Mathematisches Forschungsinstitut Oberwolfach

Report No. 49/2006

Qualitative Assumptions and Regularization in High-Dimensional Statistics

Organised by
Lutz Dümbgen (Bern)
Jon A. Wellner (Seattle)

November 5th – November 11th, 2006

Abstract. Important and exciting developments are currently underway in nonparametric statistics involving inter-play between qualitative constraints, penalization, and regularization methods. Some of these developments are taking place on the theoretical side (with connections in the direction of empirical process theory), while other parts of the development are occurring on the algorithmic and approximation theory sides. This workshop brought together researchers from several of these groups to exchange ideas and problems, to probe further research directions.

Mathematics Subject Classification (2000): primary: 62xx, secondary: 41Axx, 52Axx, 65Kxx, 90Cxx.

Introduction by the Organisers

This workshop was well-attended with 47 participants from Europe and overseas, among them many promising young scientists. While most participants are working in mathematical statistics, several participants are experts in approximation theory or fields of application such as astrophysics or econometrics, too. The participants exchanged ideas, discussed new developments and established new projects and interactions for the subsequent tasks.

Traditional nonparametric statistics and new trends. Nonparametric statistics has undergone dramatic changes during the last two decades. At first the focus shifted from permutation and rank testing in classical settings such as comparison of two univariate samples to multi- and even infinite-dimensional problems such as density estimation and regression. Here new results and techniques from empirical process theory, a very active research area in itself, played a prominent role.

At present statisticians working on nonparametrics are facing three kinds of problems, among others: On the one hand, the research focused strongly on point estimation, whereas in applications people are in need of tests and confidence sets. A second problem is the curse of dimensionality. Roughly speaking, the number of unknown parameters for reasonable approximating models grows exponentially with the dimension. This problem results in rather slow rates of convergence in higher dimensions. In addition, many estimation problems are inverse in the sense of involving indirect measurements and being ill-posed.

Qualitative Assumptions. For all three problems introducing qualitative assumptions is turning out to be a successful strategy with further potential. That means, in many situations, restrictions on the underlying function parameters such as e.g. monotonicity, concavity/convexity, or upper bounds on the number of local extrema may be used to enhance the performance of point estimators substantially and to replace quantitative smoothness assumptions which are difficult to justify. In addition, imposing such constraints enables the construction of nonparametric tests and confidence sets, sometimes even without relying on asymptotic expansions.

Computation and Regularization. In order to deal with the qualitative assumptions algorithmically, techniques for constrained optimization come into play. Sometimes it turns out that standard solutions from optimization theory such as, for instance, quadratic programming, are not efficient for statistical purposes, and alternative procedures such as the pool-adjacent-violators algorithm have been developed by statisticians. Naturally, regularization methods are used in this context, too. Regularization methods themselves are a well-known tool for treating inverse problems. In statistics they are also important in order to produce "sparse" estimators, i.e. estimators which are easier to interpret because of few non-zero parameters, few local extrema or other characteristics. In fact, in many fields of application the underlying parameter itself is assumed to be sparse, at least approximately. This is in fact the intrinsic reason why nonparametric curve estimation is possible at all.

Dimensional Asymptotics. More recently some authors showed how to use regularization successfully in regression problems with sparse parameters but of dimension p growing almost exponentially with the number n of observations. Considerations of this type are increasingly important, showing new trade-offs between flexibility of models and stability of estimation. One obvious example is the analysis of gene expression data, where the number of parameters (gene fragments) is in the range of a few hundred to several thousand, while sample sizes are rarely larger than a few hundred. Here approaches such as penalized logistic regression turn out to be very promising.

Informal Sessions. In addition to the regular talks (see the abstracts below), we organized two informal evening sessions with the following talks:
Arnold Janssen: Regions of alternatives with high and low power for goodness-of-fit tests,
Angelika Rohde: Adaptive goodness-of-fit tests based on signed ranks,

Melanie Birke: Estimating a convex function in nonparametric regression,
Kaspar Rufibach: The log–concave density estimator as a smoother,
Nicolai Bissantz: Nonparametric testing in noisy inverse problems.

Workshop: Qualitative Assumptions and Regularization in High-Dimensional Statistics

Table of Contents

Abstracts

The Hardy-Littlewood-Pólya monotone rearrangement algorithm

Dragi Anevski
(joint work with Anne-Laure Fougéres)

Sorting a set of discrete data points is an elementary operation. The Hardy-Littlewood-Pólya [6] monotone rearrangement algorithm can be seen as a continuous version of a sorting procedure for discrete data. Thus let $(x_1, \ldots, x_n)$ be a vector, or equivalently x be a function defined on the set $\{1, \ldots, n\}$. Then the decreasing rearrangement $\hat{x}$, defined by $\hat{x}(i) = x_{(i)}$ is given by the (generalized) inverse of $\tilde{x}(s) = \#\{i : x(i) > s\}$. If x is an arbitrary function on the interval $[0, 1]$, one could do the analogue approach but replace counting measure $\#$ with Lebesgue measure λ, so then $\hat{x}$ is defined as the (generalized) inverse of $z(s) = \lambda\{t \in (0, 1) : x(t) > s\}$. Let T_I denote the map which takes a function x on it's monotone arrangement $\hat{x}$ on an interval I.

The monotone rearrangement algorithm has mainly been used as a device in analysis, see e.g. [7]. It was only recently applied to statistical estimation problems by Fougerès [4], where the algorithm was introduced for density estimation.

We introduce the following general two-step approach for estimation of a monotone function: Assume that x is a function in some infinite-dimensional function class $\mathcal{X}$ for which one defining property is that the functions are monotone on an interval $I \subset \mathbf{R}$. Assume that x_n is an estimate of x that is smooth, such as e.g. a kernel estimator, but that is not necessarily monotone. We then propose to use the monotone rearrangement $\hat{x}_n$ of x_n as an estimate of x, and we derive the limit properties for this estimator. The applications of this are e.g. to density and regression function estimate, cf. [2].

A computational advantage of the proposed estimator is its simple use in practice: Starting with the estimator x_n evaluated at equi-spaced grid points $\tau_1, \ldots, \tau_n$, the resulting estimator is obtained as the (discrete) sorting of $x_n(\tau_1)$, $\ldots, x_n(\tau_n)$. Furthermore under the assumption of process weak limit distribution results for (a localized version of) the stochastic part of x_n and that the deterministic part of x_n is asymptotically differentiable at a fixed point t_0, with strictly negative derivative, it is possible to obtain point wise limit distribution results for the final estimate $\hat{x}_n(t_0)$. The estimator is consistent if the preliminary estimator is.

Possible applications of the general results are to monotone density and regression function estimation, although not limited to these cases. These are the

problems of estimating f and m respectively in

(i) $t_1, \ldots, t_n$ stationary observations with marginal decreasing density f on $\mathbf{R}^+$,

(ii) (t_i, y_i) observations from $y_i = m(t_i) + \epsilon_i$, $t_i = i/n, i = 1, \ldots, n, m$ decreasing on $[0, 1]$, $\{\epsilon_i\}$ stationary sequence with mean zero.

Standard approaches in these two problems have been isotonic regression for the regression problem, first studied by Brunk [3], and (nonparametric) Maximum Likelihood estimation for the density estimation problem, first introduced by Grenander [5]. Lately Anevski and Hössjer [1] gave a unified approch generalizing the results by Brunk and Grenander and others.

Using kernel estimators, denoted as x_n, as preliminary estimators of f and m one can apply the monotone rearrangement map on x_n to obtain a monotone function.

The general limit distribution result is obtained as follows: Assume that $\{x_n\}_{n\geq 1}$ is is a sequence of continuous stochastic processes and decompose

$$x_n(t) = x_{b,n}(t) + v_n(t), \tag{1}$$

with v_n the stochastic and $x_{b,n}$ the deterministic part of x_n. Given a sequence $d_n \downarrow 0$ and a point t_0 (in the interior of the support of x) define the rescaled determinstic and stochastic parts respectively as

$$\tilde{v}_n(s; t_0) = d_n^{-1}\{v_n(t_0 + sd_n) - v_n(t_0)\}, \tag{2}$$

$$g_n(s) = d_n^{-1}\{x_{b,n}(t_0 + sd_n) - x_{b,n}(t_0)\}. \tag{3}$$

The next two assumptions are the main tools to obtain limit distribution results. The first is a restriction on the process that is typically satisfied by a large class of processes, such as empirical processes and partial sum processes and smoothed versions of these, and states that the rescaled process converges weakly to a nontrivial limit process. The second is a set of properties that the monotone rearrangement map satisfies, and basically states invariant properties of the map under addition and multiplication of constants and monotonicity. This is also crucial for obtaining consistency of the resulting estimator and implies that consistency of x_n necessarily gives consistency for $T(x_n)$.

(*Rescaling for the preliminary estimator*) Assume that there exists a stochastic process $\tilde{v}(\cdot; t_0) \neq 0$ such that $\tilde{v}_n(s; t_0) \xrightarrow{\mathcal{L}} \tilde{v}(s; t_0)$ on $C(-\infty, \infty)$ as $n \to \infty$, and that the functions $\{x_{b,n}\}_{n\geq 1}$ are monotone such that $\sup_{|s|\leq c} |g_n(s) - As| \to 0$, and for each $c > 0$ and for some constant $A < 0$, as $n \to \infty$.

(*Properties of the monotone rearrangement map* T) The monotone rearrangement map T_I satisfies the following:

$$
\begin{aligned}
&(i) && T_I(f+c) = T_I(f) + c, \text{ if } c \text{ is a constant},\\
&(ii) && T_I(cf) = cT_I(f), \text{ if } c > 0 \text{ is a constant},\\
&(iii) && f \le g \Rightarrow T_I(f) \le T_I(g),\\
&(iv) && \text{Let } f_c(t) = f(ct). \text{ Then } T_I(f_c)(u) = T_{I\cdot c}(f)(u/c),\\
&(v) && \text{Let } f_c(t) = f(t+c). \text{ Then } T_I(f_c) = T_{I+c}(f),
\end{aligned}
$$

proved by Anevski and Fougeres [2]. Under these assumptions it is possible to derive the limit distribution result

$$
d_n^{-1}[T_J(x_n)(t_0) - x_n(t_0)] \quad \xrightarrow{\mathcal{L}} \quad T[As + \tilde{v}(s;t_0)](0), \tag{4}
$$

as $n \to \infty$, cf. [2]. It is possible to derive the result under general dependence assumptions, demanding essentially stationarity for the underlying random parts $\{\epsilon_i\}$ and $\{t_i\}$ respectively.

The properties satisfied by T are similar to properties satisfied by the slope of greatest convex minorant map, used in monotone regression and density estimation, cf. [1]. This possibly suggests a general tool for obtaining limit distributions for non-regular problems, analogous to the assumption of Hadamard differentiability of the functional for regular problems: namely assuming that the map T satisfies the above conditions. As a matter of fact, any map T possessing the above properties will give a limit distribution result analogous to (4), under the assumptions (2) and (3) on the rescaled preliminary estimator, and the resulting estimator $T(x_n)$ will be consistent if x_n is consistent.

An interesting future problem is to find other such maps, and to study similar maps for other estimation problems, such as e.g. estimation under convexity assumptions.

References

[1] D. Anevski and O. Hössjer, *A general scheme for inference under order restrictions*, Ann. Statist. **34** (4) (2006), 1874–1930.

[2] D. Anevski and A.-L. Fougères, *Limit properties for the monotone rearrangement algorithm with applications to monotone density and regression function estimation*, Preprint (2006).

[3] H.D. Brunk, *On the estimation of parameters restricted by inequalities*, Ann. Math. Statist. **29** (1958), 437–454.

