of Southampton, UK

Peter C. B. Phillips
Cowles Foundation for Research in Economics
Yale University

and
University of Auckland & University of Southampton, UK

Abstract

Present econometric methodology of inference in cointegrating regression is extended to mildly integrated time series of the type introduced by Magdalinos and Phillips (2007, 2009). It is well known that conventional approaches to estimating cointegrating regressions fail to produce even asymptotically valid inference procedures when the regressors are nearly integrated, and substantial size distortions can occur in econometric testing. The new framework developed here enables a general approach to inference that resolves this difficulty and is robust to the persistence characteristics of the regressors, making it suitable for general practical application. Mildly integrated instruments are employed, one using system regressors and internally generated instruments, the other using external instruments. These new IV techniques eliminate the endogeneity problems of conventional cointegration methods with near integrated regressors and robustify inference to uncertainty over the precise nature of the integration in the system. The use of mildly integrated instruments also provides a mechanism for linking the conventional treatment of endogeneity in simultaneous equations with the econometric methodology for cointegrated systems. The methods are easily implemented, widely applicable and help to alleviate practical concerns about the use of cointegration methodology when roots are in the vicinity of unity rather than precisely at unity.

Keywords: Central limit theory, Cointegration, Endogeneity bias, Instrumentation, Mild integration, Mixed normality, Robustness, Simultaneity.

JEL classification: C22

The Alpha-Procedure – a non-parametric invariant method for solving tasks of reconstruction of functional dependencies and pattern recognition

Tatjana Lange, University of Applied Sciences Merseburg, Geusaer Straße, 06217 Merseburg

Abstract: The presentation deals with a common method for solving tasks of both reconstructing functional dependencies and pattern recognition. It focuses on the description of the Alpha-procedure which is based on a geometric representation of the separation of two classes by a hyperplane within a *d*-dimensional rectifying feature space. The needed dimension of the space, i.e. the number of features that is necessary for a successful classification, is gained step by step using a 2-dimensional *repère* (frame of vector space). The supplement of the feature set is performed depending on the values of the functions describing the discriminating power of both the feature and the *repère*. The transformation of the vectors (i.e. class' objects) within the two-dimensional *repère* is done towards the growth of value of the discriminating power while the invariant is preserved.

Here, the invariant is the object's affiliation with a class.

The result of the *repère's* transformation builds a fictive feature. Now, a new *repère* is built using this fictive feature and as second dimension the next real feature that owns the best value of the discriminating power. The enrichment of the feature set and the transformation of the *repères* are stopped after the classes are separated.

The advantage of the Alpha-procedure is the robustness and clarity of the process separating step by step the classes using two-dimensional *repères*.

Keywords: Alpha-procedure, reconstruction of functional dependencies, pattern recognition, supervised learning, repère, invariant

References:

1. V.I. Vasil'ev. The reduction principle in pattern recognition learning (PRL) problem. Pattern Recognition and Image Analysis 1, 1 (1991).
2. V.I. Vasil'ev. The reduction principle in problems of revealing regularities I. Cybernetics and Systems Analysis 39, 686-694 (2003).
3. V.I. Vasil'ev and T. Lange. The duality principle in learning for pattern recognition (in Russian). Kibernetika i Vytschislit'elnaya Technika 121, 7-16 (1998).
4. T. Lange and V.I. Vasil'ev. A Cluster-Analytical Approach to the Modeling of Non-Linear Systems. Operations Research Proceedings, Papers of the 25th Annual Meeting of DGOR, 1996, Springer-Verlag, (1997)

Tatyana Krivobokova

Smoothing parameter selection in two frameworks for penalized splines.

In contrast to other nonparameteric regression techniques, smoothing parameter selection for spline estimators can be performed not only by employing criteria that approximate the average mean squared error (e.g. generalized cross validation), but also by making use of the maximum likelihood (or empirical Bayes) paradigm. In the later case, the function to be estimated is assumed to be a realization of some stochastic process, rather than from a certain class of smooth functions. Under this assumption both smoothing parameter selectors for spline estimators are well-studied and known to perform similar. A more interesting problem is the properties of smoothing parameter estimators in the frequentist framework, that is if the underlying function is non-random. In this talk we discuss the asymptotic properties of both smoothing parameter selection criteria for general low-rank (or so-called penalized) spline smoothers in the frequentist framework and give also insights into their small sample performance.

