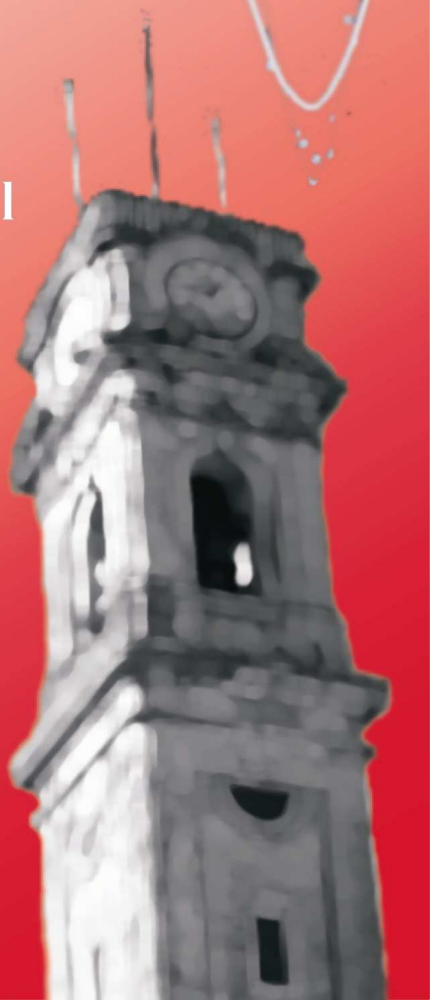
Workshop on

Nonparametric Inference - WNI2008

University of Coimbra, Portugal

June 26-28, 2008



Abstracts

Workshop on Nonparametric Inference – WNI2008

Coimbra, June 26–28, 2008

Scientific Committee

Organizing Committee

Workshop on Nonparametric Inference – WNI2008

Coimbra, June 26–28, 2008

PROGRAM

Thursday June 26	Friday June 27	Saturday June 28
9.00–10.00 Opening & Registration	9.30–10.30 **László Györfi**	9.30–10.30 **Emmanuel Candès**
10.00–11.00 **Philippe Vieu**	10.30–11.00 Coffee break	10.30–11.00 Coffee break
11.00–11.30 Coffee break	11.00–12.40 Contributed sessions 3A, 3B	11.00–12.40 Contributed session 5
11.30–12.30 **Antonio Cuevas**		
	L U N C H	
14.30–16.10 Contributed session 1	14.30–16.10 Contributed sessions 4A, 4B	14.30–16.10 Contributed session 6
16.10–16.40 Coffee break	16.10–16.40 Coffee break	16.10–16.40 Coffee break
16.40–18.00 Contributed session 2	Visit to Univ. Coimbra (historical)	16.40–17.40 **Irène Gijbels**
		17.40–18.00 Closing

Welcome reception 19.30–20.30		

WNI2008 – Scientific Program

Plenary Sessions

Room Pedro Nunes

Thursday, June 26
Chairman: Paulo Oliveira
10.00–11.00
On nonparametric functional data analysis
Philippe Vieu (p. 15)

Thursday, June 26
Chairman: Carlos Tenreiro
11.30–12.30
On nonparametric estimation of boundary measures
Antonio Cuevas (p. 16)

Friday, June 27
Chairwoman: Carla Henriques
9.30–10.30
Nonparametric prediction of time series
László Györfi (p. 17)

Saturday, June 28
Chairman: Carlos Tenreiro
9.30–10.30
Computationally tractable statistical estimation when there are more variables than observations
Emmanuel Candès (p. 18)

Saturday, June 28
Chairman: Paulo Oliveira
16.40–17.40
Smoothing non equispaced heavy noisy data with wavelets kernels
Irène Gijbels (p. 19)

Contributed Sessions

Session 1

Thursday, June 26
Room Pedro Nunes

Chairman: Jacobo de Uña-Álvarez

14.30–14.50 **The shorth plot**
J.H.J. Einmahl, M. Gantner, G. Sawitzki (p. 23)

14.50–15.10 **Spatial prediction via the kernel method with cross-validation approaches for bandwidth selection**
R. Menezes, C. Ferreira, P. García-Soidán (p. 24)

15.10–15.30 **Nonparametric estimation of a log-concave density**
M. Cule, R. Samworth, R. Gramacy, M. Stewart (p. 25)

15.30–15.50 **Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices**
J.E. Chacón, T. Duong (p. 26)

15.50–16.10 **A copula-based test for independence**
N. Neumeyer, S. Kiwitt (p. 27)

Session 2

Thursday, June 26
Room Pedro Nunes

Chairman: Arne Kovac

16.40–17.00 **Advances in data depth**
C. Agostinelli, M. Romanazzi (p. 28)

17.00–17.20 **Variable selection and weighting by nearest neighbor ensembles**
J. Gertheiss, G. Tutz (p. 30)

17.20–17.40 **Peak preserving spectral density estimation**
L. Desmet, I. Gijbels (p. 31)

17.40–18.00 **The additive-interactive nonlinear volatility model, its estimation and some testing issues**
M. Levine, T. Li (p. 32)

Session 3A

Friday, June 27
Room Pedro Nunes

Chairwoman: Cecília Azevedo

11.00–11.20 **Specification tests for the distribution of errors in nonparametric regression: a martingale approach**
J. Mora, A. Pérez-Alonso (p. 33)

11.20–11.40 **Representation of the conditional distribution function of a censored variable given a linear combination of a d-vector of covariates**
M.C. Iglesias-Pérez, W. González-Manteiga (p. 34)

11.40–12.00 **A test for comparing regression curves versus one-sided alternatives**
J.C. Pardo-Fernández, N. Neumeyer (p. 36)

12.00–12.20 **Nonlinear wavelet regression for left-truncated, dependent data**
J. Uña-Álvarez, H.-Y. Liang (p.37)

12.20–12.40 **Nonparametric Mixture Regression**
A. Rojas, C. Genovese, L. Wasserman (p. 38)

Session 3B

Friday, June 27
Room 2.4

Chairwoman: Manuela Neves

11.00–11.20 **Longitudinal modeling when response and time-dependent covariates are measured at different timepoints**
J.A. Dubin, X. Xiong (p. 39)

11.20–11.40 **IRT analysis of STAIC data**
B. Oliveiros, A. Gomes da Silva, E. Ponciano (p. 40)

11.40–12.00 **A semiparametric estimator for spatial count data regression analysis**
J.A. Santos, M.M. Neves, C. Barroso, F. Costigliola, F. Maia (p. 41)

12.00–12.20 **Forecasting M3 competition data - another approach**
C. Cordeiro, M. Neves (p. 43)

12.20–12.40 **Estimation of the restricted conditional mean gap time: the induced dependent censoring aspect**
A.-C. Andrei (p. 44)

Session 4A

Friday, June 27
Room Pedro Nunes

Chairman: Michel Delecroix

14.30–14.50 **A geometric interpretation of the multiresolution criterion**
T. Mildenberger (p. 45)

14.50–15.10 **Nonparametric Bayesian inference for integrals with respect to an unknown finite measure**
T. Erhardsson (p. 46)

15.10–15.30 **Multiresolution and model choice**
A. Kovac (p. 48)

15.30–15.50 **Weak convergence of the supremum distance for supersmooth kernel deconvolution**
B. van Es, S. Gugushvili (p. 49)

15.50–16.10 **Nonparametric model checking for dynamic load sharing models**
E. Beutner (p. 50)

Session 4B

Friday, June 27
Room 2.4

Chairwoman: Natalie Neumeyer

14.30–14.50 **A SDE growth model: nonparametric estimation of the drift and the diffusion coefficients**
P.A. Filipe, C.A. Braumann (p. 51)

14.50–15.10 **Smoothing non-stationary correlated data**
P. Foster (p. 52)

15.10–15.30 **On confidence intervals for a distribution function under association**
C. Azevedo (p. 53)

15.30–15.50 **Stability estimating in optimal sequential hypotheses testing**
E. Gordienko, A. Novikov, E. Zaitseva (p. 54)

15.50–16.10 **MAP estimation for curve modeling with free-knot splines**
L. Amate, M.J. Rendas (p. 55)

Session 5

Saturday, June 28
Room Pedro Nunes

Chairman: Juan Carlos Pardo-Fernández

11.00–11.20 **Nonparametric adaptive Bayesian oracle projection estimation in the white noise model**
A. Babenko, E. Belitser (p. 57)

11.20–11.40 **B-splines regression smoothing and difference type of penalties**
I. Gijbels, A. Verhasselt (p. 58)

11.40–12.00 **Principal points and elliptical distributions from the multivariate setting to the functional case**
G. Boente, J.L. Bali (p. 59)

12.00–12.20 **Length and surface area estimation under convexity type assumptions**
B. Pateiro-López, A. Rodríguez-Casal (p. 60)

12.20–12.40 **Functional data classification based on reproducing kernel regularization**
A. Muñoz, J. González (p. 61)

Session 6

Saturday, June 28
Room Pedro Nunes

Chairman: Peter Foster

14.30–14.50 **Smoothing parameter selection in hazard rate estimation**
F. Van Graan (p. 62)

14.50–15.10 **A copula approach to conditional density estimation**
O. Faugeras (p. 63)

15.10–15.30 **A bootstrap comparison of generalized linear models against nonparametric alternatives with binomial stimulus-response data**
K. Zychaluk, D.H. Foster (p. 64)

15.30–15.50 **Nonparametric surveillance under weak dependence to detect structural changes**
A. Steland (p. 65)

15.50–16.10 **Adaptive maximum and proxy maximum probability estimation of multidimensional Poisson intensities**
J.C.S. Miranda (p. 66)

Plenary sessions

On nonparametric functional data analysis

Frederic Ferraty
Laboratoire de Statistique et Probabilités, Univ. Paul Sabatier, Toulouse, France, ferraty@cict.fr
Philippe Vieu
Laboratoire de Statistique et Probabilités, Univ. Paul Sabatier, Toulouse, France, vieu@cict.fr

Abstract. Statistics for Functional Data is a recent field of researches that was popularized by the monographes [5] and [6]. Various statistical questions have been studied with functional data, but the previous literature (see references in [1], [5] and [6]) was concentrated around *parametric* models and methods. Starting with [2] *nonparametric* models have been developed for analyzing functional variables, and the monograph [3] presents a wide scope of the literature in this field (including theoretical and applied issues). The main difficulty in developing *nonparametric* statistics for *functional variable* is to control the dimensional effects which are much more important than in standard nonparametric statistics (since functional data are realizations of infinite-dimensional random variables). This control can be made by suitable topological considerations.
The aim of this talk is to present the main ideas going with Nonparametric Functional Data Analysis. After giving precise definitions and discussing the meaning of the words *nonparameric* and *functional variables*, it will be explained how usual kernel smoothing ideas can be adapted to infinite dimensional variables. Then, some among the most important results of [3] will be exposed. The monography [3] is accompanied by a web site [4] containing S+/R routines and applications to various curve datasets. This talk will be illustrated by means of examples extracted of [4].