[4] A.-L. Fougères, *Estimation de densités unimodales*, Canad. J. Statist. **25** (1997), 375–387.

[5] U. Grenander, *On the theory of mortality measurement. II*, Skandinavisk Aktuarietidskrift **39** (1956), 123–253.

[6] G.H. Hardy, J.E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press, (1988).

[7] E.H. Lieb and M. Loss, *Analysis*, Graduate Studies in Mathematics **32** (2001), American Mathematical Society.

Regularized Mean Estimators in the Multivariate Linear Model

Rudolf Beran

The least squares estimator for the mean response matrix in a multivariate linear model is known to perform poorly, in risk and in practice, when the rank of the design matrix equals or nearly equals the number of observed responses. A better regularized estimator of the multivariate means is constructed as follows:
(a) Express the least squares estimator as the sum of its orthogonal projections into subspaces determined by pertinent nested submodels of the multivariate linear model.
(b) Right multiply each summand by any symmetric matrix with eigenvalues in $[0, 1]$ to generate a class of affine shrinkage candidate estimators. These candidate estimators can be expressed as the closure of a set of affinely penalized least squares estimators.
(c) Find, in closed form, the candidate estimator that minimizes estimated quadratic risk.

The risk of this adaptive regularized estimator converges asymptotically to that of the candidate estimator with smallest quadratic risk. In the asymptotic theory, the rank of the design matrix tends to infinity while the number of observations equals or exceeds this rank. Examples of the adaptive regularized estimator include multivariate discrete spline estimators, multivariate submodel selection estimators, regularized MANOVA estimators, and a positive-part Efron-Morris estimator.

Iterated Regularization: Boosting and Twin Boosting for High-Dimensional Data

Peter Bühlmann

We present a statistical perspective on boosting. Special emphasis is given to estimating potentially complex parametric or nonparametric models, including generalized linear and additive models as well as regression models for survival analysis (cf. [1]).

The practical aspects of boosting procedures for fitting statistical models are illustrated by means of the open-source software package *R:mboost* ([3]).

Furthermore, we propose Twin Boosting ([2]) which has much better feature selection behavior than boosting. In addition, for cases with a few important effective and many noise features, Twin Boosting also substantially improves the predictive accuracy of boosting. Twin Boosting is as general and generic as boosting. It can be used with general weak learners and in a wide variety of situations, including generalized regression, classification or survival modeling.

References

[1] P. Bühlmann, *Boosting for high-dimensional linear models*, Annals of Statistics **34** (2006), 559–583.
[2] P. Bühlmann, *Twin Boosting: improved feature selection and prediction*, Preprint.

[3] P. Bühlmann and T. Hothorn, *Boosting: a statistical perspective*, Preprint.

Adaptive Convex Regression

Tony Cai and Mark Low

Consider the regression model

$$y_i = f(x_i) + \sigma z_i, \quad i = 0, 1, 2, ..., n$$

where $x_i = i/n$ and $z_i \overset{iid}{\sim} N(0,1)$ and the problem of estimating and providing confidence intervals for $f(x_*)$ where $x_* = i_*/n$ is a fixed point in the interval $(0,1)$. In this talk we consider the case where the regression function f is assumed to be convex.

For each j we introduce a linear estimator

$$\delta_j = \sum_{k=-2^{j-1}}^{2^{j-1}} w_{j,k}(y_{i_*-k} + y_{i_*+k})$$

where $w_{j,k}$ are all non negative and sum to 1.

An oracle "bandwidth"

$$j_{oracle} = \operatorname*{argmin}_j E_f(\delta_j - f(x_*))^2$$

is defined and related to the oracle risk

$$R_{oracle}(f) = E_f(\delta_{j_{oracle}} - f(x_*))^2.$$

These provide a natural goal for the performance of a data driven selection procedure. For $1 \le j \le J$ let

$$\bar{\delta}_j = 2^{-j} \sum_{k=1}^{2^{j-1}} (y_{i_*-k} + y_{i_*+k})$$

and set

$$T_j = \bar{\delta}_j - \bar{\delta}_{j-1}.$$

These T_j can be used as tests to empirically choose the "best" j since T_j gives an estimate of the absolute bias of $\bar{\delta}_{j+1}$. The selection rule is given as

$$\hat{j} = \operatorname*{argmin}_{1 \le j \le J} \left\{ T_j^2 + \sigma^2 2^{-j} \right\}.$$

The estimator of $f(x_*)$ is then defined as

$$\hat{f}(x_*) = \delta_{\hat{j}}.$$

and a $(1-\alpha)$-level confidence interval for $f(x_*)$ is given by

$$CI_\alpha = [\delta_{\hat{j}} - c_1 z_{\alpha/2} \sigma 2^{-\hat{j}/2},\ \delta_{\hat{j}} + c_2 z_{\alpha/2} \sigma 2^{-\hat{j}/2}].$$

The performance of these procedures are discussed in terms of two oracle bounds, a linear oracle bound and a two-function oracle bound. For the linear

oracle bound consider the class of all symmetric, nonnegative linear estimators with weights adding to 1, $\hat{f}_L = \sum c_i y_i$ where $c_i \geq 0$, $\sum c_i = 1$, $c_{i_*+j} = c_{i_*-j}$. The linear oracle bound is then given by

$$R_n(f) = \inf_{\hat{f}_L} E(\hat{f}_L - f(x_*))^2.$$

The two-function oracle bound is built from the true unknown convex function, say f_0 and a single alternative and is given by

$$R_n^*(f_0) = \sup_{f_1 \in \mathcal{F}} \inf_{\hat{f}} \max_{i=0,1} E_{f_i}(\hat{f} - f_i(x_*))^2.$$

We show that

- $E(\hat{f}(x_*) - f(x_*))^2 \leq CR_n(f)$.
- $E(\hat{f}(x_*) - f(x_*))^2 \leq CR_n^*(f)$.

For confidence intervals, let $L_n(f)$ be that minimum expected length for confidence intervals of $f(x_*)$ which have coverage probability of at least $1 - \alpha$ over the collection of convex functions. Then our confidence procedure not only has a given level of coverage probability but also satisifies $EL(CI) \leq CL_n(f)$ where C is a small constant.

Asymptotics, shape regularization and local adaptivity

Patrick Laurie Davies

We consider firstly the standard non–parametric regression model

$$Y(t) = f(t) + \sigma Z(t),\ t \in [0,1]. \tag{1}$$

Given a sample $\boldsymbol{Y}_n = \{(t_i, Y(t_i)) : I = 1, \ldots, n\}$ of size n the standard measure of performance for an estimator $\hat{f}_n$ based on $\boldsymbol{Y}_n$ is the mean integrated squared error MISE

$$MISE = \boldsymbol{E}(\|\hat{f}_n - f\|_2^2)$$

and its rate of convergence to zero. procedures which attain or almost attain the optimal rate of convergence are regarded as superior to procedures which do not have this property. The attainable rate of convergence depends on the smoothness of f which is taken to be unknown. Some procedures such as wavelets are know to be able to adapt to the unknown smoothness and hence to automatically converge faster for smoother functions (Donoho and Johnstone [2]). Other methods such as the taut string of Davies and Kovac [3] do not have this adaptivity property. As an example we take the infinitely differentiable function $f_0(t) = \sin(2\pi t)$. The Daubechies least-asymmetric orthonormal compactly supported wavelet with 10 vanishing moments satisfy a Hölder condition of a least 2.9 and hence the rate of convergence of the MISE is at least $n^{-0.8529}$. The taut string procedure has a rate of convergence of at best $n^{-2/3}$ and simulations show that for sample of size up to

$4 \cdot 10^6$ the wavelets have a superior performance in terms of MISE. We used Nason [4] to calculate the wavelet reconstruction. If however we put

$$f_1(t) = \sin(2\pi t) + 0.5 \exp(-5000(t - 1/2)^2) \tag{2}$$

then the rates of convergence are not altered but for finite sample the roles are almost reversed with the taut string performing better for sample sizes of between $16 \cdot 10^4$ and $4 \cdot 10^6$. An examination of individual results indicates that the small bump in (2) effects all those wavelet coefficients for which the wavelet contains the bump which in turn effects the reconstruction away from the bump. This is not the case for the taut string where the effect of the bump on the reconstruction is essentially limited to the bump itself. A possible explanation is that the taut string is a form of shape regularization. To analyse the shape regularization we define a universal, honest and non-asymptotic region for f based on a sample $\boldsymbol{Y}_n$ as follows. We put

$$w(\boldsymbol{Y}_n, g, I) = \frac{1}{\sqrt{|I|}} \sum_{t_i \in I} (Y(t_i) - g(t_i))$$

and

$$\mathcal{A}(\boldsymbol{Y}_n, \sigma_n, \tau_n) = \{g : \max_{I \subset [0,1]} |w(\boldsymbol{Y}_n, g, I)| \leq \sigma_n \sqrt{\tau_n \log n}\}$$

where

$$\sigma_n = \frac{1.4826}{\sqrt{2}} \text{median}\,(|Y(t_2) - Y(t_1)|, \ldots, |Y(t_n) - Y(t_{n-1})|).$$

and it can be shown that for an appropriate value of τ_n which depends only on n and α the region $\mathcal{A}_n$ is a universal, honest and non-asymptotic region for f with coverage probability of at least α. If we now regularize in $\mathcal{A}_n$ by minimizing the number of local extremes then it can be shown that the local rate of convergence of resulting estimator to f at the point t depends only on the behaviour of f in a small neighbourhood of t. Moreover the estimator converges automatically at the rate $(\log n/n)^{-1/2}$ on intervals where f is constant and at $(\log n/n)^{-1/3}$ intervals where f is monotone. Corresponding results hold if the number of intervals of convexity or concavity is minimized. In the case of wavelets the effects disappear as n tends to infinity and cannot therefore be demonstrated asymptotically if the function f in (1) is kept fixed. In order to overcome this it is intended to use an idea of Dahlhaus [1] and to consider a sequence of models

$$Y_n(t) = f_n(t) + \sigma Z(t),\ t \in [0, 1]. \tag{3}$$

where the f_n become increasingly complex as n increases. First simulations indicate that in this situation shape regularized procedures with a slow rate of convergence can be consistent $\lim_{n\to\infty} \boldsymbol{E}(\|\hat{f}_n - f_n\|_2^2) = 0$ whilst others with a faster rate of convergence for fixed f may not be consistent $\liminf_{n\to\infty} \boldsymbol{E}(\|\hat{f}_n - f_n\|_2^2) > 0$..

References

[1] R. Dahlhaus, *Small sample effects in time series analysis: a new asymptotic theory and a new estimate*, Ann. Statist. **16** (1988), 808–841.

[2] D.L. Donoho and I.M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc. **90** (1995), 1200–1224.
[3] P.L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (with discussion)*, Ann. Statist. **29** (2001), 1–65.
[4] G.P. Nason, *WaveThresh3 Software*, Dept. of Mathematics, University of Bristol, Bristol, UK (1998).