Optimal Rank-Based Testing for Common Principal Components

M. Hallin*†, D. Paindaveine* and T. Verdebout‡

*ECARES, Université libre de Bruxelles
†ORFE, Princeton University
‡Université Lille 3

We provide optimal testing procedures for the null hypothesis of Common Principal Components (CPC). We first show that the pseudo-Gaussian test previously introduced in Hallin et al. (2008a) is indeed locally and asymptotically optimal, in the Le Cam sense, under Gaussian densities. Although asymptotically valid under non-Gaussian and possibly heterokurtic elliptical families with finite fourth-order moments, this test, which involves the estimation of kurtoses, remains poorly robust. Moreover, it is highly desirable for principal component methods based on scatter matrices to extend to arbitrary elliptical populations, irrespective of any moment assumptions. We therefore propose rank-based procedures that remain valid under any possibly heterokurtic m-tuple of elliptical densities (without any moment assumptions). In the homogeneous case, the normal-score version of our signed-rank tests uniformly dominate, in the Pitman sense, the optimal pseudo-Gaussian test. The results are obtained via a nonstandard application of Le Cam's LAN methodology in the context of curved statistical experiments.

Bootstrap for the first order Random Coefficient Autoregressive Model

Thorsten Fink (joined work with Jens-Peter Kreiss)
TU Braunschweig

We consider autoregressive processes of order one with a random coefficient and develop two bootstrap approaches that work for the distribution of the estimated autoregressive parameter as well as for the distribution of the estimated variances of the innovation noise and of the stochastic parameter.

For a bootstrap procedure, we usually need to have i.i.d. random variables, but we do not have these directly when observing a (random coefficient) autoregressive process. We first propose a wild bootstrap that uses the estimated densities of the innovations and the disturbance noise of the stochastic parameter to generate a bootstrap replicate of the process. Thereafter, we show how to obtain approximative residuals for the process and to separate between the innovation and the disturbance noise even though the standard method for autoregressive processes does not work in this context since one then would obtain convoluted residuals of the innovation and disturbance noises. This enhances the classical residual bootstrap for autoregressive processes. The consistency of the bootstrap approaches is established and their performance is illustrated by a simulation study.

Keywords: Random Coefficient Autoregressive Model, Deconvolution, Wild Bootstrap, Residual Bootstrap

Tim McMurry

Title: Robust Empirical Bayes With Application to Genome-Wide Association Studies

Abstract:
Large scale technologies such as gene expression microarrays and genome-wide association studies measure a large number of parallel parameters on a usually much smaller number of subjects. Bayesian and empirical Bayes analyses are natural for large scale data because of their ability to infer the collective structure of the many underlying parameters and to borrow information from the other observations. This talk proposes a rank-conditioned procedure in which the inference is based on the conditional distribution of the error given the rank of the (raw) estimate among all other estimates as opposed to conditioning on the raw estimate itself. Our method is particularly suited for correcting ranking bias in large scale estimation and for constructing valid confidence intervals for selected top-ranked parameters. The new method is almost as efficient as the corresponding Bayesian analysis when the prior is correctly specified. When the prior is incorrectly specified, the new method can be much more robust in the sense that it continues to provide accurate point and interval estimates.