1. Bosq, D. (2000) *Linear processes in functions spaces. Theory and Applications*. Lecture Notes in Statistics 149, Springer-Verlag, New York.
2. Ferraty, F., Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés (in french). *Comptes Rendus Acad. Sci. Paris* 330, 403–406.
3. Ferraty, F., Vieu, P. (2006). *Nonparametric functional data analysis*. Springer Series in Statistics, New York.
4. Ferraty, F., Vieu, P. (2006). NPFDA in practice. Free access on line at *http://www.lsp.ups-tlse.fr/staph/npfda/*
5. Ramsay, J., Silverman, B. (1997). *Functional data analysis*. Springer Series in Statistics, New York.
6. Ramsay, J. and Silverman, B. (2005). *Functional data analysis* (Second edition). Springer Series in Statistics, New York.

On nonparametric estimation of boundary measures

Antonio Cuevas
Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain, antonio.cuevas@uam.es

Abstract. The measure of the boundary ∂G of a compact body $G \subset [0,1]^d$ can be expressed in terms of the Minkowski content defined by

$$L_0(G) = \lim_{\epsilon \to 0} \frac{\mu\left(B(\partial G, \epsilon)\right)}{2\epsilon}.$$

Here μ denotes the Lebesgue measure on $\mathbb{R}^d$ and $B(G,\epsilon)$ stands for the parallel set $B(G,\epsilon) = \cup_{x \in G} B(x,\epsilon)$, where $B(x,\epsilon)$ is the closed ball with center $x \in \mathbb{R}^d$ and radius $\epsilon > 0$.
This concept is less general than the $(d-1)$-dimensional Hausdorff measure of ∂G (which coincides with $L_0(G)$ in regular cases) but it is more suitable for the estimation of the boundary measure with statistical methods.
To be more concrete, our methods can be used in those cases where the sampling information consist of random observations in $[0,1]^d$ in such a way that for each observation we are able to decide whether or not it belongs to G.
In these situations a natural nonparametric estimator of $L_0(G)$ can be defined. We will present some results concerning consistency, convergence rates and asymptotic normality of such estimator.
The practical aspects of these ideas in image analysis will be also briefly commented.
Most of this talk is a summary of recent joint work with Inés Armendáriz, Ricardo Fraiman (both from Universidad de San Andrés, Buenos Aires, Argentina) and Alberto Rodríguez-Casal (Universidad de Santiago de Compostela, Spain).

1. Armendáriz, I., Cuevas, A., Fraiman, R. (2008)̇Nonparametric estimation of boundary measures and related functionals: Asymptotic results. *Submitted.*
2. Cuevas, A., Fraiman, R., Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* 35, 1031–1051.
3. Pateiro-López, B., Rodríguez-Casal, A. (2008). Length and surface area estimation under convexity-type restrictions. *Submitted.*

Nonparametric prediction of time series

László Györfi
Budapest University of Technology and Economics, Hungary,
gyorfi@szit.bme.hu

Abstract. We study the problem of sequential prediction of a real valued sequence. At each time instant $t = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome Y_t of a sequence of real numbers $Y_1, Y_2, \ldots$ with knowledge of the pasts $Y_1^{t-1} = (Y_1, \ldots, Y_{t-1})$ (where Y_1^0 denotes the empty string) and the side information vectors $X_1^t = (X_1, \ldots, X_t)$, where $X_t \in \mathbb{R}^d$. Thus, the predictor's estimate, at time t, is based on the value of X_1^t and Y_1^{t-1}. A prediction strategy is a sequence $g = \{g_t\}_{t=1}^{\infty}$ of functions

$$g_t : \left(\mathbb{R}^d\right)^t \times \mathbb{R}^{t-1} \to \mathbb{R}$$

so that the prediction formed at time t is $g_t(X_1^t, Y_1^{t-1})$.
In this paper we assume that $(X_1, Y_1), (X_2, Y_2), \ldots$ is a stationary and ergodic process. After n time instants, the *normalized cumulative prediction error* is

$$L_n(g) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} (g_t(X_1^t, Y_1^{t-1}) - Y_t)^2.$$

We show a universally consistent prediction strategy g such that for any stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$ with $\mathrm{E}Y_0^4 < \infty$,

$$\lim_{n\to\infty} L_n(g) = L^* \quad \text{almost surely,}$$

where

$$L^* \stackrel{\text{def}}{=} \mathrm{E}\big(Y_0 - \mathrm{E}Y_0 \big| X_{-\infty}^0, Y_{-\infty}^{-1}\big)^2$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $X_{-\infty}^0, Y_{-\infty}^{-1}$.
The previous prediction may result in universally consistent classification rule as follows.Here Y_i is binary valued, and the classifier formed at time t is $f_t(X_1^t, Y_1^{t-1})$. We show a classification strategy f such that for any stationary ergodic process $\{(X_n, Y_n)\}_{-\infty}^{\infty}$, the *normalized cumulative* $0-1$ *loss* converges to the minimal error probability:

$$\begin{aligned} R_n(f) &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} I_{\{f_t(X_1^t, Y_1^{t-1}) \neq Y_t\}} \\ &\to R^* \stackrel{\text{def}}{=} \mathrm{E}\Big\{\min\big(\mathrm{P}\{Y_0 = 1 | X_{-\infty}^0, Y_{-\infty}^{-1}\}, \mathrm{P}\{Y_0 = 0 | X_{-\infty}^0, Y_{-\infty}^{-1}\}\big)\Big\}, \end{aligned}$$

where $I_{\{\cdot\}}$ denotes the indicator function.

Computationally tractable statistical estimation when there are more variables than observations

Emmanuel Candès
California Institute of Technology, USA, emmanuel@acm.caltech.edu

Abstract. In many important statistical applications, the number of variables or parameters is much larger than the number of observations. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the unknown number of pixels. Examples in functional MRI and tomography immediately come to mind. Other examples of high-dimensional data in genomics, signal processing and many other fields abound. In the context of multiple linear regression for instance, this setup raises the question of whether or not it is possible to estimate a vector of parameters of size p from a vector of observations of size n when $n \ll p$, or whether it is possible to estimate the mean response reliably under the same circumstances.
This talk will survey very recent progress in this area showing that L^1-methods such as the Dantzig selector and/or the lasso enjoy remarkable statistical properties. For instance, we will show that under reasonable sparsity assumptions, the Dantzig selector achieves an accuracy which nearly equals that one would achieve with an oracle that would supply perfect information about which coordinates of the unknown parameter vector are nonzero and which were above the noise level. This is connected with the important model selection problem since we will show that one can effectively tune L^1-based methods as to automatically select the subset of covariates with nearly the best predictive power, by solving convenient optimization programs. We will discuss a few engineering applications where this could have a large pay-off.

Smoothing non equispaced heavy noisy data with wavelets kernels

Anestis Antoniadis
Laboratoire de Modélisation et Calcul (LMC), IMAG, Université de Grenoble I, France, Anestis.Antoniadis@imag.fr
Irène Gijbels
Katholieke Universiteit Leuven, Belgium, irene.gijbels@wis.kuleuven.be
Jean-Michel Poggi
Laboratoire de Mathématiques, Université de Paris XI, France, jean-michel.poggi@math.u-psud.fr

Abstract. We consider a nonparametric noisy data regression model where the unknown regression function is assumed to belong to a wide range of function classes (including discontinuous functions). The distribution of the noise is assumed to be unknown and satisfying some weak conditions. We allow for error distributions that may have heavy tails, so that, for example, no moments of the noise exist. The error is assumed to have zero median. The design points are assumed to be deterministic points, not necessarily equispaced. Since the functions can be nonsmooth and the noise may have heavy tails, traditional estimation methods cannot be applied directly in this situation. We first use local medians to construct variables that are structured as a Gaussian nonparametric regression, as in Brown et al. (2008). The difference here is though that the resulting data are not equispaced. We therefore rely on a wavelet block penalizing procedure (see Amato et al. (2006)) adapted to non equidistant designs to construct an estimator of the regression function. Under mild assumptions on the design it is shown that the estimator simultaneously attains the optimal rate of convergence over a wide range of Besov classes, without prior knowledge of the smoothness of the underlying functions or prior knowledge of the error distribution. The performances of the procedure is illustrated via a simulation study covering a broad variety of settings. Applications to real data examples are also given.

1. Amato, U., Antoniadis, A., Pensky, M. (2006). Wavelet kernel penalized estimation for non-equispaced design regression. *Stat. Comp.* 16, 37-56.
2. Brown, L.D., Cai, T., Zhou, H. (2008). Robust nonparametric estimation via wavelet median regression. *Ann. Statist*, to appear.