Simple Monotone Regression

Holger Dette
(joint work with Natalie Neumeyer, Kay Pilz)

Consider the common nonparametric regression model

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

where $\{(X_i, Y_i)\}_{i=1}^n$ is a bivariate sample of i.i.d. observations such that X_i has a positive two times continuously differentiable density f with compact support, say $[0,1]$. The variance function $\sigma : [0,1] \to \mathbb{R}^+$ and the regression function $m : [0,1] \to \mathbb{R}$ are assumed to be continuous and two times continuously differentiable, respectively and the regression function m is strictly increasing. We define for $N \in \mathbb{N}$

$$\hat{m}_I^{-1}(t) := \frac{1}{Nh_d} \sum_{i=1}^{N} \int_{-\infty}^{t} K_d\Big(\frac{\hat{m}(\frac{i}{N}) - u}{h_d}\Big) du$$

as an estimate of $m^{-1}(t)$, where

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_r\Big(\frac{X_i - x}{h_r}\Big) Y_i}{\sum_{i=1}^n K_r\Big(\frac{X_i - x}{h_r}\Big)}$$

is the classical Nadaraya-Watson estimate [see Nadaraya (1964) or Watson (1964)], K_d and K_r denote symmetric kernels with compact support, say $[-1,1]$, existing second moment and h_d, h_r are the corresponding bandwidths converging to 0 with increasing sample size n. Note that $\hat{m}_I^{-1}(t)$ is strictly increasing and consequently the same is true for its inverse denoted by $\hat{m}_I$. This statistic is a strictly isotone and smooth estimate of the regression function m.

Under the certain assumptions of regularity the following results can be proved if $n, N \to \infty$:

- If $\lim_{n\to\infty} h_r/h_d = \infty$, then for all $t \in (m(0), m(1))$ with $m'(m^{-1}(t)) > 0$,

$$\sqrt{nh_r}\Big(\hat{m}_I^{-1}(t) - m^{-1}(t) + \kappa_2(K_r) h_r^2 \Big(\frac{m''f + 2m'f'}{fm'}\Big)(m^{-1}(t))\Big) \stackrel{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \tilde{g}^2(t)),$$

where $\kappa_2(K_r) = \int u^2 K_r(u) du/2$ and the asymptotic variance is given by

$$\tilde{g}^2(t) = \frac{\sigma^2(m^{-1}(t))}{\{m'(m^{-1}(t))\}^2 f(m^{-1}(t))} \int K_r^2(u) du.$$

- If $\lim_{n\to\infty} h_r/h_d = \infty$, then for every $t \in (0,1)$ with $m'(t) > 0$,

$$\sqrt{nh_r}\Big(\hat{m}_I(t) - m(t) - \kappa_2(K_r)h_r^2\Big(\frac{m''f + 2m'f'}{f}\Big)(t)\Big) \stackrel{\mathcal{D}}{\Rightarrow} \mathcal{N}(0, \tilde{s}^2(t)),$$

where the asymptotic variance is given by

$$\tilde{s}^2(t) = \frac{\sigma^2(t)}{f(t)} \int K_r^2(u)du.$$

References

[1] H. Dette, N. Neumeyer, K.F. Pilz, *A simple nonparametric estimator of a monotone regression function*, Bernoulli **12** (2006), 469–490.

[2] H. Dette, N. Neumeyer, K.F. Pilz, *A note on nonparametric estimation of the effective dose in quantal bioassay*, J. Amer. Statist. Assoc. **100** (2005), 503–510.

[3] E.A. Nadaraya, *On estimating regression*, Theory of Probability and its Applications **15** (1964), 134–137.

[4] G.S. Watson, *Smooth regression analysis*, Sankya, Ser. A **26** (1964), 359–372.

An Empirical Comparison of Nonparametric Regression Methods

Ursula Gather

(joint work with P. Laurie Davies, Henrike Weinert)

Nonparametric regression can be considered as a problem of model choice. We present the results of a simulation study in which several nonparametric regression techniques are compared with respect to their behaviour on different test beds.

Consider paired data $\mathcal{Y}_n = \{(t_i, y(t_i))\}_{i=1}^n$ where the design points are ordered $0 \le t_1 < \cdots < t_n \le 1$ but not necessarily equidistant. The problem is to use the data to derive a function f_n which can be regarded as an adequate denoised representation of the data. The model we assume for the data is

$$Y(t_i) = f(t_i) + \sigma\epsilon(t_i), \quad i = 1, \ldots, n, \tag{1}$$

which represents a *signal* f corrupted by *noise* ϵ which we take to be standard Gaussian white noise. In the context of nonparametric regression the problem of model choice becomes: estimate f by a function $f_n^* \in \mathcal{F}$ that minimizes an expected distance or risk:

$$\mathrm{E}\left[d\left(f, f_n^*\right)\right] = \inf_{f_n \in \mathcal{F}} \mathrm{E}\left[d\left(f, f_n\right)\right], \tag{2}$$

where $\mathcal{F}$ is some specified class of functions and $d(\cdot,\cdot)$ is an appropriate loss function. In addition some model selection rules require the optimization in (2) to be conducted under constraints, whereby some measure of the complexity of a model is included in the term to be minimized. We review the following methods for signal approximation: wavelet regression (WH for hard and WS for soft thresholding, Donoho and Johnstone, 1994), the unbalanced haar method (UH, Fryzlewicz, 2006), minimum description length denoising (MDL, Rissanen, 2000),

a kernel plug-in estimator (PL, Herrmann, 1997) and adaptive weights smoothing (AWS, Polzehl and Spokoiny, 2000, 2003).

Another approach to nonparametric regression is based on the concept of *data approximation* described by Davies (1995, 2003). Although the concept makes use of properties of the model it does not operate solely within it but poses the question as to whether the model can be regarded as an adequate approximation to the data. Risk minimization such as (2) is not involved nor does it make assumptions about the existence of a true underlying function f. The model with parameters (f_n, σ_n) is regarded as an adequate approximation if typical data generated under the model *look like* the observed data $\mathcal{Y}_n$. Within the set of parameter values (f_n, σ_n) which give an adequate approximation we then select an f_n which minimizes one or more measures of complexity. The "taut-string" (TS and TV) nonparametric regression method of Davies and Kovac (2001) is an example of this idea. The measure of complexity is the number of peaks. The definition of approximation is based on the residuals.

The test beds we use are those introduced in Donoho and Johnstone (1994) and which are known as Blocks, Bumps, Heavisine and Doppler. We also include a constant signal as well as a heavily oscillating sine-function which terminates with a constant.

The loss functions we consider are the empirical versions $d_2(f, g)$ and $d_\infty(f, g)$ of the L_2– and L_∞–norms (see Donoho and Johnstone, 1994). For any given test bed with function f and for any given procedure resulting in some f_n the measures of performance are the average values of $d_2(f, f_n)$ and $d_\infty(f, f_n)$ over the simulations.

We also introduce a new loss which measures how well the extremes (e.g., peaks and troughs) of an estimate f_n match those of the test signal f. There are two possible errors. The reconstruction f_n can fail to have a local extreme of the correct type at a point where a target signal f exhibits one. The second type of error is that f_n exhibits a local extreme at a point where the test bed function does not have one. We propose a peak identification loss (PID):

$$\begin{aligned} PID(f, f_n) &= \mathrm{sgn}(\mathrm{n_{est}} - \mathrm{n_{extr}})\,((\mathrm{n_{extr}} - \mathrm{n_{id}}) + (\mathrm{n_{est}} - \mathrm{n_{id}})) \\ &= \mathrm{sgn}(\mathrm{n_{est}} - \mathrm{n_{extr}})\,(\mathrm{n_{extr}} + \mathrm{n_{est}} - 2\mathrm{n_{id}})\,, \end{aligned} \tag{3}$$

where the counts $(n_{extr} - n_{id})$ and $(n_{est} - n_{id})$ measure the extent of the two errors described above, since n_{extr} denotes the number of local extremes of the signal function f, n_{est} the number of local extremes of a reconstruction f_n of f, and n_{id} the number of local extremes of f that are correctly identified by f_n. We use $\mathrm{sgn}(\mathrm{n_{est}} - \mathrm{n_{extr}})$ so that it is possible to see if too many (positive sign) or too few (negative sign) local extremes are identified.

The count n_{id} is the number of correctly identified local extremes by a reconstruction f_n.

We summarize the results of the simulation studies as follows:

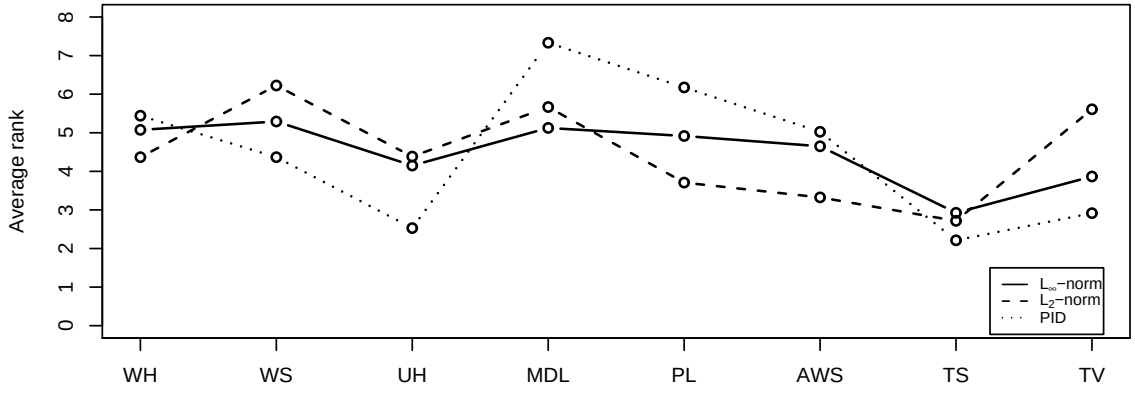

FIGURE 1. Average rank of the eight methods.

- With the exception of TS and TV the performance with respect to the peak identification deteriorates as the sample size increases.
- The MDL method often produces too many local extremes.
- The reconstructions produced by kernel and wavelet methods (WH, WS, AWS and PL) often fails to reproduce the magnitudes of the peaks for the Bumps function.
- As the Blocks function is piecewise constant all the methods apart from UH, TS and TV perform poorly as they are designed to give smooth reconstructions. UH performs very well on the Blocks function.
- There are two types of behaviour for the sine function. Either the signal is not recognized at all or the reconstruction is reasonable.
- The MDL method performs extremely badly on the white noise test bed.
- Overall TS performs the best. This can also be seen in Figure 1: TS has the smallest average rank for all performance measures.

REFERENCES

[1] P.L. Davies, *Data features*, Statistica Neerlandica **49** (1995), 185–245.
[2] P.L. Davies, *Approximating data and Statistical procedures – I. Approximating data*, Technical Report 7/2003, SFB 475, Dept. of Statistics, Univ. of Dortmund, Germany.
[3] P.L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (with discussion and rejoinder)*, Ann. Statist. **29** (2001), 1–65.
[4] D.L. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), 425–455.
[5] P. Fryzlewicz, *Unbalanced Haar technique for nonparametric function estimation*, Preprint (2006).
[6] E. Herrmann, *Local bandwidth choice in kernel regression estimation*, J. Comp. Graph. Statist. **6** (1997), 35–54.
[7] J. Polzehl and V.G. Spokoiny, *Adaptive weights smoothing with applications to image restoration*, J. Royal Statist. Soc. B **62** (2000), 335–354.

[8] J. Polzehl and V.G. Spokoiny, *Varying coefficient regression modeling*, Preprint (2003).
[9] J. Rissanen, *MDL Denoising*, IEEE Trans. Information Theory **46** (2000), 2537–2760.

Convex Hulls of Samples in $\mathbb{R}^d$

Piet Groeneboom

In [4] a central limit theorem for the number of vertices N_n of the convex hull of a uniform sample from the interior of convex polygon was derived. To be more precise, it was shown that $\{N_n - \frac{2}{3}r\log n\}/\{\frac{10}{27}r\log n\}^{1/2}$ converges in law to a standard normal distribution, if r is the number of vertices of the convex polygon from which the sample is taken. This paper also gives a central limit theorem for uniform samples in the interior of a circle (which can be extended to a central limit theorem for uniform samples from convex figures with a smooth boundary), where both the asymptotic expectation and the variance of the number of vertices are of order $n^{1/3}$ instead of order $\log n$. Somewhat remarkably, these different rates can be conjectured on the basis of results on the concave majorant of Brownian motion *without* drift and *with* (negative) parabolic drift, given in [3] and [5], respectively.