Estimation of a density using real and artificial data *

Luc Devroye[1], Tina Felber[2], and Michael Kohler[2,†]

[1] *School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6, email: lucdevroye@gmail.com*

[2] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: tfelber@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

January 5, 2012

Abstract

Let $X, X_1, X_2, \ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Given a sample of the distribution of (X, Y) and additional independent observations of X we are interested in estimating f. We apply a regression estimate to the sample of (X, Y) and use this estimate to generate additional artificial observations of Y. Using these artificial observations together with the real observations of Y we construct a density estimate of f by using a convex combination of two kernel density estimates. It is shown that if the bandwidths satisfy the usual conditions and if in addition the supremum norm error of the regression estimate converges almost surely faster towards zero than the bandwidth of the kernel density estimate applied to the artificial data, then the convex combination of the two density estimates is L_1–consistent. The performance of the estimate for finite sample size is illustrated by simulated data, and the usefulness of the procedure is demonstrated by applying it to a density estimation problem in a simulation model.

AMS classification: Primary 62G07; secondary 62G20.

Key words and phrases: Density estimation, L_1–error, nonparametric regression, consistency.

*Running title: *Density estimation using real and artificial data*

†Corresponding author. Tel: +49-6151-16-6846

Bootstrap for a class of discretely observed continuous time series

Tobias Niebuhr
Technische Universität Braunschweig, Institut für Mathematische Stochastik, Germany.
e-mail: t.niebuhr@tu-bs.de

We develop a bootstrap proposal for Lévy-driven continuous-time autoregressive (CAR) processes sampled in discrete time. It is well-known that a sample of an Ornstein-Uhlenbeck process, that is a CAR process of order one, has a discrete-time autoregressive representation with i.i.d. noise. Based on this representation a simple bootstrap approach can be found. Since higher order CAR samples lose this nice representation as a linear autoregressive process and instead only obey an autoregressive representation with dependent noise, a more general bootstrap proposal has to be developed. We consider statistics depending on a fixed grid and use the solution of the stochastic differential equation (SDE) that corresponds to the original CAR process. Based on discrete-time observations and some auxiliary observations on a finer grid, which lead to approximations of the derivatives of the continuous time process and which, at least asymptotically, possess a multivariate autoregressive representation with i.i.d. residuals, a valid residual based bootstrap can be defined which allows to imitate the CAR process on the underlying discrete time grid. We show that this approach is consistent for empirical autocovariances and autocorrelations of CAR processes.
This is joint work with Peter J. Brockwell and Jens-Peter Kreiß.

Maximal variances of order statistics

Tomasz Rychlik

Institute of Mathematics
Polish Academy of Sciences
Chopina 12, 87100 Toruń, Poland
e-mail: trychlik@impan.gov.pl

Abstract

Optimal evaluations of variances of order statistics expressed in terms of variances of parent random variables were derived by Papadatos (1995, 1997) for independent identically distributed samples from the general and symmetric populations, respectively. We deliver a more precise solution in the symmetric case. Also, we present solutions to the similar problems for arbitrarily dependent identically distributed random variables. The sharp bounds for the order statistics from dependent samples are identical for numerous dispersion measures connected with the notion of M-functionals of location.

Title: Matrix Completion and Large-Scale SVD Computations

Trevor Hastie, Stanford University

The Singular Value Decomposition (SVD) is a fundamental tool in all branches of data analysis - arguably one of the most widely used numerical tools. Over the last few years, partly inspired by the "Netflix problem", the SVD has again come into focus as a solution to the "matrix completion" problem. One partially observes a very large matrix, and would like to impute the values not observed. By assuming a low-rank structure, the SVD is one approach to the problem - a SVD with large amounts of missing data. In this talk we discuss an approach for building a path of solutions of increasing rank via nuclear-norm regularization. An integral part of this algorithm involves repeatedly computing low-rank SVDs of imputed matrices. We show how these tasks can be efficiently handled by parallel computational algorithms, allowing the method to scale to very high-dimensional problems.

Complete case estimators for semiparametric regression models

Ursula U. Müller[1]
Texas A&M University

uschi@stat.tamu.edu
http://www.stat.tamu.edu/∼uschi

The fastest and simplest method of dealing with missing data is listwise deletion, i.e. using only cases that are completely observed. It is well known, though, that a statistical analysis based on those data does not always perform well. Approaches which impute missing values often give better results. However, we can identify situations where a complete case analysis is appropriate and sometimes even optimal.