Contributed sessions

The shorth plot

John H.J. Einmahl
Tilburg University, Netherlands, j.h.j.einmahl@uvt.nl
Maria Gantner
Tilburg University, Netherlands, m.gantner@uvt.nl
Günther Sawitzki
University of Heidelberg, Germany, gs@stablab.uni-heidelberg.de

Abstract. For a probability measure P on $\mathbb{R}$, the length of the shorth at a point $x \in \mathbb{R}$ and for a coverage level $\alpha \in (0,1)$ is defined as

$$S_\alpha(x) = \inf\{|I| : P(I) \geq \alpha, I \in \mathbb{I}_x\},$$

where $\mathbb{I}_x$ is the class of closed intervals which contain $x \in \mathbb{R}$. Then the shorth plot is defined as the graph of the function

$$x \mapsto S_\alpha(x), \quad x \in \mathbb{R}$$

for (all or) a selection of coverages α. The empirical shorth plot is the graph of

$$x \mapsto S_{n,\alpha}(x), \quad x \in \mathbb{R},$$

where $S_{n,\alpha}(x)$ is the empirical counterpart of $S_\alpha(x)$.
The shorth plot is a tool to investigate probability mass concentration. Because of its monotonicity in α, different choices of α can be plotted in one picture and give a good overview over local as well as global features of the probability distribution simultaneously. Compared to nonparametric kernel density estimators, the shorth has the advantage of avoiding bandwidth selection problems. Its easy computation makes it a valuable tool for a first exploration of the data.
Under weak assumptions, we can show that the rate of convergence of the localized empirical length of the shorth to the theoretical length is $n^{-1/2}$, uniformly in α and the point of localization x.
Some real-data examples as well as a sketch of the proof will be shown.

1. Einmahl, J.H.J., Gantner, M. and Sawitzki, G. (2008), "The Shorth Plot", *CentER Discussion Paper*, 2008-24, Tilburg University.
2. Einmahl, J.H.J., and Mason, D.M. (1992), "Generalized Quantile Processes", *The Annals of Statistics*, 20, 1062–1078.
3. Sawitzki, G. (1994), Diagnostic Plots for One-Dimensional Data, In: R. Ostermann and P. Dirschedl, ed., "Computational Statistics, 25th Conference on Statistical Computing at Schloss Reisensburg". Physica-Verlag/Springer, Heidelberg, 237–258.

Spatial prediction via the kernel method with cross-validation approaches for bandwidth selection

Raquel Menezes
University of Minho, Portugal, rmenezes@mct.uminho.pt
Célia Ferreira
University of Minho, Portugal, celia40991@yahoo.com.br
Pilar García-Soidán
University of Vigo, Spain, pgarcia@uvigo.es

Abstract. In this work, a nonparametric predictor will be considered, based on the kernel method, which will be applied to the stochastic processes where a random design is assumed for choice of the spatial locations.
The kriging techniques are typically used for the latter purpose, providing us with predictors that are optimal in some sense. In fact, the referred approaches are derived by minimizing the mean-squared prediction error, subject to some constraints that are dependent on the hypotheses assumed from the random process.
However, the results of the kriging equations rely on the validity of the conditions required, so that a failure in the hypotheses may have a significative effect. For instance, misspecification of the distribution, the mean or the second-order structure may lead to poor predictions.
Taking the above in mind, an alternative may be obtained via the kernel method, which will be proved to be valid under rather general conditions. In particular, the nonparametric kernel predictor satisfies that the mean-squared prediction error tends to be negligible, as the sample size increases. In addition, an adequate estimation of the bandwidth could be achieved by asymptotically minimizing the corresponding error.
The use of the optimal bandwidth in practice demands estimation on unknown quantities, dependent on the first and second order structures of the random process. Implementation of the latter approximations in an accurate way often turns out to be difficult. Hence, alternative cross-validation approaches will be provided for selection of the bandwidth, more easily attainable for a given data set.
Finally, we will describe some numerical studies carried out in order to analyze the performance of the nonparametric predictor, when adopting different selections of the bandwidth, which will be compared with the results achieved by kriging predictors, for gaussian and non-gaussian data. An application of the proposed predictor to a real data set is also included.

Nonparametric estimation of a log-concave density

Madeleine Cule
University of Cambridge, UK, mlc40@cam.ac.uk
Richard Samworth
University of Cambridge, UK, rjs57@cam.ac.uk
Robert Gramacy
University of Cambridge, UK, bobby@statslab.cam.ac.uk
Michael Stewart
University of Sydney, Australia, michaels@maths.usyd.edu.au

Abstract. We will show that, given an i.i.d. sample $X_1, \ldots, X_n$ in $\mathbb{R}^d$ from a distribution with log-concave density f, a unique nonparametric maximum likelihood estimator of f exists. We will see that a simple reformulation allows us to compute the estimator using non-differentiable convex optimization techniques. Unlike kernel density estimation (where the choice of bandwidth can be critical), this is a fully automatic procedure, and no additional tuning parameters are required. The method will be illustrated with simulated data in one and two dimensions, and we will briefly consider asymptotic performance. Finally, an example application to clustering (using the EM aglorithm) of breast cancer data will be discussed.

Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices

J. E. Chacón
Departamento de Matemáticas, Universidad de Extremadura, Spain, jechacon@unex.es

T. Duong
Imaging and Modeling Group, Institut Pasteur, Paris, France, tduong@pasteur.fr

Abstract. Multivariate kernel density estimation is an important technique in statistical exploratory data analysis. Its utility relies on its ease of interpretation, especially by graphical means. The crucial factor which determines the performance of kernel density estimation is the bandwidth matrix. Research in finding optimal bandwidth matrices began with restricted parameterizations of the bandwidth matrix which mimicked univariate selectors. Progressively these restrictions were relaxed to develop more flexible selectors. We propose a plug-in bandwidth selector with unconstrained parametrizations of both the final and pilot selectors. We quantify its asymptotic and finite sample properties. For target densities whose structure is corrupted by pre-sphering (or pre-whitening) transforms, our unconstrained selector shows the most improvement over the existing plug-in selectors.

A copula-based test for independence

Natalie Neumeyer
University of Hamburg, Department of Mathematics, Germany,
neumeyer@math.uni-hamburg.de

Sebastian Kiwitt
University of Hamburg, Department of Mathematics, Germany,
kiwitt@math.uni-hamburg.de

Abstract. We suggest a new nonparametric estimator for a bivariate copula density, which is based on an orthogonal series expansion, and is itself a copula density. As application we consider a new asymptotically distribution-free test for independence of the components of bivariate random variables, which applies methods of order-selection tests. We deduce the asymptotic distribution and investigate the small sample performance by means of a simulation study. As further applications of the copula density estimator we discuss the estimation of bivariate densities in situations where informations about the marginals are available.

Advances in data depth

Claudio Agostinelli
Department of Statistics, Ca' Foscari University, Venice, Italy, claudio@unive.it

Mario Romanazzi
Department of Statistics, Ca' Foscari University, Venice, Italy, romanaz@unive.it

Abstract. Data depth provides several tools for nonparametric multivariate analysis. Of particular interest here are geometrical depth functions, like Tukey's halfspace depth, Liu's simplicial depth and Oja's simplicial volume depth, which derive information from simple geometrical structures. The main applications are a center-outward ordering of multivariate observations according to the values of the depth functions, location estimators with the meaning of multivariate medians, depth-weighted and depth-trimmed estimators of multivariate parameters and some new graphical presentations (scale curve, DD-plot) for both single-sample and paired-samples inspection and comparison. The present contribution is a review on some recent developments in the field.

The first topic is the robustness of depth ranks. Using the explicit expressions of the influence functions of both halfspace and simplicial depth values, a more precise understanding of the two definitions and an evaluation of their relative merits is now possible.

Next, there is an increased evidence that depth functions provide information not just on location but also on dispersion of a multivariate distribution. It has been proved that the Lebesgue integral of simplicial depth is the expected volume of a random simplex whose vertices are $p + 1$ independent observations from the reference distribution. This measure is known as the multivariate Gini's index, because in the scalar case it gives Gini's mean difference. The same result holds for Oja's and Tukey's depth functions, that is, the Lebesgue integral of both of them is equal to a specific multivariate dispersion measure. This opens a new research direction because depth methods appear to be successful both in location and dispersion investigations.

The final topic is local depth. By definition, depth provides a measure of centralness which is monotonically decreasing along any given ray from the deepest point. This implies that any depth function is unable to account for multimodality. To overcome this problem, a generalized notion of depth is required. We suggest to evaluate the centrality of a point conditional on a bounded neighbourhood. For example, the local version of simplicial

depth is the ordinary simplicial depth, conditional on random simplices whose volume is not greater than a prescribed threshold. Local versions of Oja's and Tukey's depths are similarly defined. These generalized depth functions are indeed able to record local fluctuations of the density function and can be useful in mode detection and cluster analysis.

Variable selection and weighting by nearest neighbor ensembles

Jan Gertheiss
Department of Statistics, LMU Munich, Germany,
jan.gertheiss@stat.uni-muenchen.de
Gerhard Tutz
Department of Statistics, LMU Munich, Germany,
tutz@stat.uni-muenchen.de

Abstract. In the field of statistical discrimination nearest neighbor methods are a well known, quite simple but successful nonparametric classification tool. In higher dimensions, however, predictive power normally deteriorates. In general, if some covariates are assumed to be noise variables, variable selection is a promising approach. In this presentation feature selection problems of small scale are investigated. In a first step the benefit of an extended forward as well as backward variable selection procedure for nearest neighbor classifiers is examined. The main focus however is on the development and testing of a nearest neighbor ensemble with implicit variable selection: Let $\hat{y}_{(j)}$ be the k nearest neighbor prediction, if the distance measure in the p-dimensional predictor space is only based on predictor x_j. Now the prediction $\hat{y}$ is constructed as an ensemble, i.e. a weighted average

$$\hat{y} = \sum_{j=1}^{p} c_j \hat{y}_{(j)}, \ \sum_{j=1}^{p} c_j = 1, \ c_j \geq 0 \, \forall j.$$

The weights - or *coefficients* - c_j have to be estimated, variable selection means setting $c_j = 0$ for some j. The latter can be done for example by hard thresholding. Coefficient estimation is performed via certain loss functions and quadratic programming. Finally the set of predictions can be augmented by including interactions of predictors. That means adding all nearest neighbor predictions based on two or even three predictors.
In contrast to other nearest neighbor approaches we are not primarily interested in classification, but in estimating the (posterior) class probabilities. So the used loss functions are mainly motivated that way. Ensemble adjustment is not necessary: If single predictions $\hat{y}_{(j)}$ are replaced by estimated probabilities, the resulting ensemble can be interpreted in a similar way.
In simulation studies and for real world data the investigated methods are compared to alternative well established classification tools (that offer probability estimates as well). Despite their simple structure, the proposed methods' performance is quite good - especially if relevant covariates can be separated from noise variables.