In the unpublished preprint [6] a central limit result for the joint distribution of N_n and A_n is given, where A_n is the area of the convex hull, using a coupling of the sample process near the border of the polygon with a Poisson point process as in [4], and representing the remaining area in the Poisson approximation as a union of a doubly infinite sequence of independent standard exponential random variables.

We derive this representation from the representation in [4] and also prove the central limit result of [6], using this representation. The relation between the variances of the asymptotic normal distributions of number of vertices and the area, established in [6] corresponds to a relation between the actual sample variances of N_n and A_n in [2]. Moreover, in [1] an exact formula for the number of vertices of the convex hull of a uniform sample from a triangle is announced, which has as corollary an asymptotic formula for the variance of N_n, corresponding to the scaling in the central limit theorem of [4]. I will briefly discuss the relation of these results, announced in [1], with the results in [4] and [6].

If time permits, I will also discuss the generalization of these results to convex hulls of samples of points, generated by probability distributions on $\mathbb{R}^d$, $d > 2$.

References

[1] C. Buchta, *On the distribution of the number of vertices of a random polygon*, Anzeiger Abt. II (Math.-Naturwiss. Klasse II, Österreichische Akademie der Wissenschaften) **139** (2004), 17–19.
[2] C. Buchta, *An identity relating moments of functionals of convex hulls*, Discrete and Computational Geometry **33**, 125–142.
[3] P. Groeneboom, *The concave majorant of Brownian motion*, Annals of Probability **11** (1983), 1016–1027.
[4] P. Groeneboom, *Limit theorems for convex hulls*, Probab. Theory Rel. Fields **79** (1988), 327–368.

[5] P. Groeneboom, *Brownian motion with a parabolic drift and Airy functions*, Probab. Theory Rel. Fields **81** (1989), 79–109.
[6] A.V. Nagaev, I.M. Khamdamov, *Limit theorems for functionals of random convex hulls* (in Russian), Preprint of Institute of Mathematics, Academy of Sciences of Uzbekistan. Tashkent (1991).
[7] A.V. Nagaev, *Some properties of convex hulls generated by homogeneous Poisson point processes in an unbounded convex domain*, Ann. Inst. Statist. Math. **47**, 21–29.

Total Variation Regularization for Bivariate Density Estimation

Roger Koenker
(joint work with Ivan Mizera)

L_1 penalties have proven to be an attractive regularization device for non-parametric regression, image reconstruction, and model selection. For function estimation, L_1 penalties, interpreted as roughness of the candidate function measured by their total variation, are known to be capable of capturing sudden changes in the target function while still maintaining a general smoothing objective. We explore the use of penalties based on total variation of the estimated density, its square root, and its logarithm – and their derivatives – in the context of univariate and bivariate density estimation, and compare the results to some other density estimation methods including L_2 penalized likelihood methods. Connections to maximum entropy and taut string methods can be established via conjugate duality methods.

Our objective is to develop a unified approach to total variation penalized density estimation offering methods that are: capable of identifying qualitative features like sharp peaks, extendible to higher dimensions, and computationally tractable. In bivariate settings we focus on discretizations in which log densities are represented by piecewise linear functions on Delone triangulations, where total variation of candidate functions is easily evaluated by summing gradient gaps along edges of the triangulation. Modern interior point methods for solving convex optimization problems play a critical role in achieving the final objective, as do piecewise linear finite element methods that facilitate the use of sparse linear algebra.

References

[1] R. Koenker and I. Mizera, (2006), Density Estimation by Total Variation Regularization, *A Festschrift for Kjell Doksum*, World Scientific: Singapore.

Curves and Modality

Arne Kovac

Given noisy bivariate observations $(x_i, y_i), i = 1, \ldots, n$ at n different time points $t_1, \ldots, t_n$ we consider the problem of specifying a smooth curve $f = (f^X, f^Y)$ such that f approximates the data and is simple in the sense that the number of local extreme values in the curvature function is as small as possible. In Figure 1 the

top left panel shows a spiral with added bivariate Gaussian noise and the right panel a reconstruction obtained from a kernel estimator. The curve is smooth, but does not approximate the data very well as can be seen in the bottom panel where the residuals in x- and y-direction are plotted.

We adopt a bivariate version of the multiresolution criterion by Davies and Kovac (2001) and require the sums of the residuals in x- and y-direction over stretches of different sizes and locations all to be smaller than what would be expected from white noise. More specifically we require an approximation to satisfy simultaneously

$$\left|\sum_{i\in I}(y_i - f_i^y)\right| < w_I \cdot \sigma, \quad \left|\sum_{i\in I}(x_i - f_i^x)\right| < w_I \cdot \sigma$$

with $w_I = \sqrt{|I| \cdot 2\log(2n)}$ for all intervals I of some family $\mathcal{I}$ of subintervals of $\{1, \ldots, n\}$.

One choice for $\mathcal{I}$ is to take all possible subintervals

$$\mathcal{I}_1 = \{\ \{j, j+1, \ldots, k\} \text{ for all } 1 \le j \le k \le n\}.$$

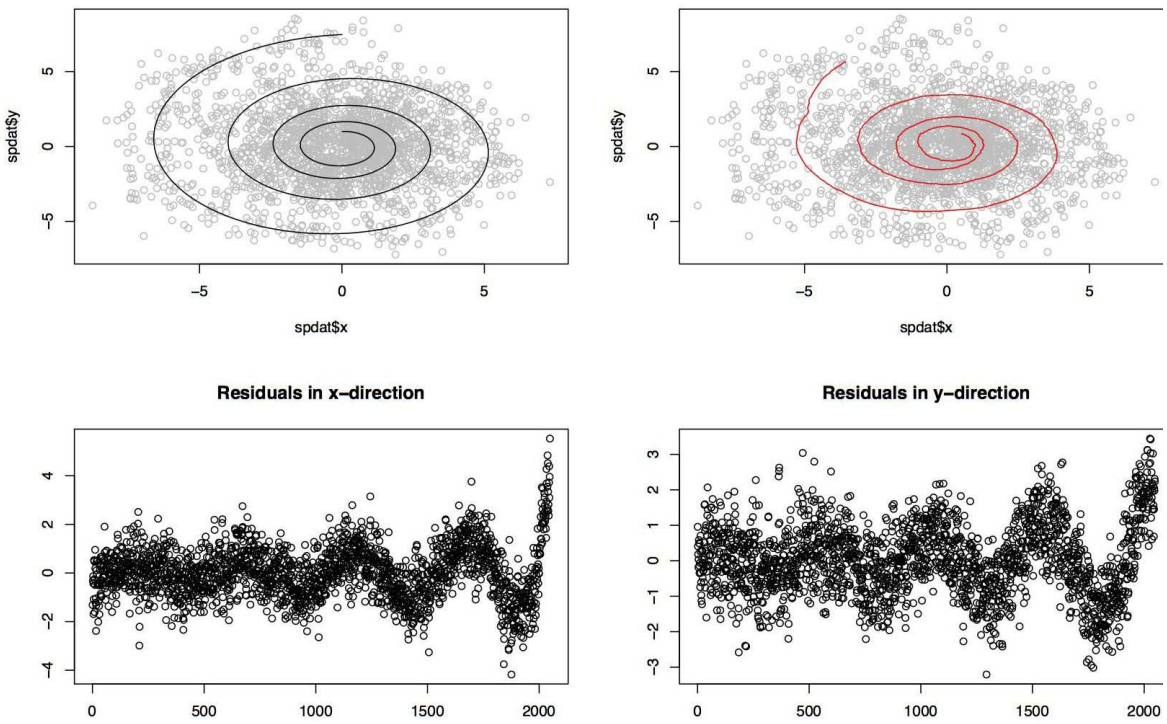

FIGURE 1. Noisy spiral and kernel estimator. Top left: Original spiral, Top right: Kernel estimator, Bottom left and right: Residuals in x- and y-direction

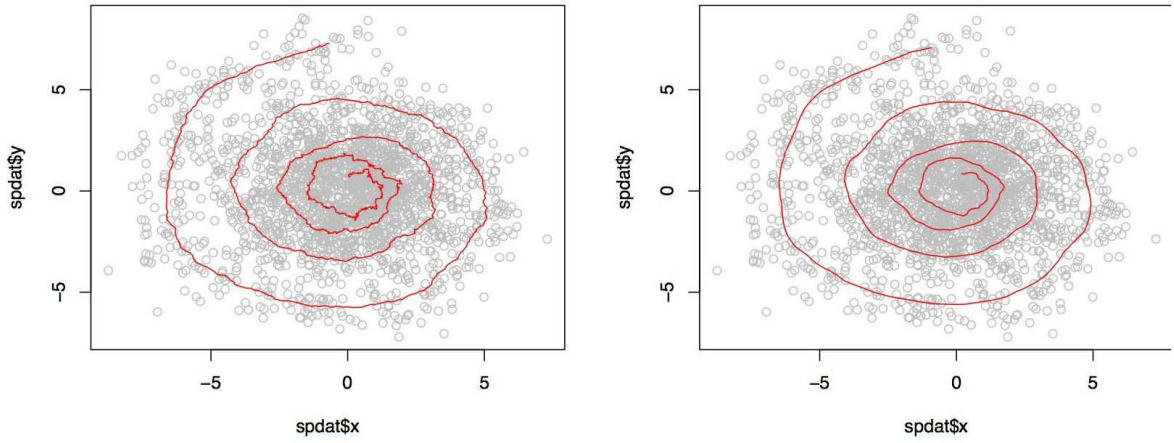

FIGURE 2. Noisy spiral and two approximations that just satisfy the multiresolution criterion. Left: Kernel estimator, Right: Total variation penalty

Computational complexity can be reduced by considering a smaller collection like all intervals with dyadic end points

$$\mathcal{I}_2 = \{\ \{2^j k + 1, \ldots, 2^j(k+1)\} \text{ for all } 0 \le j \le \lfloor \log_2(n) \rfloor, k = 0, 1, \ldots, \lceil \frac{n}{2^j} \rceil \}.$$

This collection has been used for the examples below. The multiresolution criterion requires the true value of σ. This may be estimated from the data by putting

$$\sigma = \frac{1.4826}{\sqrt{2}} \text{median}\,(|y_2 - y_1|, |x_2 - x_1|, \ldots, |x_n - x_{n-1}|)$$

(Davies and Kovac, 2001; Donoho et al, 1995)

We aim to find a curve f that satisfies this multiresolution criterion and is at the same time as simple as possible. Figure 2 shows in its left panel another approximation from an kernel estimator, but this time choosing the largest bandwidth such that the kernel estimate satisfies the multiresolution criterion above. Although this estimate approximates the data much better it contains a large number of spurious local extreme values.

In the univariate setting of non-parametric regression regularisation techniques based on total variation like the taut string method (Mammen and van de Geer, 1997; Davies and Kovac, 2001) and its generalisations (Dümbgen and Kovac, 2005) or quantile regression using total variation penalties (Koenker et al, 1994) have been shown to produce simple estimates.