I present a general method for obtaining limiting distributions of complete case statistics for missing data models from those of the corresponding statistics when all data are observed. This provides a convenient method of adapting established methods without (reproducing) lengthy proofs.

The methodology is illustrated by analyzing inference procedures for partially linear regression models with responses missing at random: we derive asymptotically efficient estimators of the slope parameter and present an asymptotically distribution free test for fitting a normal distribution to the errors.

This talk is based on joint work with Hira L. Koul and Anton Schick.

References

[1] H.L. Koul, U.U. Müller and A. Schick (2012). Complete case analysis revisited.
`http://www.stat.tamu.edu/∼uschi/research/cca.pdf`

[2] U.U. Müller (2009). Estimating linear functionals in nonlinear regression with responses missing at random, *Ann. Statist.*, 37, 2245-2277.

[3] U.U. Müller, A. Schick and W. Wefelmeyer (2006). Imputing responses that are not missing, *Probability, Statistics and Modelling in Public Health* (M. Nikulin, D. Commenges and C. Huber, eds.), 350-363, Springer.

[4] U.U. Müller, A. Schick and W. Wefelmeyer (2012). Estimating the error distribution function in semiparametric additive regression models. *J. Statist. Plann. Inference*, 142, 552-566.

Prepared on February 22, 2012

[1]Ursula U. Müller was supported by the US National Science Foundation, grant DMS-0907014.

Copulas and covariates

Noël Veraverbeke
Universiteit Hasselt, Belgium
noel.veraverbeke@uhasselt.be

Studying the relationship between two (or more) random variables in the presence of a covariate can be done based on a conditional version of Sklar's theorem: there exists a copula function expressing the joint conditional distribution as a function of the one dimensional conditional marginal distributions. We discuss recent results on several estimators for this unknown copula function. First of all there is the nonparametric method which uses empirical estimators with weights that smooth over the covariate space. An application is the asymptotic theory for association measures like the conditional Kendall's tau [2,4]. A second method is semi-parametric in nature: it starts from a parametric family of copulas in which the parameter depends on the covariate. This parameter function is estimated by local likelihood [1]. A third method provides a smooth estimator by the use of Bernstein polynomials [3].

References

[1] F. Abegaz, I. Gijbels and N. Veraverbeke, Semi-parametric estimation of conditional copulas. *J. Multivariate Analysis* (to appear) 2012.

[2] I. Gijbels, N. Veraverbeke and M. Omelka, Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis*, **55**, 1919-1932, 2011.

[3] P. Janssen, J. Swanepoel and N. Veraverbeke, Large sample behavior of the Bernstein copula estimator. *J. Statist. Planning and Inference*, **142**, 1189-1197, 2012.

[4] N. Veraverbeke, M. Omelka and I. Gijbels, Estimation of a conditional copula and association measures. *Scandinavian J. Statistics*, **38**, 766-780, 2011.

Towards Frequency Domain Analysis of Stationary Functional Data

Victor M. Panaretos

Ecole Polytechnique Fédérale de Lausanne (EPFL)

We consider the problem of drawing statistical inferences on the second-order structure of weakly dependent functional time series. When functional data are independent, the entire second order structure is captured by the covariance operator. For dependent functional data, one needs to consider covariance operators relating different lags of the series, as is the case in multivariate time series. In the functional case, most work has focused on inference for stationary time series that are linear - a case that is now well understood. More recent work has focused on the estimation of the mean, long-run covariance operator and principal components on the basis of moment type mixing conditions for time series that are not necessarily linear. In this talk we consider the problem of inferring the complete second-order structure of stationary functional time series without a priori structural modeling assumptions. Our approach is to formulate a frequency domain framework for weakly dependent functional data, employing suitable generalisations of finite-dimensional notions. We introduce the basic ingredients of such a framework, propose estimators, and study their asymptotics under functional cumulant-type mixing conditions. (Based on joint work with Shahin Tavakoli, EPFL).

CLT on one dimensional stratified spaces

This is a talk in the invited session on
Nonparametric Statistics on Manifolds and their Applications.