Peak preserving spectral density estimation

Lieven Desmet
Dept. of Mathematics, K.U. Leuven, Belgium,
lieven.desmet@wis.kuleuven.be
Irène Gijbels
Dept. of Mathematics, K.U. Leuven, Belgium,
irene.gijbels@wis.kuleuven.be

Abstract. Building on ideas developed in Gijbels, Lambert and Qiu, 2007 and on Gijbels and Desmet, 2007, on jump-preserving and peak-preserving regression, we look at the problem of spectral density estimation in stationary time series with short range dependence. The focus is on the estimation of prominent peaks which indicate the presence of important periodic components in the time series.

We propose a modification of the maximum likelihood estimator (based on the Whittle likelihood) of the log spectral density as developed by Fan and Kreutzberger, 1998. Those authors take a maximum likelihood approach and exploit the known distribution of the log periodogram ordinates Y_k, obtaining an estimator $\hat{\alpha}$ of the log spectral density in some frequency ω by maximizing the weighted sum

$$L^{\alpha,\beta}(\omega) = \sum_{i=1}^{n} [-\exp\{Y_k - \alpha - \beta(\omega_k - \omega)\} + Y_k - \alpha - \beta(\omega_k - \omega)] K_h(\omega_k - \omega)$$

where localization is achieved by means of kernel weights (h is a suitable bandwidth).

We introduce one-sided versions of the weighted sum (based on one-sided versions of the kernel) which lead to an enhanced estimation near peaks, and also provide an objective quantity indicating when this alternative estimation should be considered.
An extensive numerical study is provided as well as illustrations on real data.

1. Desmet, L., Gijbels, I. (2007) Peak preserving regression using local linear fitting. *Manuscript.*
2. Fan, J., Kreutzberger, E. (1998). Automatic Local Smoothing for Spectral Density Estimation. *Scand. J. Statist.* 25, 359–369.
3. Gijbels, I., Lambert, A., Qiu, P. (2007). Jump-preserving regression and smoothing using local linear fitting: a compromise. *Ann. Inst. Statist. Math.* 59, 235–272.

The additive-interactive nonlinear volatility model, its estimation and some testing issues

Michael Levine
Department of Statistics, Purdue University, USA,
mlevins@stat.purdue.edu
Tony (Jinguang) Li
Department of Statistics, Purdue University, USA, li115@stat.purdue.edu

Abstract. We consider a new separable nonparametric volatility model that allows for "interactions" in both mean and conditional variance (volatility) function. It can be concisely described as an additive-interactive nonlinear ARCH model. We propose this model as a possible alternative to the generalized additive nonlinear ARCH (GANARCH) model of Kim and Linton (2004), with which it shares the common origin. Unlike the GANARCH model, it does not assume the known link function but includes second-order interaction terms in both mean and variance functions instead. This ensures a much more data-driven model compared to GANARCH of Kim and Linton (2004) since our assumptions do not assume that anything know about the data distribution. This is very beneficial since in practice the data distribution has to be selected based on the exploratory data analysis which is very difficult for multivariate data. Thus, the proposed model is much more flexible compared to GANARCH.
Motivated by the local instrumental variable estimation method (LIVE), also introduced in Kim and Linton (2004), we propose instrumental variable-based estimators of the components of the mean and volatility functions. The estimators are shown to be consistent and asymptotically normal. Explicit expressions for asymptotic means and variances of these estimators are also obtained. Several simulation experiments are conducted that show a very good performance of our algorithm for moderate sample sizes. Finally, the method is applied to the real data set of currency exchange rates where it leads to some interesting conclusions.
Historically, multiple functional component testing in nonparametric models has been a fairly difficult problem. We introduce a novel F-type approach to testing the significance of the two-way interactive terms in the mean function based on the unbalanced design ANOVA with unequal variances. Simulation studies show that the method performs very well for sample sizes of about 5000 which are easily available in financial applications.

Specification tests for the distribution of errors in nonparametric regression: a martingale approach

Juan Mora
Universidad de Alicante, Spain, juan@merlin.fae.ua.es
Alicia Pérez-Alonso
Universidad Carlos III de Madrid, Spain, aperez1@eco.uc3m.es

Abstract. We discuss how to test whether the distribution of regression errors belongs to a parametric family of continuous distribution functions, making no parametric assumption about the conditional mean or the conditional variance in the regression model. More specifically, let (X, Y) be a bivariate continuous random vector such that $\mathrm{E}(Y^2)$ is finite, denote $m(x) \equiv \mathrm{E}(Y|X = x)$, $\sigma^2(x) \equiv \mathrm{Var}(Y|X = x)$ and consider the error term $\varepsilon \equiv \{Y - m(X)\}/\sigma(X)$, which is, by definition, a zero-mean unit-variance random variable. If $F_\varepsilon(\cdot)$ denotes the c.d.f. of ε and $\mathcal{F} \equiv \{F(\cdot, \theta),\, \theta \in \Theta \subset \mathbb{R}^m\}$ denotes a parametric family of zero-mean unit-variance continuous c.d.f.'s, each of them known except for the parameter vector θ, we propose a testing procedure to face the hypotheses

$$\begin{aligned} \mathrm{H}_0 &: \exists\, \theta_0 \in \Theta \text{ such that } F_\varepsilon(\cdot) = F(\cdot, \theta_0), \qquad \text{vs.} \\ \mathrm{H}_1 &: F_\varepsilon(\cdot) \notin \mathcal{F}, \end{aligned}$$

when independent and identically distributed observations $\{(X_i, Y_i)\}_{i=1}^n$, with the same distribution as (X, Y), are available. In principle, one could think of using a Kolmogorov-Smirnov or a Cramér-von Mises statistic, constructed replacing errors by residuals and parameters by estimates. However, using the results derived in Akritas and Van Keilegom (2001, *Scand. J. Statist.*) the asymptotic distribution of these residual-based statistics can be derived, and it proves to be not asymptotically distribution-free, a property that is already well-known in the literature. Then, we follow the methodology introduced in Khmaladze (1993, *Ann. Statist.*) to derive asymptotically distribution-free martingale-transformed test statistics. Finally, we derive the asymptotic distribution and the consistency of these martingale-transformed statistics under appropriate conditions. Two Monte Carlo experiments show that the transformed statistics work reasonably well in terms of size and power, and that their behaviour is not very sensitive to the choice of the smoothing value.

Representation of the conditional distribution function of a censored variable given a linear combination of a d-vector of covariates

María Carmen Iglesias-Pérez
Dpto. Estadística e I.O., Univ. Vigo, Spain, mcigles@uvigo.es
Wenceslao González-Manteiga
Dpto. Estadística e I.O., Univ. Santiago, Spain, wenceslao@usc.es

Abstract. An important aim in survival analysis is to study how an explanatory covariate vector, X, influences the survival or duration time, Y, when this duration time is not completely observed because, for example, the presence of censoring. This dependence may be modeled in a different numbers of ways but, typically, the key idea is to assume some kind of functional relationship for the conditional distribution function (cdf) or some other conditional curve, as the conditional hazard function or the conditional hazard rate (see Cao and González Manteiga, 2007). We focus our attention on the estimation of the cdf when we have a random d-vector of covariates with $d > 1$. Even for small values of d, the conventional nonparametric estimators can suffer poor accuracy and then a solution to this difficulty can be achieved via the estimation of the cdf of Y given $\theta_0^t X$, where θ_0 is a certain vector of parameters (see Hall and Yao (2005) for this approach with complete data).
Here we consider the estimation of the cdf of Y given $\theta_0^t X$ for censored data. We obtain a representation when θ_0 is replaced by some $n^{1/2}$-consistent estimator, $\hat{\theta}$. Specifically, in the multivariate context ($d > 1$) and under several conditions which include $\hat{\theta} - \theta_0 = o_P\left(n^{-2/5}\right)$, we prove that

$$\hat{F}\left(y \mid \hat{\theta}^t x\right) = \hat{F}\left(y \mid \theta_0^t x\right) + o_P\left(n^{-2/5}\right),$$

where the estimator $\hat{F}(\bullet \mid x)$ of $F(\bullet \mid x)$, the conditional distribution function of $Y \mid X = x$, is the estimator introduced by Iglesias-Pérez and González-Manteiga (1999) in the single covariate context ($d = 1$).
This representation and the asymptotic properties of $\hat{F}(\bullet \mid x)$ prove that $\hat{F}\left(y \mid \hat{\theta}^t x\right)$ and $\hat{F}\left(y \mid \theta_0^t x\right)$ have the same asymptotic distribution.
Some examples are included to illustrate this estimation.

1. Cao, R., González-Manteiga, W. (2008). Goodness-of fit tests forconditional models under censoring and truncation. *J. of Econometrics* 143, 166–190.

2. Hall, P., Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.* 33, 1404–1421.
3. Iglesias-Pérez, M.C., González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *J. Nonparametr. Statist.* 10, 213–244.

A test for comparing regression curves versus one-sided alternatives

Juan Carlos Pardo-Fernández
Universidade de Vigo, Spain, juancp@uvigo.es
Natalie Neumeyer
Universität Hamburg, Germany, neumeyer@math.uni-hamburg.de

Abstract. Consider two pairs of random variables (X_j, Y_j), for $j = 1, 2$, related via nonparametric regression models

$$Y_j = m_j(X_j) + \varepsilon_j,$$

where $m_j(x) = E(Y_j|X_j = x)$ is the conditional mean. We assume that the covariates X_1 and X_2 have common support R_X. Under this setup it is of practical interest to test for the equality of the regression curves, which means that effect of the covariate over the response is the same in both populations. In some situations, additional information can be given about the alternative hypothesis. In this talk, we study the problem of testing for the null hypothesis

$$H_0 : \quad m_1(x) = m_2(x) \text{ for all } x \in R_X,$$

versus *one-sided alternatives*, that is, one function is always equal or bigger than the other:

$$\begin{aligned} H_1 : \quad & m_1(x) \leq m_2(x) \text{ for all } x \in R_X \text{ and} \\ & m_1(x) < m_2(x) \text{ in a set of positive measure.} \end{aligned}$$

We propose a very simple testing procedure, which is based on the following idea. Let m be any function verifying $m_1(x) \leq m(x) \leq m_2(x)$, for all $x \in R_X$. Define the random variables, for $j = 1, 2$,

$$\varepsilon_{j0} = Y_j - m(X_j),$$

which can also be expressed as

$$\varepsilon_{j0} = \varepsilon_j + (m_j(X_j) - m(X_j)).$$

Obviously, under the null hypothesis, $m_1(x) = m(x) = m_2(x)$, and $\varepsilon_{j0} = \varepsilon_j$. However, under the alternative hypothesis it happens that

$$E(\varepsilon_{10}) < 0 \quad \text{and} \quad E(\varepsilon_{20}) > 0.$$

Therefore the comparison of the expectations of the regression errors under the null hypothesis can be used to detect the alternative hypothesis H_1. In practice, the regression errors are replaced by residuals estimated in a nonparametric way.
In this talk, we will explain the testing procedure, and we will show some related theoretical results and simulations.