We consider a two-dimensional total variation penalty and consider minimising the functional

$$T(f) = \sum_{i=1}^{n} (x_i - f_i^X)^2 + \sum_{i=1}^{n} (y_i - f_i^Y)^2 + \sum_{i=1}^{n-1} \lambda_i \sqrt{(f_{i+1}^X - f_i^X)^2 + (f_{i+1}^Y - f_i^Y)^2}$$

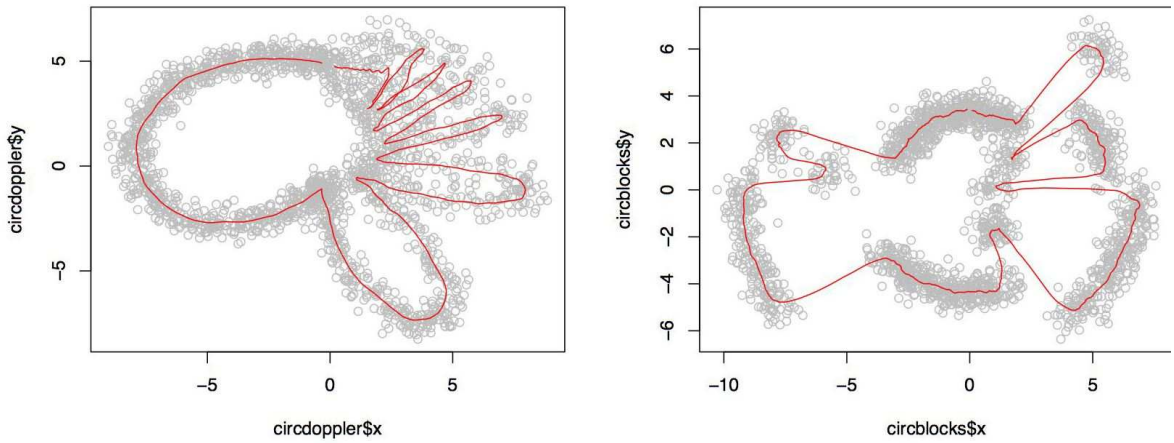

FIGURE 3. Circular noisy versions of Donoho and Johnstone's Doppler and Bumps signals and approximations using total variation penalties and the multiresolution criterion.

The smooth taut string functional by Kovac (2006) can be regarded as a special case of this functional in one dimension.

In order to obtain approximations that are as smooth and simple as possible we start with a large global penalty $\lambda_1 = \cdots = \lambda_{n-1}$ and successively reduce λ on intervals where the multiresolution criterion is not yet satisfied. This local squeezing technique has been described by Davies and Kovac (2001) and Dümbgen and Kovac (2005) in more detail. The application of this technique to the spiral data can be seen in the right panel of Figure 2. The approximation is much smoother than the kernel estimate.

Finally, Figure 3 shows approximations obtained from circular versions of the well known Doppler and Blocks functions by Donoho and Johnstone (1994). These were generated as $x_j = \cos(2\pi j/n)r_j$ and $y_j = \sin(2\pi j/n)r_j$ where $r_j = f(j/n) - \min_i(f(i/n), i = 1, \ldots, n) + 1$ and where f was successively the Doppler and the Blocks signal. The bivariate total variation penalties generate sharp discontinuities for the Blocks signal while the functions look smooth and simple elsewhere and approximate the data very well.

REFERENCES

[1] P. L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (with discussion)*, Ann. Statist. **29** (2001), 1–65.
[2] D. L. Donoho, I. .M. Johnstone, G. Kerkyacharian and D. Picard, *Wavelet shrinkage: asymptopia?*, J. Royal Statist. Soc. B **57** (1995), 371–394.
[3] D. L. Donoho and I. .M. Johnstone, *Ideal spatial adaption by wavelet shrinkage*, Biometrika **81** (1994), 425–455.
[4] L. Dümbgen and A. Kovac, *Extensions of Smoothing via Taut Strings*, Technical report.
[5] R. Koenker, P. Ng and S. Portnoy, *Quantile Smoothing Splines*, Biometrika **81** (1994), 673–680.
[6] A. Kovac, *Smooth functions and local extreme values*, Computational Statistics and Data Analysis (to appear).

[7] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist. **25** (1997), 387–413.

Limiting Distribution of the MLE for Current Status Data with Competing Risks

Marloes H. Maathuis

(joint work with Piet Groeneboom, Jon A. Wellner)

We study current status data with competing risks. Such data arise naturally in cross-sectional studies with several failure causes. Moreover, generalizations of these data arise in HIV vaccine trials (see [3]). The general framework is as follows. We analyze a system that can fail from K competing risks, where $K \in \mathbb{N}$ is fixed. The random variables of interest are (X, Y), where $X \in \mathbb{R}$ is the failure time of the system, and $Y \in \{1, \ldots, K\}$ is the corresponding failure cause. We cannot observe (X, Y) directly. Rather, we observe the 'current status' of the system at a single random time $T \in \mathbb{R}$, where T is independent of (X, Y). This means that at time T, we observe whether or not failure occurred, and if and only if failure occurred, we also observe the failure cause Y.

We consider nonparametric estimation of the sub-probability distribution functions $F_{0k}(s) = P(X \leq s, Y = k)$, $k = 1, \ldots, K$. The sub-distribution functions are related to each other, since their sum is the overall failure time distribution, i.e., $\sum_{k=1}^{K} F_{0k}(s) = P(X \leq s)$. Thus, we need to estimate a *system* of functions.

This problem, or close variants thereof, has been recently studied by [3], [4] and [5]. These papers introduce various nonparametric estimators, provide algorithms to compute them, and show simulation studies that compare them. However, until now, very little was known about the large sample properties of the estimators.

We have started to fill this gap by developing the local asymptotic theory for two estimators: the nonparametric maximum likelihood estimator (MLE) and the 'naive estimator' of [5]. We study the MLE because it is a natural estimator that often exhibits good behavior. The naive estimator is a simpler estimator that was suggested in the literature, and we consider it for comparison.

In [1] we prove that both the MLE and the naive estimator are consistent and converge globally and locally at rate $n^{1/3}$. We also show that the local rate of convergence is optimal in a minimax sense. The proof of the local rate of convergence of the MLE uses new methods, and relies on a rate result for the sum of the MLEs of the sub-distribution functions which holds uniformly on a fixed neighborhood of a point.

In [2] we use these rate of convergence results to derive the local limiting distributions of the estimators. The limiting distribution of the naive estimator is given by the slopes of the convex minorants of K correlated Brownian motion processes with parabolic drifts. The limiting distribution of the MLE involves a *new self-induced* process. We prove that this process exists and is almost surely unique. Finally, we present a simulation study showing that the MLE is superior

to the naive estimator in terms of mean squared error, both for small sample sizes and asymptotically.

References

[1] P. Groeneboom, M.H. Maathuis, J.A. Wellner, *Current status data with competing risks: consistency and rates of convergence of the MLE*, Technical Report 500 (2006), Dept. of Statistics, Univ. of Washington, Submitted.

[2] P. Groeneboom, M.H. Maathuis, J.A. Wellner, *Current status data with competing risks: limiting distribution of the MLE*, Technical Report 501 (2006), Dept. of Statistics, Univ. of Washington, Submitted.

[3] M.G. Hudgens, G.A. Satten, I.M. Longini, *Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation*, Biometrics **57** (2001), 74–80.

[4] N.P. Jewell, J.D. Kalbfleisch (2004), *Maximum likelihood estimation of ordered multinomial parameters*, Biostatistics **5** (2004), 291–306.

[5] N.P. Jewell, M.J. van der Laan, T. Henneman, *Nonparametric estimation from current status data with competing risks*, Biometrika **90** (2003), 183–197.

Additive Isotone Regression

Enno Mammen

(joint work with Kyusang Yu)

This talk is about optimal estimation of the additive components of a nonparametric, additive isotone regression model. We discuss a backfitting estimator that is based on iterative application of the pool adjacent violator algorithm to the additive components of the model. Our main result states the following oracle property. Asymptotically up to first order, each additive component is estimated as well as it would be (by a least squares estimator) if the other components were known. This goes beyond the classical finding that the estimator achieves the same rate of convergence, independently of the number of additive components. The result states that the asymptotic distribution of the estimator does not depend on the number of components.

We have two motivations for considering this model. First of all we think that this is a useful model for some applications. But our main motivation comes from statistical theory. We think that the study of nonparametric models with several nonparametric components is not fully understood. The oracle property that is stated in this paper for additive isotone models has been shown for smoothing estimators in some other nonparametric models. This property is expected to hold if the estimation of the different nonparametric components is based on local smoothing where the localization takes place in different scales. An example are additive models of smooth functions where each localization takes place with respect to another covariate. In Mammen, Linton and Nielsen (1999) the oracle property has been verified for the local linear smooth backfitting estimator. As local linear estimators also isotonic least squares is a local smoother. The estimator is a local average of the response variable but in contrast to local linear estimators

the local neighborhood is chosen by the data. This data adaptive choice is automatically done by the least squares minimization. This understanding of isotonic least squares as a local smoother was our basic motivation to conjecture that for isotonic least squares the oracle property should hold as for local linear smooth backfitting.

The study of the oracle property goes beyond the classical analysis of rates of convergence. Rates of convergence of nonparametric estimators depend on the entropy of the nonparametric function class. If several nonparametric functions enter into the model the entropy is the sum of the entropies of the classes of the components. This implies that the resulting rate coincides with the rate of a model that only contains one nonparametric component. Thus, rate optimality can be shown for a large class of models with several nonparametric components by use of empirical process theory. Rate optimality for additive models was first shown in Stone (1985).

It may be conjectured that the oracle property holds for a much larger class of models. In Horowitz, Klemela and Mammen (2006) a general approach was introduced to applying one-dimensional nonparametric smoothers to an additive model. The procedure consists of two steps. In the first step, a fit to the additive model is constructed using the projection approach of Mammen, Linton and Nielsen (1999). This preliminary estimator uses an undersmoothing bandwidth, so its bias terms are of asymptotically negligible higher order. In a second step, a one-dimensional smoother operates on the fitted values of the preliminary estimator. For the resulting estimator the oracle property was shown: This two step estimator is asymptotically equivalent to the estimator obtained by applying the one-dimensional smoother to a nonparametric regression model that only contains one component. It was conjectured that this result also holds in more general models where several nonparametric components enter into the model. Basically, a proof could be based on this two step procedures. The conjecture has been verified in Horowitz and Mammen (2004, 2006) for generalized additive models with known and with unknown link function.

For more details on additive isotone regression see also Mammen and Yu (2006).

References

[1] J. Horowitz, J. Klemela and E. Mammen, *Optimal estimation in additive regression models*, Bernoulli **12** (2006), 271–298.

[2] J. Horowitz and E. Mammen, *Nonparametric estimation of an additive model with a link function*, Ann. Statist. **32** (2004), 2412–2443.

[3] J. Horowitz and E. Mammen, *Nonparametric estimation of an additive model with an unknown link function*, Working paper (2006).

[4] E. Mammen, O. B. Linton and J. P. Nielsen, *The existence and asymptotic properties of a backfitting projection algorithm under weak conditions*, Ann. Statist. **27** (1999), 1443–1490.

[5] E. Mammen and K. Yu, *Additive Isotone Regression*, Working paper (2006).

[6] C.J. Stone, *Additive regression and other nonparametric models*, Ann. Statist. **13** (1985), 689–705.

Lasso-Type Recovery of Sparse Representations for High-Dimensional Data

Nicolai Meinshausen

The Lasso was introduced by Tibshirani (1996) and has since been proven to be very popular and well studied (Knight 2000, Zhao 2005, Zou 2005, Wright 2006). Some reasons for the popularity might be that the entire regularization path of the Lasso can be computed efficiently (Osborne 2000, Efron et al. 2004), is able to handle more predictor variables than samples, and selects sparse interpretable models. Several extensions and variations have been proposed (Yuan 2005, Zhao 2004, Zou 2005, Candes 2005). The LARS algorithm of Efron et al. (2004) produces the exact solutions of the Lasso for all penalty parameters in a computationally very efficient manner. See also the original homotopy algorithm of Osborne (2000). These solutions have undoubtedly contributed much to the popularity of the Lasso.