Abstract.
Recent investigations showed that object data analysis
can be characterized as data analysis on a metric space
with a manifold stratification. Many of these were
initiated by a Working Group (http://www.samsi.info/working-groups/data-analysis-sample-spaces-manifold-stratification)
at the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2010/2011
program on analysis of object data, and are continued at a
2012 workshop at the Mathematical Biosciences Institute
(http://www.mbi.osu.edu/2011/pgcdescription.html)
The asymptotic behavior of the Frechet sample means from
a probability distribution with a Frechet mean at a
regular point on such a sample space was essentially established by
Bhattacharya and Patrangenaru (2006); however little is
known, if the Frechet mean is on its singular part. In this
talk we will investigate the asymptotics of the Frechet
sample means on a one dimensional stratified space.

Non-parametric adaptive estimation by dependent observations

Vyacheslav A. Vasiliev
Tomsk State University, Russia

This talk presents the truncated estimation method with applications for non-parametric and parametric problems.

The main aim is to obtain estimators with a given accuracy by fixed sample size.

We consider two problems in detail:

– estimation of the regression function using the sequential truncated estimation method (joint results with D.Politis);

– investigation of ratio type estimators using the general truncated estimation method.

In the first problem observed inputs and noises of the regression model are supposed to be dependent and form sequences of dependent numbers. Two types of estimators are considered. Both estimators are constructed on the base of the Nadaraya–Watson kernel estimator.

First, the sequential estimators with a given bias and mean square error are defined. According to sequential approach the duration of observations is defined as a special stopping time. Then on the basis of these estimators, truncated sequential estimators are constructed. They are defined on the time interval of a fixed length. At the same time the variance of these estimators is also known, and both estimators have optimal (as compared to the case of independent inputs) rates of convergency.

In the second problem, the general truncated estimation method for investigation of basic ratio type estimators constructed by dependent sample of finite size is presented. This method gives a possibility to obtain estimators with guaranteed accuracy in the sense of L_m-norm, $m \geq 2$. As an illustration, parametric and non-parametric estimation problems on a time interval of a fixed length are considered. In particular, parameters of linear (autoregressive) and non-linear (ARARCH) discrete-time processes are estimated. Moreover, parameter estimation problem of the non-Gaussian Ornstein-Uhlenbeck process by discrete-time observations and the estimation problem of a logarithmic derivative of a noise density of an autoregressive process with guaranteed accuracy are solved.

It is shown, in addition, that all parametric truncated estimators have rates of convergence of basic estimators. In particular, non-parametric estimator above has optimal rate of convergence (as compared to the case of independent inputs).

Wei Biao Wu

Testing parametric assumptions of trends of nonstationary time series

I will discuss testing whether the mean trend of a nonstationary time series is of certain parametric forms. A central limit theorem for the integrated squared error is derived, and with that a hypothesis-testing procedure is proposed. The method is illustrated in a simulation study, and is applied to assess the mean pattern of lifetime-maximum wind speeds of global tropical cyclones from 1981 to 2006. I will also revisit the trend pattern in the central England temperature series. The work is joint with Ting Zhang.

Statistical Inference for Directional Data

Authors: Graciela Boente[(2)], Rosa M. Crujeiras[(1)], María Oliveira[(1)], Daniela Rodriguez[(2)], Alberto Rodríguez-Casal[(1)] and **Wenceslao Gonzalez-Manteiga**[(1)].

[(1)] University of Santiago de Compostela (Spain)
[(2)] Universidad de Buenos Aires – CONICET (Argentina)

Abstract

Two inferential problems on a directional density are considered in this talk, specifically, nonparametric estimation of the density and testing the adequacy of a parametric family.

In the particular case of circular data, nonparametric kernel density estimation is revised, with special attention to the bandwidth selection problem. A new plug—in rule is introduced, and its performance is illustrated by a simulation study. For the more general case of directional data, the problem of testing if the density belongs to a certain parametric family is considered. A bootstrap procedure to approximate the distribution of the test statistic is developed and its performance under the null and different alternatives is explored through a simulation study. Both techniques are illustrated with real data.