Nonlinear wavelet regression for left-truncated, dependent data

Jacobo de Uña-Álvarez
University of Vigo, Spain, jacobo@uvigo.es
Han-Ying Liang
University of Vigo, Spain, and Tongji University, China, hyliang83@yahoo.com

Abstract. Left-truncated data appear in a number of applications, including Astronomy, Survival Analysis, and Economics. Most papers dealing with left-truncated data assume that the data are independent. However, when sampling cluster of individuals (e.g. family members, or repeated measurements taken on the same subject), the observations will be typically correlated. In this talk we introduce a new nonlinear wavelet-based estimator of the regression function when the response variable is left-truncated. It is assumed that the observations form a stationary α-mixing sequence. The nonlinear wavelet-based estimator of the covariate's density is considered as well. We establish asymptotic results for the new estimators, as an asymptotic expression of the mean integrated squared error (MISE). Under standard conditions, it is seen that the rate of convergence of the MISE is not affected by the presence of discontinuities in the underlying curves, nor by the dependence structure of the data.

Nonparametric mixture regression

Alex Rojas
Carnegie Mellon University in Qatar, Qatar, arojas@qatar.cmu.edu
Chris Genovese
Carnegie Mellon University, USA, genovese@stat.cmu.edu
Larry Wasserman
Carnegie Mellon University, USA, larry@stat.cmu.edu

Abstract. In this talk we consider the problem of estimating conditional densities functions. Conditional density estimation is a technique that, compared to usual regression methods, allows for a better understanding of the relation- ship between a response variable and a set of covariates. We present a new conditional density estimator based on finite mixture models. This estimator summarizes the relationship between the covariates and the response with a set of *parameter functions*, which describes the conditional behavior succinctly. This feature gives the proposed model the advantage of being parsimonious and easily interpretable. We consider two methods for fitting the model: local likelihood and a conditional minimum distance approach. We apply the proposed estimator to study the role environment plays in the process of galaxy evolution.

Longitudinal modeling when response and time-dependent covariates are measured at different timepoints

Joel A. Dubin
University of Waterloo, Canada, jdubin@uwaterloo.ca
Xiaoqin Xiong
University of Waterloo, Canada

Abstract. In this talk, we will discuss a flexible method to handle both association and temporal sequencing of distinct longitudinal measures, where the measures may be of mixed type (e.g., one continuous, the other binary) and recorded on non-uniform grids and different time points from one another. A smoothing step will be involved. The approach will be demonstrated on a dataset of hemodialysis patients, where longitudinal measures of health outcomes (e.g., infection) were recorded at different time points than longitudinal physiologic measures such as serum C-reactive protein levels (a marker for inflammation). An interesting scientific question to answer is whether experiences of infection follow or predate inflammation.

IRT analysis of STAIC data

Barbara Oliveiros
IBILI, Faculdade de Medicina de Coimbra, Portugal, bpaiva@ibili.uc.pt
Alexandre Gomes da Silva
ISCAC, Instituto Politécnico de Coimbra, Portugal, asilva@iscac.pt
Emanuel Ponciano
IBILI, Faculdade de Medicina de Coimbra, Portugal, ponciano@ibili.uc.pt

Abstract. STAIC studies like other studies involving Likert-type surveys are commonly analyzed compiling data and reporting means and standard deviations. More recently confirmatory factor analysis is also used to explore this kind of data. This study presents the analysis of STAIC type data using item response theory. The data consist of Portuguese students. Individual performances and items consistency were analyzed. Properties of item theory are discussed.

A semiparametric estimator for spatial count data regression analysis

José António Santos
Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Portugal, jsantos@isegi.unl.pt

M. Manuela Neves
ISA, Technical University of Lisbon, Portugal, manela@isa.utl.pt

Carlos Barroso
Cesam, Departamento de Biologia, Universidade de Aveiro, Portugal, cmiguez@ua.pt

Francesco Costigliola
Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Portugal, francesco.costigliola@gmail.com

Francisco Maia
INRB - IPIMAR, CRIP Centro, Portugal, maia.francisco@gmail.com

Abstract. We present a semiparametric estimator for count data regression analysis based on two covariates. This is an extension of the model concerning only one covariate that was already developed [1].
A local maximum likelihood estimator based on Poisson regression is presented as well as its asymptotic bias, variance and distribution.
This semiparametric estimator is an alternative to the parametric count data regression models that does not depend on regularity conditions and model specification accuracy.
Consider Y as a count random variable with support $\mathbb{N}_0$ and two covariates X_1, X_2. The conditional mean of Y is:

$$\mathrm{E}[Y|X_1 = x_{1i}, X_2 = x_{2i}] = \lambda(x_{1i}, x_{2i}) = \exp\left[m(x_{1i}, x_{2i})\right], \qquad (1)$$

where $m(x_{1i}, x_{2i})$ is an unknown function of interest to be estimated through local polynomial smoothing.
Considering a Taylor development of degree one as an approximation to $m(x_{1i}, x_{2i})$, where (x_{1i}, x_{2i}) is in a neighborhood of (x_1, x_2), we have:

$$\lambda(x_{1i}, x_{2i}) \approx \exp\left[\beta_0 + \beta_1(x_{1i} - x_1) + \beta_2(x_{2i} - x_2)\right]. \qquad (2)$$

In the context of the Poisson regression the logarithm of the local likelihood function is:

$$\mathcal{L}_1(\beta_0, \beta_1, \beta_2|\mathbf{X}, \mathbf{y}, (x_1, x_2), h) =$$

$$h^{-2}\sum_{i=1}^{n}\Big\{(-\lambda(x_{1i},x_{2i})+y_i\ln\lambda(x_{1i},x_{2i})-\ln(y_i!)) \times K\left((x_1-x_{1i})/h,(x_2-x_{2i})/h\right)\Big\}, \tag{3}$$

where $\lambda(x_{1i},x_{2i})=\exp\left[\beta_0+\beta_1(x_{1i}-x_1)+\beta_2(x_{2i}-x_2)\right]$.
Additionally other topics deserve special interest here such as the bandwidth selection procedure (sample splitting), the spatial dependence of the observations and the spatial nature of the covariates.
The main motivation for the model we present is to study the spatial distribution of *Nassarius reticulatus* in Ria de Aveiro (NW Portugal). This species is an abundant mollusc of the Portuguese coast that has been widely used as a bioindicator of tributyltin (TBT) pollution, a serious problem caused by ship antifouling paints. The data are counts of *N. reticulatus* gastropods that were collected randomly along the eight channels of Ria de Aveiro and the covariates refer to the geographical position (east/west, north/south) where each sample was collected.
The results are compared to those arising from other approaches particularly spatial statistics.

1. Santos, J. A., Neves, M. M. (2007). A Local Maximum Likelihood Estimator for Poisson Regression. *Metrika*, to appear.

Forecasting M3 competition data - another approach

Clara Cordeiro
FCT, University of Algarve, Portugal, ccordei@ualg.pt
Manuela Neves
ISA, Technical University of Lisbon, Portugal, manela@isa.utl.pt

Abstract. The M3 competition data is often used by many researchers in their investigation. These 3003 time series are a good database for testing new methodologies using previous results. The results in Makridakis and Hibon (2000) are frequently used as a benchmark in comparative studies. This article presents 24 forecasting methods and several accuracy measures are used to analyze and classify the performance of the various methods.
The method proposed in this work combines the use of exponential smoothing methods with the resampling technique bootstrap to forecast time series and it works in an automatic way. The procedure starts by selecting the best exponential smoothing method according to the characteristics that a times series reveals. After adjusting the best method, our attention is drawn to the residual part. The bootstrap is then used after an autoregressive adjustment, selected by AIC criterion. The time series is then reconstructed, adding the initial characteristics (if they exist) to the bootstrapped residuals, see Cordeiro and Neves (2007a, b) for more details. Forecasts are finally obtained using the model initially selected.
To evaluate this procedure some accuracy measures such as the symmetric mean absolute percentage error and the root mean squared error are calculated in order to compared with the competition results in Makridakis and Hibon (2000). All this computational work is performed with the software R 2.6.2 (R Development Core Team (2008)).

1. Cordeiro, C., Neves, M. (2007a). Bootstrap prediction intervals: a case-study. In: Joan del Castillo, Anna Espinal and Pere Puig (Eds.): Proceedings of the 22nd International Workshop on Statistical Modelling (IWSM2007), 191-194.
2. Cordeiro, C., Neves, M. (2007b). Resampling techniques in time series prediction: a look at accuracy measures. In: Gomes, M.I., Pestana, D., Silva, P.(eds): Proceedings of the 56th Session of the International Statistical Institute(ISI 2007), pag. 353 extended abstrat in CD-ROM.
3. Makridakis, S., Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *Internat. J. Forecasting* 16, 451-476.

Estimation of the restricted conditional mean gap time: the induced dependent censoring aspect

Adin-Cristian Andrei
Department of Biostatistics and Medical Informatics, University of Wisconsin, USA, andrei@biostat.wisc.edu

Abstract. In numerous clinical trials, information is available on a series of successive landmark events. Such sequences may include: randomization time, date of first and second hospitalization and death date. The time elapsed between two successive events is called a gap time. In the presence of censoring, one may not observe the gap times of interest in their entirety. Even under independent censoring, all but the first gap time are subject to induced dependent censoring (IDC). Via simulation studies, we investigate the magnitude of the IDC problem in the estimation of the restricted conditional mean of the most recent gap time, given all prior gap times. We propose, as a possible solution, estimators based on inverse probability-of-censoring weighting techniques, and show that they are consistent and asymptotically normal. Simulations are performed in a variety of scenarios and an example is used to illustrate these methodological developments.