The Lasso estimator, as introduced by Tibshirani (1996), is given by

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \ \|Y - X\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_1}, \tag{1}$$

where $X = (X_1, \ldots, X_p)$ is the $n \times p$ matrix whose columns consist of the n-dimensional predictor variables X_k, $k = 1, \ldots, p$. The vector Y contains the n-dimensional set of real-valued observations of the response variable.

The distribution of Lasso-type estimators has been studied in Knight (2000). Variable selection and prediction properties of the Lasso have been studied extensively for high dimensional data with $p \gg n$, a frequently encountered challenge in modern statistical applications.

Most of the work (e.g. Greenshtein 2003, van de Geer 2006) has focused on the behavior of prediction loss, even though van de Geer (2006) obtains also bounds on the ℓ_1-norm distance between the true coefficient vector. We focus exclusively on the properties of the estimate of the coefficient vector under squared error loss and try to understand the behavior of the estimate under a violated *irrepresentable condition*. The aim is to see whether meaningful models can be build. Some examples of other recent work in this direction are Meinshausen (2004), Donoho (2006), Zhao (2005) and Candes (2005).

An estimator is *sign consistent* if and only if

$$P\{\text{sign}(\beta) = \text{sign}(\hat{\beta})\} \to 1 \qquad n \to \infty.$$

It was recently discovered (Zhao 2005, Zou 2005, Meinshausen 2004) that the Lasso estimator can only be sign consistent if the design matrix satisfies the so-called *irrepresentable condition*. The latter condition can easily be violated in applications due to the presence of highly correlated variables.

We examine the behavior of the Lasso estimators if the *irrepresentable condition* is violated. Our main result will show ℓ_2-consistency of the Lasso, even if the *irrepresentable condition* is not fulfilled. An estimator is said to be ℓ_2-consistent if

$$\|\hat{\beta} - \beta\|_{\ell_2} \to 0 \qquad n \to \infty.$$

Convergence rates will also be derived. An ℓ_2-consistent estimator is attractive, as important variables are chosen with high probability and falsely chosen variables have very small coefficients.

Even though the Lasso cannot recover the correct sparsity pattern, we show that the estimator is still consistent in the ℓ_2-norm sense for fixed design under the conditions on (a) the number s of non-zero components of the vector β, (b) the minimal eigenvalues of covariance matrices that are induced by selecting of order s variables. The results are extended to vectors β in weak ℓ_q-balls with $0 < q < 1$.

We also show that the Lasso estimator can be made *sign consistent*, even under a violated irrepresentable condition, by two-stage procedures, by thresholding small coefficients. Preferably, selected variables are first re-estimated with less bias by relaxing the penalty parameter. These results support procedures like Lars-OLS hybrid (Efron et al., 2004), Gauss-Dantzig selector (Candes 2005), or Relaxed Lasso (Meinshausen 2006).

To summarize, it is known the Lasso is bound to select some noise variables under a violated *irrepresentable condition*. The obtained results are encouraging as they show that the Lasso selects at least all important variables with high probability.

References

[1] E. Candes and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Arxiv preprint math.ST/0506081 (2005).

[2] D. Donoho, M. Elad, and V. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Transactions on Information Theory **52** (2006), 6–18.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics **32** (2004), 407–451.

[4] E. Greenshtein and Y. Ritov, Persistence in high-dimensional predictor selection and the virtue of over-parametrization. Bernoulli **10** (2004), 971–988.

[5] K. Knight and W. Fu, *Asymptotics for lasso-type estimators*, Annals of Statistics **28** (2000), 1356–1378.

[6] N. Meinshausen, *Relaxed Lasso*, Computational Statistics and Data Analysis (2006), to appear.

[7] N. Meinshausen and P. Bühlmann (2006), *High dimensional graphs and variable selection with the lasso*, Annals of Statistics **34** (2006), 1436–1462.

[8] M. Osborne, B. Presnell and B. Turlach, *On the lasso and its dual*, J. Comp. Graph. Statist. **9** (2000), 319–337.

[9] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Statist. Soc. B **58** (1996), 267–288.

[10] S. van de Geer, *High-dimensional generalized linear models and the lasso*, Technical Report 133 (2006), ETH Zürich.

[11] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Royal Statist. Soc. B (2006), to appear.

[12] P. Zhao and B. Yu, *Boosted lasso*, Technical Report 678 (2004), University of California, Berkeley.

[13] P. Zhao and B. Yu, *On model selection consistency of lasso*, Technical Report 702 (2006), Department of Statistics, UC Berkeley, to appear in J. Mach. Learning Research.

[14] H. Zou, *The adaptive lasso and its oracle properties*, Technical Report 645 (2005), School of Statistics, University of Minnesota, to appear in J. Amer. Statist. Assoc.

Inference using Shape-Restricted Regression Splines

Mary Meyer

Nonparametric function estimation methods are appealing because they require only minimal, qualitative assumptions. Often the only theoretically valid assumptions can be stated in terms of smoothness or shape restrictions; for example, it might be known that the mean response is a decreasing and convex function of the predictor, or the trend function is increasing and smooth, or the density is unimodal, etc. In practice, however, parametric methods are typically preferred: the usual software packages provide easy estimation and inference methods, and parameters are readily interpretable. Therefore, an important class of inference methods involves the test of a desired parametric form against the valid, qualitative assumptions.

Scatterplot smoothers are among the popular nonparametric methods. One of the difficulties inherent in inference with these methods stems from sensitivity of the fits to user-defined parameters such as bandwidth or smoothing parameter. These may be somewhat arbitrary, and it is undesirable for the inference conclusions to vary with these choices. On the other hand, ordinary shape-restricted methods such as isotonic or convex regression do not require any such subjective input, but inference methods are cumbersome due to the large dimensions of the models, and the fits are often unsatisfying because they are not smooth, and "spike" at the endpoints.

Regression splines are a popular nonparametric function estimation method because they are smooth, flexible, parsimonious, and are very easy to compute. The unrestricted versions are known to be sensitive to knot number and placement, but if assumptions such as monotonicity or convexity may also be imposed on the regression function, the shape-restricted regression splines are robust to knot choices. Further, shape-restricted regression splines are more computationally efficient than the ordinary shape-restricted regression estimators and do not have the spiking problem at the end points. The relatively small degrees of freedom and the insensitivity of the fits to the knot choices allow for practical inference methods, even with small sample sizes.

Tests of constant versus increasing and linear versus convex regression function have been established using ordinary shape-restricted regression for the alternative fit. The distribution of the test statistic under the null hypothesis is known to be that of a mixture of beta random variables. These tests, when implemented with shape-restricted regression splines, have higher power than the standard version. The derivation of the test statistic is similar, and again the distribution under the null hypothesis is a mixture of beta distributions, but the number of distributions in the mix is substantially smaller.

The extension of these tests to the more general test where the null hypothesis is that the regression function has a specific parametric form, versus a qualitative

alternative involving shape, is possible using the regression splines for the alternative fit. A test for linear versus increasing regression function is presented, and some nice properties demonstrated.

Primal and Dual Formulations for the Estimation of a Probability Density via Regularization: Divergences, Entropies, and Likelihoods

Ivan Mizera
(joint work with Roger Koenker)

General schemes relevant for the estimation of a probability density via regularization are investigated—the primal and dual formulations of the discretized version. The primal formulation,

$$\text{(P)} \qquad -w^T Eg + J(-Dg) + \sum_j s_j \psi(h_j) = \min_g ! \quad \text{subject to } h_j \geq g_j$$

is motivated by the case when $\psi(h) = \mathrm{e}^h$ (when we can put $g = h$, the constraint is not needed); in this case, it is a discretization of the usual penalized maximum likelihood scheme

$$-\frac{1}{n}\sum_{i=1}^{n} \log f(x_i) + \lambda J(-\log f) + \int f = \min_f !$$

The (strong) dual of (P) is

$$\text{(D)} \qquad \begin{aligned} -\sum_j s_j \psi^*(f_j) - J^*(u) &= \max_{f,u} ! \\ \text{subject to } Sf = (E^T w + D^T u) &\succeq 0, \end{aligned}$$

where $S = \mathrm{diag}(s)$, and ψ^*, J^* are conjugates of convex functions ψ and J. The instances of J and D cover the usual regularization prescriptions using quadratic and L^1 (total variation) penalties, the Lagrange as well as the constrained version, and also some instances of regularization by shape constraints [5]. For L^1 penalties in particular, (D) becomes

$$\begin{aligned} -\sum_j s_j \psi^*(f_j) &= \max_{f,u} ! \\ \text{subject to } Sf = (E^T w + D^T u) &\succeq 0, \quad \|u\|_\infty \leq \lambda, \end{aligned}$$

the formulation that can be interpreted as the maximization of some entropy function, or, alternatively, minimization of some information divergence over a sieve given by the constraints [1].

We review special cases that yield various Rényi α-entropies [4] in the dual formulation: the Shannon entropy (the Kullback-Leibler divergence), with $\psi(u) = \mathrm{e}^u$ and primal penalizing $\log f$; the Simpson-Gini entropy (the χ^2 divergence), with $\psi(u) = u^2/2$ and primal penalizing f; the case corresponding to the minimum Hellinger distance estimation, with $\psi(u) = -1/u$ and primal penalizing $-1/\sqrt{f}$;

and other [3]. Conditions assuring that the dual minimizer is a density are discussed; finally, the connections of solutions penalizing the total variation of g to the stretched (taut) string density estimators are presented [2].

References

[1] R. Koenker and I. Mizera, *Density estimation by total variation regularization*, Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum (Vijay Nair, ed.), World Scientific, 2006 (in press).

[2] R. Koenker and I. Mizera, *The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and stretched strings*, Prague Stochastics 2006: Proceedings of the joint session of 7th Prague Symposium on Asymptotic Statistics and 15th Prague conference on Information Theory, Statistical Decision Functions and Random Processes (M. Huskova and M. Janzura, eds.), Prague, Matfyzpress, 2006.

[3] R. Koenker and I. Mizera, *Primal and dual formulations relevant for the numerical estimation of a probability density via regularization*, (submitted)

[4] A. Rényi, *On measures of entropy and information*, Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (J. Neyman, ed), University of California Press, Berkeley, 1961.

[5] K. Rufibach and L. Dümbgen, *Maximum likelihood estimation of a log-concave density: basic properties and uniform consistency*, Preprint.

Excess Mass and Related Statistical Methods

Wolfgang Polonik

The excess mass approach has been introduced independently by Hartigan (1987) and Müller and Sawitzki (1991). Hartigan used the approach to construct a nonparametric estimate of a convex density level set. Müller and Sawitzki introduced the excess mass functional, a measure of concentration of a distribution, and constructed tests for modality. First extension have been proposed in Nolan (1991) and Polonik (1995).

Connections of the excess mass approach to a surprising number of other statistical methods exist. This includes nonparametric maximum likelihood density estimation under order restrictions (Polonik, 1998), majorization (Hardy, Littlewood and Polya, 1929, 1952), the so-called Hartigan's dip test of unimodality (Hartigan and Hartigan, 1985), binary classification (Tsybakov, 2004, Steinwart et al., 2005), and split point estimation in decision trees (Bühlmann and Yu, 2002, Banerjee and McKeague, 2006). Further connections exist to estimating minimum volume sets, including the classical 'shorth' and the minimum volume ellipsoid as special cases, and to generalized quantiles (Polonik, 1997). The spirit underlying the excess mass approach can also be found in PRIM, a bump hunting algorithm proposed by Friedman and Fisher (1999) and analyzed by Polonik and Wang (2006), and also in vertical density representation (Troutt, 1991, Troutt, Pang and Hou, 2004).