Bootstrap Methods for Lasso-Type Estimators Under A Moving-Parameter Framework

Wenlong Cai and Stephen M.S. Lee
The University of Hong Kong

Abstract

We study the distributions of Lasso-type regression estimators in a moving-parameter asymptotic framework, and consider various bootstrap methods for estimating them accordingly. We show, in particular, that the distribution functions of Lasso-type estimators, including even those possessing the oracle properties such as the adaptive Lasso and the SCAD, cannot be consistently estimated by the bootstraps uniformly over the space of the regression parameters, especially when some of the regression coefficients lie close to the origin. Such lack of uniform consistency poses difficulties in practical applications of the bootstraps for making Lasso-based inferences. In the light of this seemingly negative result, we seek, however, to develop criteria for assessing the relative risks, phrased in terms of their uniform consistency properties, of the various bootstrap methods, based on which an optimal bootstrap strategy may be formulated in an adaptive manner. A simulation study is provided to demonstrate the non-normal nature of the distributions of Lasso-type estimators, and to assess the performances of various bootstrap estimates of such distributions across different values of regression parameters.

A Semiparametric Threshold Model for Censored Longitudinal Data Analysis

Wenyang Zhang
Department of Mathematics
The University of York, UK

May 14, 2012

Abstract

Motivated by an investigation of the relationship between blood pressure change and progression of microalbuminuria (MA) among individuals with Type I diabetes, we propose a new semiparametric threshold model for censored longitudinal data analysis. We also study a new semiparametric BIC-type criterion for identifying the parametric component of the proposed model. Cluster effects in the model are implemented as unknown fixed effects. Asymptotic properties are established for the proposed estimators. A quadratic approximation used to implement the estimation procedure renders the method very easy to implement by avoiding the computation of multiple integrals and the need for iterative algorithms. Simulation studies show that the proposed methods work well in practice. An illustration using the Wisconsin Diabetes data suggests some interesting findings.

The Approximate F-Test Under Random Censorship

Winfried Stute
University of Giessen

In the linear regression model the F-test constitutes a classical approach to test for linear hypotheses among the parameters. In survival analysis the data are often subject to censorship so that the full information required to perform the F-test is not available. It is the aim of my talk to discuss a modification of the F-test under censored regression.

Speaker: Wolfgang Polonik

Title: Estimation of filamentary structure

Abstract: In dimension 2, a filament is a curve in the plane. A filamentary structure is a collection of filaments that can intersect. In practice such filamentary structures are observed for instance in medical images of blood vessels, remote sensing (road detection), or in the locations of galaxies (cosmic web).
In this talk we will discuss procedures for estimating filamentary structures. Our definition of a filament is that of a ridge line corresponding to a probability density. We will use integral curves to characterize points on a filament and use estimates of these integral curves to construct estimates of the underlying filamentary structure. Two different estimation procedures will be discussed. One of them is supported by theoretical results involving suprema of Gaussian processes on manifolds. The methods are illustrated by an application to a galaxy data set. This is joint work with Wanli Qiao.

SIEVE INFERENCE ON SEMI-NONPARAMETRIC TIME SERIES MODELS

By

Xiaohong Chen, Zhipeng Liao and Yixiao Sun

February 2012

COWLES FOUNDATION DISCUSSION PAPER NO. 1849

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Sieve Inference on Semi-nonparametric Time Series Models*

Xiaohong Chen†, Zhipeng Liao‡, and Yixiao Sun§

First Draft: October 2009; This draft: February 2012

Abstract

The method of sieves has been widely used in estimating semiparametric and nonparametric models. In this paper, we first provide a general theory on the asymptotic normality of plug-in sieve M estimators of possibly irregular functionals of semi/nonparametric time series models. Next, we establish a surprising result that the asymptotic variances of plug-in sieve M estimators of irregular (i.e., slower than root-T estimable) functionals do not depend on temporal dependence. Nevertheless, ignoring the temporal dependence in small samples may not lead to accurate inference. We then propose an easy-to-compute and more accurate inference procedure based on a "pre-asymptotic" sieve variance estimator that captures temporal dependence. We construct a "pre-asymptotic" Wald statistic using an orthonormal series long run variance (OS-LRV) estimator. For sieve M estimators of both regular (i.e., root-T estimable) and irregular functionals, a scaled "pre-asymptotic" Wald statistic is asymptotically F distributed when the series number of terms in the OS-LRV estimator is held fixed. Simulations indicate that our scaled "pre-asymptotic" Wald test with F critical values has more accurate size in finite samples than the usual Wald test with chi-square critical values.