A geometric interpretation of the multiresolution criterion

Thoralf Mildenberger
Technische Universität Dortmund, Germany,
mildenbe@statistik.uni-dortmund.de

Abstract. A recent approach to choosing the amount of smoothing or regularization in nonparametric regression is to select the simplest estimate for which the residuals "look like noise". This can be checked with the so-called multiresolution criterion, which Davies and Kovac introduced in connection with their taut-string procedure [Davies and Kovac (2001): Local extremes, runs, strings and multiresolution (with discussion and rejoinder), AOS(29) 1-65]. It has also been used in several other nonparametric procedures like spline smoothing [Davies and Meise (2008): Approximating Data with weighted smoothing splines, to appear in JNPS; Davies, Kovac and Meise (2008): Nonparametric regression, confidence regions and regularization, to appear in AOS] or piecewise constant regression [Boysen, Kempe, Munk, Liebscher and Wittich (2007): Consistencies and rates of convergence of jump penalized least squares estimators, to appear in AOS]. We show that this criterion is related to a norm, the *multiresolution norm* (MR-norm). We point out some important differences between this norm and p-norms: The MR-norm is not invariant w.r.t. sign changes and permutations and this makes it useful for detecting runs of residuals of the same sign. We also give sharp upper and lower bounds for the MR-norm in terms of p-norms.

Nonparametric Bayesian inference for integrals with respect to an unknown finite measure

Torkel Erhardsson
Department of Mathematics, Linköping University, Sweeden,
toerh@mai.liu.se

Abstract. In this talk we consider the following problem:
Problem 1. Let μ be an unknown finite measure on a measurable space $(S, \mathcal{S})$. Define $g : S \to \mathbb{R}^n$ by $g = (g_1, \ldots, g_n)$, where the functions $g_i : S \to \mathbb{R}$, $i = 1, \ldots, n$, are measurable and linearly independent. Let $T(\mu) = \int_S g(x) d\mu(x) \in \mathbb{R}^n$. We observe an ($\mathbb{R}^k$-valued) random variable Y whose distribution given the value of $T(\mu)$ is known. The goal is to estimate $T(\mu)$ (or μ itself).
Problem 1 belong to the large class of inverse problems, which often arise in applications. An inverse problem involves a mapping $T : S_1 \to S_2$. For an unknown $\mu \in S_1$ one observes the value Y , which is $T(\mu)$ contaminated with random "noise". Typically, the problem of solving the "equation" $T(\mu) = Y$ is ill-posed, since either $Y \notin Im(T)$, T is not invertible, or T^{-1} exists but is very irregular.
The approach taken in this talk, which has grown in importance in recent years, is to view inverse problems as statistical problems, for which it is natural to use Bayesian inference. Thus, both μ and Y are considered to be random quantities, the probability distribution of which (called the "prior" distribution) represents initial beliefs about μ and Y. Bayesian inference consists in computing the conditional distribution of μ given Y (called the "posterior" distribution), which represents beliefs about μ after observing Y.
A number of examples of successful applications of Bayesian methods to inverse problems have appeared in recent years. Recently, Wolpert et al. (2003) proposed a nonparametric Bayesian approach to Problem 1, in which a random measure with independent increments is used as a prior for μ, and where a Metropolis-Hastings type MCMC algorithm is used to sample from an approximation to the posterior distribution.
The main result presented in this talk is a new method to carry out Bayesian inference for $T(\mu)$ in Problem 1. In the case when μ is a probability measure, we use a Dirichlet process as a prior for μ, and construct an approximation to the posterior distribution of the integrals using the SIR algorithm and samples from a new multidimensional version of a Markov chain by Feigin and Tweedie. The method can be modified to handle the case when μ is a finite measure; we then use a Gamma process as a prior

for μ. We prove that the Markov chain is positive Harris recurrent, and that the approximating distribution converges weakly to the posterior as the sample size increases, under a mild integrability condition. Furthermore, the rate of convergence seems quite fast in typical applications. The method is therefore a promising alternative to the previously suggested approach to Problem 1.

Multiresolution and model choice

Arne Kovac
University of Bristol, UK, a.kovac@bristol.ac.uk

Abstract. We consider various settings of the nonparametric regression problem and consider the multiresolution criterion where for given data $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$ we require an approximation f to satisfy

$$|\sum_{i \in I} y_i - f_i| \leq c_I$$

for all $I \in \mathcal{I}$ where $\mathcal{I}$ is some family of subintervals of $\{1, \ldots, n\}$. Overall aim is to find functions that satisfy this criterion while at the same being smooth and simple in the sense of avoiding artificial local extreme values. We explore this concept in the usual regression context, but also expand it to inverse problems, estimation of parameters in differential equations, bivariate curves and online data.

1. Davies, P. L., Kovac, A., Meise, M. (2008). Confidence Regions, Regularization and Non-Parametric Regression. *Ann. Statist.*, (to appear).
2. Davies, P. L., Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Ann. Statist.* 29, 1–65.
3. Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J., (2007). Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. Royal Statist. Soc., Series B* 69, 741–796.

Weak convergence of the supremum distance for supersmooth kernel deconvolution

Bert van Es
Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, The Netherlands, vanes@science.uva.nl
Shota Gugushvili
Eurandom, Technische Universiteit Eindhoven, The Netherlands, gugushvili@eurandom.tue.nl

Abstract. Let $f_{nh}(x)$ denote the deconvolution kernel density estimator. We establish the asymptotic distribution of the supremum distance $\sup_{x\in[0,1]} |f_{nh}(x) - \mathrm{E}[f_{nh}(x)]|$, which provides a global measure of performance of the deconvolution kernel density estimator. We consider the supersmooth deconvolution problem, in particular deconvolution for error distributions with characteristic functions that have an exponential tail like the characteristic function of a normal density. It turns out that the asymptotics are essentially different from corresponding results in the ordinary smooth deconvolution. We also briefly discuss the method of construction of the uniform confidence intervals for the target density f.

Nonparametric model checking for dynamic load sharing models

Eric Beutner
Department of Quantitative Economics, Maastricht University, Netherlands, e.beutner@ke.unimaas.nl

Abstract. Recently, nonparametric statistical methods have been successfully applied to dynamic reliability models and sequential k-out-of-n systems. Load-share models assume that the failure rates of the components depend on the operating status of the other system components. An important element of the load-share model is the rule which governs how failure rates change after some components failed in the system. In this talk, we focus on nonparametric model checks for dynamic reliability models where the conditional cumulative hazard rates of the failure times are proportional. It will be shown that for a $(n-1)$-out-of-n dynamic reliability model the asymptotic distribution of our test statistic is the sup of a time transformed Brownian bridge. We establish consistency of the test. For the general case our test statistic will be based on weighted martingale residuals. It turns out that the asymptotic distribution of the test statistic is rather intractable. Using Khmaladze's goodness-of-fit idea we construct a test statistic which has asymptotically a much more tractable distribution.

1. Beutner, E., (2008). Nonparametric inference for sequential k-out-of-n systems. *Ann. Inst. Statist. Math.*, to appear
2. Hollander, M., Peña, E.A. (1995). Dynamic reliability models with conditional proportional hazards. *Lifetime Data Analysis* 1, 377–401.
3. Kvam, P.H., Peña, E.A. (2005). Estimating load-sharing properties in a dynamic reliability system. *J. Amer. Statist. Associ.* 100, 262–272.

A SDE growth model: nonparametric estimation of the drift and the diffusion coefficients

Patrícia A. Filipe
CIMA - Universidade de Évora, Portugal, pasf@uevora.pt
Carlos A. Braumann
CIMA - Universidade de Évora, Portugal, braumann@uevora.pt

Abstract. We study a stochastic differential equation (SDE) growth model to describe individual growth in random environments. In particular, in this work, we discuss the estimation of the drift and the diffusion coefficients using non-parametric methods. We illustrate the methodology by using bovine growth data.
Considering the diffusion process X_t, describing the weight of an animal at age t, characterized by the stochastic differential equation $dX_t = a(X_t)dt + b(X_t)dW_t$, with W_t being the Wiener process, we estimate the infinitesimal coefficients $a(x)$ and $b(x)$ nonparametrically. Our goal was to analyse which of the parametric models (with specific functional forms for $a(x)$ and $b(x)$) previously used by us to describe the evolution of bovine weight were good choices and also to see whether some alternative specific parametrized functional forms of $a(x)$ and $b(x)$ might be suggested for further parametric analysis of this data.

Smoothing non-stationary correlated data

Peter Foster
School of Mathematics, University of Manchester, UK,
peter.foster@manchester.ac.uk

Abstract. The smoothing technique we will be focussing on is local linear regression while the non-stationary covariance model we assume is the structured ante-dependance model of order one, denoted SAD(1), which will be fully explained. An expression for the asymptotic MISE of the smoother is derived and shown to be dependent, amongst other things, on parameters defining the SAD(1) model. A new non-linear least squares approach is used to nonparametrically fit the SAD(1) model and this will be described in detail. The strategy is to choose the smoothing parameter for the regression by minimizing the asymptotic MISE which we can do in a practical way by plugging-in estimates of the values of unknown quantities. Finally, simulations show the benefits of modeling the non- stationarity in the data rather than erroneously assuming stationarity.