The excess mass idea has also been utilized in Priebe and Marchette (2000) for density estimation, in Fisher and Marron (2001) for mode testing, for statistically

analyzing a well-known computer vision algorithm, the Hough transform, in Goldenshluger and Zeevi (2004), and also in the construction of classification methods for locally stationary time series (Chandler and Polonik, 2006). Related ideas also are used in Minotte and Scott (1993) and Minotte (1997), again in the context of investigating modality.

In this talk the excess mass approach will be reviewed, and some of the indicated connections will be presented from a certain unified point of view, with the goal of providing a more comprehensive understanding of the methods involved.

References

[1] M. Banerjee and I.W. McKeague, *Confidence sets for split point in decision trees*, Ann. Statist. (2006), to appear.

[2] P. Bühlmann and B. Yu, *Analyzing bagging*, Ann. Statist. **30** (2002), 927–961.

[3] G. Chandler and W. Polonik, *Discrimination of locally stationary time series based on the excess mass functional*, J. Amer. Statist. Assoc. **101** (206), 240–253.

[4] N.I. Fisher and J.S. Marron, *Mode testing via the excess mass estimate*, Biometrika **88** (2001), 499–517.

[5] J.H. Friedman and N.I. Fisher, *Bump hunting in high-dimensional data*, Statist. Comput. **9** (1999), 123–143.

[6] A. Goldenshluger and A. Zeevi, *The Hough transform estimator*, Ann. Statist. **32** (2004), 1908–1932.

[7] G.H. Hardy, J.E. Littlewood and G. Pólya, *Some simple inequalities satisfied by convex functions*, Messenger Math. **58** (1929), 145–152.

[8] G.H. Hardy, J.E. Littlewood and G. Pólya, *Inequalities*, Cambridge University Press (1952).

[9] J.A. Hartigan and P.M. Hartigan, *The dip test of unimodality*, Ann. Statist. **13** (1985), 70–84.

[10] J.A. Hartigan, *Estimation of a convex density contour two dimentions*, J. Amer. Statist. Assoc. **82** (1987), 267–270.

[11] C. Minotte and D.W. Scott, *The mode tree: A tool for visualization of nonparametric density features*, J. Comput. Graph. Statist. **2** (1993), 51–68.

[12] C. Minotte, *Nonparametric testing for the existence of a mode*, Ann. Statist. **25** (1997), 1646–1660.

[13] D.W. Müller and G. Sawitzki, *Excess mass estimates and tests for multimodality*, J. Amer. Statist. Assoc. **86** (1991), 738–746.

[14] D. Nolan, *The excess-mass ellipsoid*, J. Multivariate Anal. **39** (1991), 348–371.

[15] W. Polonik, *Measuring mass concentrations and estimating density contour clusters - an excess mass approach*, Ann. Statist. **23**, 855–881.

[16] W. Polonik, *Minimum volume sets and generalized quantile processes*, Stoch. Process. Appl. **69** (1997), 1–24.

[17] W. Polonik, *The silhouette, concentration functions and ML-density estimation under order restrictions*, Ann. Statist. **26** (1998), 1857–1877.

[18] W. Polonik and Z. Wang, *Estimation of regression contour clusters: an application of the excess mass approach to regression*, J. Multivariate Anal. **94** (2005), 227–249.

[19] W. Polonik and Z. Wang, *PRIM Analysis*, Preprint.

[20] C.E. Priebe and D.J. Marchette, *Alternating kernel and mixture density estimates*, Comput. Statist. Data Anal. **35** (2000), 43–65.

[21] M.D. Troutt, *A theorem on the density of the density ordinate and an alternative derivation of the Box-Müller method*, Statistics **22** (1991), 436–466.

[22] M.D. Troutt, W.K. Pang and S.H. Hou, *Vertical density representation and its applications*, World Scientific (2004).

[23] I. Steinwart, D. Hush and C. Scovel, *A classification framework for anomaly detection*, J. Machine Learning Research **6** (2005), 211–232.

Multivariate Regression Estimation with Monotonicity Constraints

Regine Scheder
(joint work with Holger Dette)

In a recent paper [1], a smooth monotonizing procedure is proposed for a one-dimensional nonparametric regression setting. In this talk, I describe two applications of this procedure in multivariate settings. We address monotonicity constraints in traditional regression models and in quantile regression models.
The problem of estimating a strictly monotone and smooth regression function in two or more variables is discussed in a nonparametric regression context. Our method starts with an unconstrained estimator and uses successively the one-dimensional isotonization procedure. The procedure calculates the monotonized inverse of the function. After that an inversion of the function is necessary. To use the method stepwise for each variable, we show that the monotonizing procedure applied in one variable does not destroy the monotonicity of the regression function in another variable. There are several advantages of this step-by-step monotonizing procedure. In each step, a comparison with the original estimator is possible. The method can be easily adjusted for a given problem, e.g., if the regression function is assumed to be strictly increasing in one variable and strictly decreasing in another one. Sometimes, it might be necessary or sufficient to monotonize only one direction. Although the procedure monotonizes one direction per step, the actual order of monotonizing has no substantial influence. All these features turn the customized usage of this procedure into an easily applicable and comprehensible method. An implementation of this isotonization is provided in the R package *monoProc* [5].
The second part of the talk focuses on conditional Quantile Regession estimation. There are two ways to define conditional quantiles. One method uses the check function $\rho_\alpha(u) = |u| + (2\alpha - 1)u$ to formulate a minimization problem (see [4] for more details). The other method relies on the fact that the conditional quantile function is the inverse of the conditional distribution function of Y given $\mathbf{x}$. There are some problems in this approach. First of all to get the inverse you need to have an isotone estimate of the conditional distribution function. For the usual Nadaraya-Watson-Estimator, this is the case as long as a positive kernel function is used, but this estimate has some asymptotic disadvantages compared to local linear estimators which are on the other hand not necessarily monotone (see [3]). A perfect workaround is the above motivated method of monotonization. The monotonizing procedure for the regression model is a two-stage method in each step. For the quantile regression model, the first stage is sufficient to get the monotonized inverse of the estimated distribution function. This is a very appealing technique, because the asymptotic behavior of local linear estimators can

be preserved, and it is easy to implement. Another drawback of estimating the conditional quantile function through the inverse of the distribution function is the curse of dimensionality which occurs easily when considering many covariates. A common way to deal with that is to assume an additive model and use marginal integration. We can show that this estimator of the additive component has a one-dimensional rate of convergence. We illustrate this approach with a data example.

References

[1] H. Dette, N. Neumeyer, K. Pilz, *A simple nonparametric estimator of a monotone regression function*, Bernoulli **12** (2006), 469–490.
[2] H. Dette, R. Scheder, *Strictly monotone and smooth nonparametric regression for two or more variables*, Can. J. Statist. **34** (2006), to appear.
[3] P. Hall, R. Wolff, Q. Yao, *Methods for estimating a conditional distribution funciton*, J. Amer. Statist. Assoc. **94** (1999), 154–163.
[4] R. Koenker, G. Bassett, *Regression quantiles*, Econometrica **46** (1978), 33–50.
[5] R. Scheder, *monoProc Package for R*, http://homepage.ruhr-uni-bochum.de/Regine.Scheder/ work.html (2005).

An Introduction to the 3U Method and Its Application to Measurement Error Problems

Jiayang Sun
(joint work with Xiaofeng Wang and Michael Woodroofe)

Many interesting practical problems can be formulated as studies about data with measurement errors. For example, density estimation from astronomical data with additive errors is related to a deconvolution problem; regression estimation with errors in covariates arises naturally in a general linear model with a repeated measures design. In this talk, we present our new research in these areas when the errors are homogeneous or inhomogeneous. Our new density estimator is "suboptimal", stable and easy to compute. In particular, no Fourier transformation (as it is for most deconvolution estimators) is needed in our computational formulae. The idea is to start from the case when errors are uniformly distributed. It then proceeds to the case when errors are distributed as a mixture of uniforms, hence approximating a large class of error distributions (including the normal distribution), in the spirit of how most random numbers are generated. Based on a representation of our new density estimator for data with measurement errors, we demonstrate how a new nonparametric regression estimator for data with errors in covariates can be established based on a "density-clone". We also discuss their implications to symbolic data analysis. This opens up a new line of research on deconvolution and errors in covariates problems.

Sparsity oracle inequalities in high-dimensional regression and density estimation

Alexander B. Tsybakov
(joint work with F. Bunea, M.H. Wegkamp)

We study regression and density estimation methods based on penalized empirical risk minimization with an ℓ_1 penalty. In regression problems, we observe $(X_1, Y_1), \dots, (X_n, Y_n)$ a sample of independent random pairs distributed as a random variable $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. We denote the probability measure of X by μ. The goal is the estimation of the unknown regression function $f(X) = \mathbb{E}(Y|X)$. In density estimation problems, we observe independent random variables $X_1, \dots, X_n$ with common probability density f on $\mathbb{R}^d$ and we want to estimate f.

In both cases the estimators that we propose are linear combinations, with data-dependent weights, of functions $f_j : \mathcal{X} \to \mathbb{R}$ in a given dictionary $\mathcal{F}_M = \{f_1, \dots, f_M\}$. We show below that our estimators mimic unknown sparse approximations of f within the dictionary $\mathcal{F}_M$, if such approximations exist. We will characterize sparsity in the following way. For any $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$, define $\mathsf{f}_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x)$. Let $M(\lambda) = \sum_{j=1}^M I_{\{\lambda_j \neq 0\}} = \text{Card } J(\lambda)$ denote the number of non-zero coordinates of λ, where $I_{\{\cdot\}}$ denotes the indicator function, and $J(\lambda) = \{j \in \{1, \dots, M\} : \lambda_j \neq 0\}$. The value $M(\lambda)$ characterizes the *sparsity* of the vector λ: the smaller $M(\lambda)$, the "sparser" λ. We are particularly interested in the case where M is very large, $M \gg n$. We obtain results in the form of sparsity oracle inequalities, i.e., oracle inequalities involving the "true" small dimension $M(\lambda)$ in place of the huge redundant dimension M. The suggested methods and sparsity oracle inequalities can be applied in the following three scenarios:

(i) sparse parametric models with $M \gg n$ parameters: in this case $f = \mathsf{f}_{\lambda^*}$ for some $\lambda^* \in \mathbb{R}^M$, and we get that our estimators converge with the L_2-rate $O(n^{-1} M(\lambda^*) \log M)$;

(ii) nonparametric regression and density estimation in classical settings where $\{f_1, \dots, f_M\}$ are the first M functions of an orthonormal basis: in this case typically $M \leq n$ and the obtained oracle inequalities imply near minimax convergence rates of our procedures on various functional classes;

(iii) aggregation of arbitrary regression or density estimators: in this case $f_1, \dots, f_M$ are preliminary estimators of f constructed from a training sample independent of the current observations, and our results imply that the suggested procedures lead to near optimal rates of aggregation.

We refer to [1, 2, 3, 4] for a more detailed exposition of our results and discussion of the implications (i) – (iii). Sparsity oracle inequalities for some settings different from ours are given in [6, 8], see also [7]. We now define our estimators and present some selected sparsity oracle inequalities separately for each of the two problems.