Keywords: Weak Dependence; Sieve M Estimation; Sieve Riesz Representor; Irregular Functional; Misspecification; Pre-asymptotic Variance; Orthogonal Series Long Run Variance Estimation; F Distribution

*We acknowledge useful comments from T. Christensen, J. Hahn, L. Hansen, J. Hidalgo, M. Jansson, D. Kaplan, O. Linton, P. Phillips, D. Pouzo, J. Powell, H. White, and other participants at 2011 Econometrics Society Australasian Meeting in Adelaide, the Pre-conference Workshop of 2011 Asian Meeting of the Econometric Society in Seoul, econometrics workshops at Yale, UC Berkeley, UCLA, UCSD and Stanford. Chen and Sun acknowledge financial support from the National Science Foundation under the respective grant numbers: SES-0838161 and SES-0752443. Any errors are the responsibility of the authors.

†Department of Economics, Yale University, 30 Hillhouse, Box 208281, New Haven, CT 06520. Email: xiaohong.chen@yale.edu

‡Department of Economics, UC Los Angeles, 8379 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095. Email: zhipeng.liao@econ.ucla.edu

§Department of Economics, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508. Email: yisun@ucsd.edu

Adaptive GAM models for Day-Ahead and Intra-Day Electricity Consumption Forecasts

Yannig Goude
ElectricitØde France R&D service

Generalized Additive Models have been investigated recently to forecasts day-ahead electricity consumptions at EDF R&D. These models achieve an interesting trade-off between accuracy of forecasts and adaptation to different data sets thanks to their semi-parametric structures. We propose here a new method based on QR decomposition to learn these models on-line as we receive new data. This allows GAM models to react to smooth changes in the data generation process: economic crisis, loss or gain of customers EUR*ƒ*We illustrate it on different data sets and real forecasts.

Yichao Wu
Department of Statistics, North Carolina State University.
Email: wu@stat.ncsu.edu

Title: Two Dimensional Solution Surface of Weighted Support Vector Machine

Abstract: The support vector machine (SVM) is one of the most popular tools for binary classification. The weighted SVM is considered as its natural extension by imposing a weight on each class (rather than individual). We show in thisarticle the joint piecewise-linearity of the solutions in the weighted SVM forthe regularization parameter and the weight parameter, respectively denoted by λ and π . An efficient algorithm to obtain the entire solution surfaces is proposed by taking advantages of its joint piecewise-linearity. As every solutions for arbitrary λ and π are completely obtained, we believe that there are huge potential applicability for both theoretical and practical applications. We applied our proposed method to tune λ for the probability estimation method originally developed by Wang et al. (2008). The numerical studies via both simulated and real data sets show that the proposed adaptive grid obtained from the entire two dimensional solution surfaces significantly improves the performance of the probability estimator.

This is a joint work with Seung Jun Shin and Hao Helen Zhang.

Title:
Joint statistical modeling of multiple high dimensional data

Yufeng Liu, University of North Carolina at Chapel Hill

Abstract:
With the abundance of high dimensional data, shrinkage techniques are very popular for simultaneous variable selection and estimation. In this talk, I will present some new shrinkage techniques for joint analysis of multiple high dimensional data. Applications on cancer gene expression data and micro-RNA data will be presented.