On confidence intervals for a distribution function under association

Cecília Azevedo
Universidade do Minho, Portugal, cecilia@math.uminho.pt

Abstract. In this paper we construct confidence sets for a marginal distribution function F of a strictly stationary sequence of associated random variables.
We address the question of the nonparametric estimation of the asymptotic variance of $\sqrt{n}\,\widehat{F}_n(x)$, where $\widehat{F}_n(x)$ is the kernel estimator of $F(x)$. We consider two estimators and study their asymptotic properties. The first estimator (denoted by $\widehat{\sigma}_n(x)$) is obtained by using a family of kernel estimators for bivariate distribution functions $F_k(x,y)$, with fixed $k = 1, 2, \ldots, n-1$, of pairs of associated random variables (X_j, X_{k+j}), $j \geq 1$ [1] and the other one, which can be viewed as an empirical variance (denoted by $\widetilde{\sigma}_n(x)$), is inspired by the paper of Guillou and Merlevède [3]. Results on weak consistency of $\widehat{F}_n(x)$ [2] and of the proposed estimator for the bivariate distribution function, $\widehat{F}_{n,k}(x,y)$ [1], and also the asymptotic normality of $\widehat{F}_n(x)$ [2], enable us to construct approximate confidence intervals for $F(x)$ for every x. We denote these intervals by $\widehat{I}$ in the first case and by $\widetilde{I}$ in the second case.
A simulation study is presented to assess the finite sample performance of the proposed confidence intervals, $\widehat{I}$ and $\widetilde{I}$. We consider segments of associated random variables from two different populations, of size n (large), and for each of them we generate m (much larger than n) independent sets of associated Monte Carlo random segments from each population. Also, for each population and each value of x and sample size n, based on m confidence intervals from simulation, we evaluate the lower and upper confidence limits of $\widehat{I}$ and $\widetilde{I}$ with the same confidence level. We also present several illustrative graphics.

1. Azevedo, C., Oliveira, P.E. (2000). Kernel-type estimation of bivariate distribution function for associated random variables. *New Trends in Probability and Statistics, Proceedings of the* 6^{th} *Tartu Conference, VSP* 5, 17-25.
2. Cai, Z., Roussas,G.G. (1996). Berry-Esseen Bounds for Smooth Estimator of a Distribution Function under Association, *J. Nonparametr. Statist.* 11, 79-106.
3. Guillou, A., Merlevède, F. (2001). Estimation of the Asymptotic Variance of Kernel Density Estimators for Continuous Time Processes, *J. Multivariate Anal.* 79, 114-137.

Stability estimating in optimal sequential hypotheses testing

Evgueni Gordienko
Autonomous Metropolitan University - Iztapalapa, Mexico City, Mexico, gord@xanum.uam.mx

Andrey Novikov
Autonomous Metropolitan University - Iztapalapa, Mexico City, Mexico, an@xanum.uam.mx

Elena Zaitseva
Autonomous Metropolitan University - Iztapalapa, Mexico City, Mexico, lenagordi@hotmail.com

Abstract. We study the stability of the classical optimal sequential probability ratio test based on independent identically distributed observations $X_1, X_2, \ldots$ when testing two simple hypotheses about their common density f: $f = f_0$ versus $f = f_1$. As a functional to be minimized, it is used a weighted sum of the average (under f_0) sample number and the two types error probabilities. We prove that the problem is reduced to stopping time optimization for the likelihood ratio process generated by $X_1, X_2, \ldots$ with the density f_0. For τ_* being the corresponding optimal stopping time we consider asituation when this rule is applied for testing between f_0 and an alternative $\tilde{f}_1$, where $\tilde{f}_1$ is some approximation to f_1. An inequality is obtained which gives an upper bound for the cost excess, when τ_* is used instead of the rule $\tilde{\tau}_*$ optimal for the pair $(f_0, \tilde{f}_1)$. The inequality found also estimates the difference between the minimal costs for optimal tests corresponding to the pairs (f_0, f_1) and $(f_0, \tilde{f}_1)$.

MAP estimation for curve modeling with free-knot splines

Laure Amate
Laboratoire I3S, CNRS-UNSA, amate@i3s.unice.fr
Maria João Rendas
Laboratoire I3S, CNRS-UNSA, rendas@i3s.unice.fr

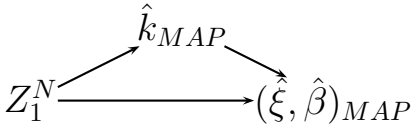
Abstract. In the context of curve modeling, splines have been widely used and studied ([1], [2], [3], [4]). Main arguments for using spline functions as function approximators are their ability to fit complex forms with arbitrary good accuracy, and the existence of basis functions, the B-splines, with attractive numerical properties. Our use of splines is motivated by the desire to find a parsimonious parametric description of a curve $s(t)$ of which we observe noisy samples:

$$Z_n = s(t_n) + \epsilon(t_n) \simeq f(t_n; \hat{\theta}(Z_1^N)) + \varepsilon(t_n), \qquad n = 1, \ldots, N, \qquad t_n \in [0, 1]$$

In this equation, $f(t;\theta)$ is a spline function, and $\varepsilon(t)$ represents observation noise ($\epsilon(t)$) and modeling errors. The approximating spline is an element of

$$\mathcal{G} = \cup_{k \in \{k_{min}, k_{max}\}} \$_k^m \ ,$$

where, inspired by [1], $\$_k^m$ is the space of splines of degree m with k knots. Vector $\theta = (k, \xi, \beta)$ identifies $f(\cdot;\theta)$ as a member of $\mathcal{G}$: the number of knots, k, identifies $\$_k^m$, the knot vector ξ indicates that $f \in \$_{k,\xi}^m$, and β are the B-Spline.
We use a Bayesian approach and $\hat{\theta}(Z_1^N)$ are the MAP estimates in $\mathcal{G}$ for a postulated prior. While MAP estimation of $s(t)$ has been addressed by other authors, leading to a model $\hat{s}(t) \notin \mathcal{G}$, the problem of building a parametric model for s in $\mathcal{G}$ has received much less attention. An exception is [5], where in a slightly different context, MAP estimation in "union models" is also addressed. We depart in a significant way from the work of these authors, in the crucial issue of defining MAP estimates for spaces with this composite structure. A careful interpretation of the notation $p(\theta|Z_1^N)$ – that we prefer to decompose as $p(\xi, \beta|k, Z_1^N)P(k|Z_1^N)$ – is required: for different values of k, densities $p(\xi, \beta|k, Z)$ are defined with respect to distinct measures, and their direct comparison is meaningless, invalidating direct maximization of the numbers $p(\theta|Z_1^N)$.
We advocate that MAP estimation in models with the union structure of $\mathcal{G}$ must proceed in a stepwise manner:



Our implementation is inspired of BARS [4], using MC methods to sample from the joint posterior on (k, ξ). By marginalizing over ξ, we approximate $P(k|Z_1^N)$. Optimization is based on a simulated annealing chain over ξ for k fixed at $\hat{k}_{MAP}$ ($\hat{\beta}_{MAP}$ is determined analytically). We note that we also estimate the noise variance. In the experiments presented, the approximation accuracy (MSE) of our MAP model $f(t; \hat{\theta}_{MAP})$ is comparable to that of BARS's estimates $\hat{s}_{BARS}$, with the advantage of providing a compact data description with $2\hat{k}_{MAP} + 1$ parameters.

1. DeBoor, C. (1978). *A Practical Guide to Splines.* Springer, 1978.
2. Luo, Z., Wahba, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* 92, 107-116.
3. Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *J. Comput. Graph. Statist.* 8, 333-352.
4. DiMatteo, I., Genovese, C., Kass, R. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* 88, 1055-1071.
5. Andrieu, C., Freitas, N. D., Doucet, A. (2001). Robust full bayesian learning for radial basis networks. *Neural Comput.* 13, 2359-2407.

Nonparametric adaptive Bayesian oracle projection estimation in the white noise model

Aleksandra Babenko
Utrecht University, Netherlands, babenko@math.uu.nl
Eduard Belitser
Utrecht University, Netherlands, belitser@math.uu.nl

Abstract. We consider an oracle projection estimator in Gaussian white noise model. Under appropriate hierarchical prior we construct a Bayes estimator for the cut-off parameter of the projection estimator and show that the resulting adaptive projection estimator for the signal satisfies an oracle quadratic risk inequality. The main tool in our approach is the exponential inequality for the posterior probabilities of the cut-off parameter.

B-splines regression smoothing and difference type of penalties

Irène Gijbels
Katholieke Universiteit Leuven, Belguim, irene.gijbels@wis.kuleuven.be
Anneleen Verhasselt
Katholieke Universiteit Leuven, Belgium,
anneleen.verhasselt@wis.kuleuven.be

Abstract. P-splines were first introduced by Eilers and Marx (1996) to approximate the unknown mean-function via a regression with B-splines and a penalty on the sum of squared k-th order differences of the regression coefficients. These P-splines are an extension of ordinary least squares regression which provides a solution to the problem of overfitting of the latter. Other frequently used penalties, which were first proposed in the model selection context, are LASSO, bridge and elastic net. In the optimization problem of B-splines regression with penalties, some parameters such as the smoothing parameter, the degree of the B-splines, the number of knots and the penalty should be chosen. A particular method, based on the minimization of the Akaike information criterion, to choose the order of the penalty and the smoothing parameter for generalized linear models is presented. Notably, a theoretical 'best' value of the difference order is investigated.

1. Eilers, P., Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.*, 11, 89–102.

Principal points and elliptical distributions from the multivariate setting to the functional case

Juan Lucas Bali
Universidad de Buenos Aires - FONCyT, Argentina,
lucasbl3@yahoo.com.ar
Graciela L. Boente Boente
Universidad de Buenos Aires-CONICET, Argentina, gboente@dm.uba.ar

Abstract. The k principal points of a random vector $\mathbf{X}$ are defined as a set of points which minimize the expected squared distance between $\mathbf{X}$ and the nearest point in the set. They are thoroughly studied in [1], [2], [5] and [7]. For their treatment, the examination is usually restricted to the family of elliptical distributions ([3] and [4]), requiring some auxiliary results regarding self-consistency of distributions([6]). In this talk we will present an extension of the previous results to the functional case. That is, we think of random elements over a separable Hilbert space $\mathcal{H}$ and we generalize the concepts of principal points, self-consistent points and of elliptical distributions so as to fit them in this functional framework. Results linking self-consistency and the eigenvectors of the covariance operator are re-obtained in this new setting. We also generalize the explicit formula for the $k = 2$ case in [1] so as to include elliptically distributed random elements in $\mathcal{H}$.