Bayesian Analysis of Moment Condition Models Using Nonparametric Priors

Yuichi Kitamura and Taisuke Otsu

Yale University

Abstract

This paper develops semiparametric Bayesian methods for moment condition models. Recent methodologies in nonparametric and semiparametric Bayesian analysis play a key role. A moment condition model possesses a feature that poses a significant challenge in application of semiparametric Bayesian methods, due to parameter restrictions induced by the model. A new method is proposed to address this problem, and a convenient algorithm that is a variant of the Metropolis-Hastings method is developed to implement it. Also, a semiparametric Bernstein-von Mises theorem is obtained; that is, the posterior distribution of the finite dimensional parameter of the model is shown to converge to the asymptotic distribution of semiparametrically efficient estimators. Preliminary numerical experiments indicate that the new algorithm is surprisingly efficient, dispite the high dimensional nature of the problem.

Regularized same-realization prediction for time series

Yulia R. Gel
University of Waterloo, Canada

Abstract

In this talk we discuss how regularization can be employed for estimation and prediction of the same-realization of time series. In particular, we show that banding enables us to employ an approximating model of a much higher order than typically suggested by AIC, while controlling how many parameters are to be estimated precisely and the level of accuracy. We present results on asymptotic consistency of banded autocovariance matrices under the Frobenius norm and provide a theoretical justification on optimal band selection using cross-validation, which can be viewed as an alternative model selection criterion for the same-realization prediction. We illustrate our procedure by simulations and case studies. This is a joint work with Peter Bickel, University of California, Berkeley.

Consistency of community detection in networks under degree-corrected block models

Yunpeng Zhao, Elizaveta Levina and Ji Zhu

University of Michigan

Community detection is a fundamental problem in network analysis, with applications in many diverse areas. The stochastic block model is a common tool for model-based community detection. However, the block model is limited by its assumption that all nodes within a community are stochastically equivalent, and provides a poor fit to networks with hubs or highly varying node degrees within communities, which are common in practice. The degree-corrected block model was proposed to address this shortcoming, and allows variation in node degrees within a community while preserving the overall block model community structure. In this talk, we present general theory for checking consistency of community detection under the degree-corrected block model, and compare several community detection criteria under both the standard and the degree-corrected block models. We show which criteria are consistent under which models and constraints, as well as compare their relative performance in practice.

Semiparametrically Nonlinear Time Series Modelling under Near Epoch Dependence

Zudi Lu

School of Mathematical Sciences
The University of Adelaide, Australia

Semiparametric methodologies have achieved quite a lot of success in nonlinear statistical and econometric modelling; see, for example, Fan and Gijbels (1996), Fan and Yao (2003) and Li and Racine (2007). Under various mixing stochastic processes (in particular the alpha--mixing, i.e., strong mixing, that covers many other mixings such as phi-mixing, beta-mixing as special cases), these techniques, including the popular local linear fitting, have been well studied in the literature by many researchers in time series modelling, c.f., Liebscher (1996), Masry (1996), Bosq (1998), Fan and Yao (2003), Gao (2007), Hansen (2008), Kristensen (2009), among others.

However, from a practical point of view, the mixing (e.g., alpha—mixing) processes suffer from many undesirable features. For example, for a lot of popular processes in econometrics such as an ARMA process mixed with ARCH or GARCH errors, it is still difficult to show whether they are alpha--mixing or not except in some very special cases. Even for a very simple linear AR(1) model with innovation being independent symmetric Bernoulli random variables taking on values of -1 and 1, the stationary solution to the model is not alpha--mixing (Andrew 1984).

In this talk, I will first review some of the extensions of the stochastic processes beyond the mixings, in particular a class of generalised stable processes from mixings, or called near epoch dependence, which covers a variety of interesting stochastic processes in time series econometric modelling. I will then report some recent developments on the local linear fitting and semiparametric model averaging techniques that my co-authors and I have made under this kind of near epoch dependent processes. The results obtained include the pointwise asymptotic distributions for the probability density estimation and the local linear estimator of a nonparametric time series regression as well as their uniformly strong and weak consistencies with convergence rates under near epoch dependence. These results are showed to be useful for the study of a proposed semiparametric model averaging method with its empirical application to an annual mean temperature anomaly series presented.