1. Flury, B. A. (1990). Principal Points, *Biometrika* 77, 33–41.
2. Flury, B. A. (1993). Estimation of Principal Points, *Appl. Statist.* 42, 139–151.
3. Frahm, G. (2004). *Generalized Elliptical Distributions: Theory and Applications*. PhD thesis from the University of Cologne, Germany.
4. Muirhead R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley and Sons Canada.
5. Tarpey, T. (1995). Principal Points and Self–Consistent Points of Symmetric Multivariate Distributions. *J. Multivariate Anal.* 53, 39–51.
6. Tarpey T. y Flury B. (1996). Self-Consistency: A Fundamental Concept in Statistics, *Statist. Sci.* 11, 229–243.
7. Tarpey T., Li L., Flury B. (1995). Principal Points and Self-Consistent Points of Elliptical Distributions. *Ann. Statist.* 23, 103–112.

Length and surface area estimation under convexity type assumptions

Beatriz Pateiro-López
Universidade de Santiago de Compostela, Spain, beatriz.pateiro@usc.es
Alberto Rodríguez Casal
Universidade de Santiago de Compostela, Spain, alrodcas@usc.es

Abstract. The reconstruction of a compact set $S \subset \mathbb{R}^d$ from a finite set of points taken in it is an interesting problem that have been addressed in many fields like computational geometry. The case where S is assumed to be convex has been extensively analyzed in the literature. In this case there exists a quite natural estimator of S: the convex hull of the sample. The perimeter and surface area of the convex hull of the sample can be successfully used for estimating the length and surface area of S. In this paper a less restrictive assumption on the set we want to estimate is considered. It is assumed that S and its complementary are both α-convex, which means that a ball of radius α can roll freely outside and inside the set. Under this assumption the α-convex hull of the sample is the natural estimator. We show that this estimator performs well not only as a support estimator but also as an estimator of the surface area of the set.

Functional data classification based on reproducing kernel regularization

Alberto Muñoz
Universidad Carlos III de Madrid, Spain, alberto.munoz@uc3m.es
Javier González
Universidad Carlos III de Madrid, Spain, javier.gonzalez@uc3m.es

Abstract. Functional data are difficult to manage for most traditional pattern recognition techniques, given the very high (or intrinsically infinite) dimensionality. The reason is that functional data are functions and most algorithms are designed to work with vectors.
In this paper we propose a functional analysis technique to obtain finite dimensional representations of functional data. The key idea is to consider each functional datum as a point in a general function space and then to project these points on to a Reproducing Kernel Hilbert subspace. For this aim we will use Regularization Theory, using the Tikonov approach and the quadratic loss function.
In addition, we show some theoretical properties of the method: The representation is continuous with respect to the original representation function and the finite sample approximation of the representation converges to the true representation in terms of the eigenfunctions of the covariance function that defines the Reproducing Kernel Space. We also explain the performance of the method when the covariance function is changed in several ways.
Regarding experimental work, we illustrate the performance of the proposed representation method in several classification and clustering problems of functional data sets. In particular, we show the performance of the proposed representation when we apply the procedure to cluster the Canadian temperature series data set, and compare the results to others obtained in the literature.

Smoothing parameter selection in hazard rate estimation

Francois Van Graan
North-West University, South Africa, Francois.VanGraan@nwu.ac.za

Abstract. Kernel type estimators of ratio functions such as the hazard rate have been studied by a number of authors. The kernel estimator of a hazard function involves evaluation of a cumulative distribution function estimate. The so-called internal estimator is obtained by convolution of a kernel and a cumulative hazard estimator. However, the practical use of the estimator heavily depends on the choice of the smoothing parameter. It is known that the least squares cross-validation bandwidth is asymptotically optimal in the case of hazard rate estimation in the settings of both complete and randomly right-censored samples. However the rate of convergence is slow. An alternative to the previous method is using a bandwidth selector based on the bootstrap. The purpose of this presentation is to compare the empirical performances of the different bandwidth selectors in hazard rate estimation for moderate sample sizes and different models.

A copula approach to conditional density estimation

Olivier P. Faugeras
Laboratoire de Statistique Théorique et Appliquée (LSTA), University Pierre et Marie Curie - Paris 6, France,
olivier.faugeras@gmail.com

Abstract. We present a new kernel estimator of the conditional density of a real valued random variable Y given $X = x$. It is based on an efficient transformation of the data by an empirical approximation of the quantile transform. By use of the copula representation, the estimator turns out to have an appealing product form, whereas competitors based either on the ratio of estimations of the joint and marginal densities or on nonparametric regression have a quotient shape. Thanks to this structure, its asymptotic properties are derived by simple combination of the results obtained on (unconditional) density estimation. In particular, a comparison of its asymptotic bias and variance and simulation evidence shows that the proposed estimator does not suffer from instability issues of its quotient shaped rivals, especially when their denominator is close to zero, e.g. for large values of x in the tails of the distribution of X.

A bootstrap comparison of generalized linear models against non-parametric alternatives with binomial stimulus-response data

Kamila Żychaluk
School of Electrical and Electronic Engineering, University of Manchester, UK, kamila.zychaluk@manchester.ac.uk

David H. Foster
School of Electrical and Electronic Engineering, University of Manchester, UK, d.h.foster@manchester.ac.uk

Abstract. Many experiments involve recording a binary response at different levels of a stimulus. The underlying stimulus-response function is usually represented parametrically as a generalized linear model (GLM) with the requirement that it should be monotonic with the stimulus level x. Monotonicity can be ensured by modelling the linear predictor as $\eta(x) = ax + b$. The shape of the stimulus-response function is then determined by the chosen link function, relating probability of success to linear prediction.

There is often little to guide the choice of the link function, and testing for its adequacy is thus crucial. Here, a test is explored using the difference in deviances D_{GLM} and D_{loc} between the GLM and a local linear model, that is, $S = D_{\mathrm{GLM}} - D_{\mathrm{loc}}$. This test is a special case of the generalized likelihood ratio test proposed in [1].

The difference in deviances for nested GLMs is distributed asymptotically as χ^2_{l-k}, where k, l are the degrees of freedom for the two models [3]. The same may also be true for local models [2], with degrees of freedom for the local model defined as the trace of the hat matrix. But these asymptotic results may not hold for finite samples, and the simulations undertaken here show that the significance level for such a test is not at its nominal level.

An alternative method investigated here is to calculate the p-values using the bootstrap method, with the bootstrap samples generated from the null parametric model. Simulations suggest that these bootstrap p-values have a uniform distribution if the null model is correct. The test also has good power for the alternatives considered.

1. Fan, J., Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test* 16, 409–444.
2. Hastie, T. J., Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.
3. McCullagh, P., Nelder, J. A. (1990). *Generalized Linear Models.* Chapman & Hall, London.

Nonparametric surveillance under weak dependence to detect structural changes

Ansgar Steland
Institute of Statistics, RWTH Aachen, Germany,
University, steland@stochastik.rwth-aachen.de

Abstract. Suppose we are given a series of observations $Y_1, Y_2, \ldots$ arriving sequentially, i.e., at the n^{th} time instance the data $Y_1, \ldots, Y_n$ are available for inference. We consider the problem to analyze nonparametrically whether there is a change in the mean when the observations behave as a random walk.

This is a problem with applications in many areas, for instance in engineering and econometrics. In engineering a departure from a (reference) signal f_0 may indicate that an information channel is no longer reliable or that the current compression has to be updated. In econometrics changes may indicate that the economic regime governing the data generating process has changed. Methods to deal with this (classic) problem based on kernel smoothers related to the Nadaraya Watson estimator have been studied in detail by the author for mixing processes (Steland, 2004, Metrika) and random walks with dependent increments (Steland, 2005, JTSA) under non-standard conditions. For the related problem to detect a change from integration to stationarity, and vice versa, we refer to (Steland, 2007, JSPI) and the references given there. In the latter papers there are also extensive references to the literature.

Local polynomial fitting is nowadays a common and well established approach to estimate nonparametrically conditional means. For asymptotic results under mixing conditions see, for instance, (Masry and Fan, 1997, Scand. J. Statist.)

In the present paper we study the local linear estimator under a different sampling model, namely the classic time series approach where observations are observed at equidistant time instances. We base our detection procedure on the canonical process associated to that estimator. Assuming weak dependent errors a functional central limit theorem is established. Our result implies a central limit theorem for the proposed detection rule, which can be used to design surveillance procedures.

We discuss applications of our results, which also cover certain discretely observed continuous time models based on Brownian motion. In addition, we also present new results on the related problem of nonparametric detection of mean changes in weak dependent time series, which can be applied to the analysis of sparse signals.

Adaptive maximum and proxy maximum probability estimation of multidimensional Poisson intensities

José Carlos Simon de Miranda
Institute of Mathematics and Statistics University of São Paulo, Brazil, simon@ime.usp.br

Abstract. We propose a non parametric methodology of estimation of the intensity for Poisson point processes on $\mathbb{R}^m$. We assume the observation region, $\mathcal{O}$, is a bounded $\mathbb{R}^m$ interval. The space of positive functions formed by composition of $L^2(\mathcal{O})$-functions with the exponential is endowed with a probability induced from another one defined on the set of wavelet coefficients. This is a convenient space for the intensity to belong to and we choose as our first estimate for the intensity the function that maximizes the posterior probability, given a trajectory of the Poisson process on $\mathcal{O}$, by means of determining the wavelet coefficients of its logarithm. A second estimate is obtained by suitably writing the posterior probability as a product of functions that are maximized separately giving raise to a proxy maximum posterior probability estimate. This approach presents the desired feature of furnishing everywhere non negative amplitude-scale invariant estimates of the intensity that exhibit not only a minimization of the energy, relative to the wavelet basis, but also a maximization of the entropy of the process on $\mathcal{O}$ conditional to the realization. A novel adaptive thresholding procedure based on jointly testing hypothesis, on the wavelet coefficients, and adjusting the priors' locations is given. We define exponentially decaying invariance and, as an example of the general estimating procedure above, we specialize to self affine and self similar probability prior Poisson processes.

Participants

Participants

A

B

C

D

H

I

K

L

M

N

O

P

R

S

T

V

Z

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Centro Internacional de Matemática

CMUC
Centre for Mathematics
University of Coimbra

FCTUC DEPARTAMENTO DE MATEMÁTICA

SPSS