Regression. Our vector of data-dependent weights $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)$ is obtained via penalized least squares:

$$\widehat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \mathsf{f}_\lambda(X_i)\}^2 + \mathrm{pen}(\lambda) \right\}, \tag{1}$$

where the penalty is given by

$$\mathrm{pen}(\lambda) = r_{n,M} \sum_{j=1}^{M} \|f_j\|_n |\lambda_j|$$

with $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g^2(X_i)$ for any function $g : \mathcal{X} \to \mathbb{R}$. Here $r_{n,M}$ is a positive constant. A sensible choice is $r_{n,M} = A\sqrt{\frac{\log M}{n}}$, for a suitably large constant $A > 0$. The corresponding estimate of f is $\widehat{f} = \sum_{j=1}^M \widehat{\lambda}_j f_j$. Let $\|g\|^2 = \int g^2(x)\, \mu(dx)$.

We define the oracle set

$$\Lambda = \left\{ \lambda \in \mathbb{R}^M : \ \|f - \mathsf{f}_\lambda\|^2 \le C_f r_{n,M}^2 M(\lambda) \right\},$$

for some positive constant C_f which is allowed to depend on f. If Λ is nonempty, we say that f has the *weak sparsity property relative to the dictionary* $\mathcal{F}_M$. Informally, this definition can be related to the intuitive notion of sparsity if for all the oracle values $\lambda \in \Lambda$ we have $M(\lambda) \ll M$. Weak sparsity can be viewed as a milder version of the *strong sparsity* (or simply *sparsity*) property which commonly means that f admits an exact representation $f = \mathsf{f}_{\lambda^*}$ for some $\lambda^* \in \mathbb{R}^M$, with hopefully small $M(\lambda^*)$. Consider the "correlations"

$$\rho_M(i,j) = \frac{< f_i, f_j >}{\|f_i\| \|f_j\|}, \quad 1 \le i \ne j \le M,$$

where $< \cdot, \cdot >$ stands for the scalar product induced by $\|\cdot\|$, and define the coherence numbers (cf. [5])

$$\rho(\lambda) = \max_{i \in J(\lambda)} \max_{j \ne i} |\rho_M(i,j)|.$$

To state the result, we need the following assumptions:

Assumption (A1) $\exists\, b < \infty$ such that $\mathbb{E}\{\exp(|Y - f(X)|) \,|\, X\} \le b$ (a.s.).
Assumption (A2) $\exists\, L < \infty$ such that $\max_{1 \le j \le M} \|f_j\|_\infty \le L$, $\exists\, c_0 > 0$ such that $\|f_j\| \ge c_0$ for all $1 \le i, j \le M$ and $\|f\|_\infty = \sup_x |f(x)| < \infty$.

Fix some $L_0 < \infty$ such that $\mathbb{E}[f_i^2(X) f_j^2(X)] \le L_0$. Set $L(\lambda) = \|f - \mathsf{f}_\lambda\|_\infty$.

Theorem 1. *Assume that (A1) and (A2) hold. Then, for all $\lambda \in \Lambda$ such that $45\, M(\lambda)\rho(\lambda) \le 1$ we have, with probability at least $1 - \pi_{n,M}(\lambda)$,*

$$\|\widehat{f} - f\|^2 \le C r_{n,M}^2 M(\lambda)$$

and

$$|\widehat{\lambda} - \lambda|_1 \le C r_{n,M} M(\lambda),$$

where C is a constant only depending on c_0 and C_f, and

$$\pi_{n,M}(\lambda) \le \exp\left(-C_2 \frac{M(\lambda)}{L^2(\lambda)} n r_{n,M}^2\right) + 10M^2 \exp\left(-C_1 n \min\left\{r_{n,M}^2, \frac{r_{n,M}}{L}, \frac{1}{L^2}, \frac{1}{L_0 M^2(\lambda)}, \frac{1}{L^2 M(\lambda)}\right\}\right),$$

for some constants C_1, C_2 depending on c_0, C_f and b only.

Density estimation. In this case, we define a vector $\widehat{\lambda}$ of data-dependent weights in the form

$$\widehat{\lambda} \;=\; \operatorname*{arg\,min}_{\lambda \in \mathbb{R}^M} \Big\{ -\frac{2}{n}\sum_{i=1}^n \mathsf{f}_\lambda(X_i) + \|\mathsf{f}_\lambda\|_{\mathrm{Leb}}^2 \;+\; \sum_{j=1}^M \omega_j |\lambda_j| \Big\}$$

where

$$\omega_j = 4L_j \sqrt{\frac{2\log(M/\delta)}{n}} \quad \text{with} \quad L_j = \|f_j\|_\infty,$$

$0 < \delta < 1/2$ is a small tuning parameter, and $\|g\|_{\mathrm{Leb}}^2 = \int g^2(x)\, dx$. We call the resulting estimate of f the SPADES estimator (SPArse Density EStimator). It is given is by $f^{\spadesuit}(x) = \mathsf{f}_{\widehat{\lambda}}(x),\ \forall x \in \mathbb{R}^d$. Set $F(\lambda) = \max_{j \in J(\lambda)} \frac{\|f_j\|_\infty}{\|f_j\|_{\mathrm{Leb}}}$, for all $\lambda \in \mathbb{R}^M$, and define $\rho(\lambda)$ as above, but using $\|\cdot\|_{\mathrm{Leb}}$ and the corresponding scalar product $<\cdot,\cdot>_{\mathrm{Leb}}$ in place of $\|\cdot\|$ and $<\cdot,\cdot>$ in the expression for $\rho_M(i,j)$.

Theorem 2. *Assume that $L_j < \infty$ for $1 \le j \le M$. Then, with probability at least $1 - 2\delta$, for all $n \ge 1$, $\alpha > 1$ and all $\lambda \in \mathbb{R}^M$ that satisfy $32\, F(\lambda)\rho(\lambda)M(\lambda) \le 1$, we have the following oracle inequality:*

$$\|f^{\spadesuit} - f\|_{\mathrm{Leb}}^2 + \frac{\alpha}{2(\alpha-1)} \sum_{j=1}^M \omega_j |\widehat{\lambda}_j - \lambda_j| \le \frac{\alpha+1}{\alpha-1} \|\mathsf{f}_\lambda - f\|_{\mathrm{Leb}}^2 + \frac{(16\alpha)^2}{\alpha-1} F^2(\lambda) M(\lambda) \frac{\log(M/\delta)}{n}.$$

References

[1] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Aggregation for Gaussian regression*, Ann. Statist., to appear (2007).

[2] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Aggregation and sparsity via ℓ_1-penalized least squares*, Proceedings of 19th Annual Conference on Learning Theory, COLT-2006. Lecture Notes in Artificial Intelligence **4005** (2006), 379–391. Springer-Verlag, Heidelberg.

[3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Preprint (2006).
[4] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Sparse density estimation and aggregation with ℓ_1 penalties*, Preprint (2006).
[5] D.L. Donoho, M. Elad, and V. Temlyakov, *Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise*, IEEE Trans. Info. Theory **52** (2006), 6–18.
[6] V. Koltchinskii, *Sparsity in penalized empirical risk minimization*, Preprint (2006).
[7] A.B. Tsybakov, *Comments on "Regularization in Statistics", by P.Bickel and B.Li*, Test **15** (2006), 303–310.
[8] S.A. van de Geer, *High dimensional generalized linear models and the Lasso*, Research report No.133 (2006), Seminar für Statistik, ETH, Zürich.

A Bound for the Empirical Risk Minimizer

Sara A. van de Geer

Suppose we observe n i.i.d. copies $X_1, \ldots, X_n$ of a random variable $X \in \mathbf{X}$ with distribution P. Let $\mathbf{F}$ be a given class of functions f on $\mathbf{X}$. The empirical risk minimizer is

$$\hat{f} := \arg\min_{f \in \mathbf{F}} P_n f, \ P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Let $\mathbf{F}^{(0)} \supset \mathbf{F}$ and define the target

$$f_0 = \arg\min_{f \in \mathbf{F}^{(0)}} Pf,$$

and its best approximation

$$f^* = \arg\min_{f \in \mathbf{F}} Pf.$$

The excess risk at f is defined as

$$\mathrm{E}_0(f) = P(f - f_0).$$

In this note, we prove a bound for the excess risk $\mathrm{E}_0(\hat{f})$. The result is a slightly modified version of Theorem 2 in Koltchinskii (2006). Namely, we allow that the target f_0 is not in the model class $\mathbf{F}$.

Note that

$$\mathrm{E}_0(\hat{f}) = \mathrm{E}_0(f^*) + P(\hat{f} - f^*). \tag{1}$$

This can be seen as a decomposition into approximation error $\mathrm{E}_0(f^*)$ and estimation error $P(\hat{f} - f^*)$.

The idea is now to find good upper bounds for the right hand side of (1). By an easy argument, the estimation error can be bounded by

$$P(\hat{f} - f^*) \leq |(P_n - P)(\hat{f} - f^*)|.$$

This shows that the empirical process plays an important role. The increments of empirical process can be studied in terms of the variance of the functions involved. Since in (1), it is the excess risk which is on the left hand side, we need to bound variances in terms of excess risk. This we call the margin behavior.

To handle the empirical process behavior, we apply a concentration inequality of Bousquet (2002). Let

$$\sigma^2(f) := Pf^2 - (Pf)^2,$$

and let

$$\mathbf{F}_\sigma := \{f \in \mathbf{F} : \ \sigma(f - f_0) \leq \sigma\}, \ \sigma > 0.$$

Consider the maximal increment of the empirical process

$$Z(\sigma) := \sup_{f \in \mathbf{F}_\sigma} |(P_n - P)(f - f_0)|, \ \sigma > 0.$$

The empirical process behavior is the behavior of $\mathbf{E}Z(\sigma)$ as function of σ.

Bousquet's inequality. *Suppose* $\|f - f_0\|_\infty \leq 1$ *for all* $f \in \mathbf{F}$. *Then*

$$\mathbf{P}\left(Z(\sigma) \geq \mathbf{E}Z(\sigma) + \sqrt{2t/n}\sqrt{\sigma^2 + 2\mathbf{E}Z(\sigma)} + \frac{t}{3n}\right)$$

$$\leq \mathrm{e}^{-t} \ \forall \ t > 0. \tag{1}$$

The margin behavior of $\mathrm{E}_0(f)$ is the behavior of $\mathrm{E}_0(f)$ for $\sigma(f - f_0)$ small. This is described by

$$D(\epsilon) = \sup\{\sigma(f - f_0) : \ f \in \mathbf{F} : \ \mathrm{E}_0(f) \leq \epsilon\}, \ \epsilon > 0.$$

Define

$$\phi(\epsilon) = \sqrt{n}\mathbf{E}Z(D(\epsilon)).$$

We first present a result for the weighted empirical process (see Massart (2000), Bartlett, Bousquet and Mendelson (2005)).

Lemma 1. *Suppose that, for some* $0 < \gamma < 1$, *the function* $\phi(\epsilon)/\epsilon^\gamma$ *is non-increasing in* ϵ *for* ϵ *bigger than, or equal to, some lower bound* ϵ_n. *Then for all* $\epsilon \geq \epsilon_n$,

$$\mathbf{E}\left(\sup_{f \in \mathbf{F}, \ \mathrm{E}_0(f) > \epsilon} \frac{\sqrt{n}|(P_n - P)(f - f_0)|}{\mathrm{E}_0(f)}\right)$$

$$\leq C_\gamma \frac{\phi(\epsilon)}{\epsilon},$$

where

$$C_\gamma := \frac{\gamma^{-\gamma/(1-\gamma)}}{1 - \gamma}.$$

The conjugate of a convex increasing function G on $[0, \infty)$ with $G(0) = 0$, is defined as the function $H(v) = \sup_{u \geq 0}[uv - G(u)]$.

Lemma 2. *Suppose* $\|f - f_0\|_\infty \leq 1$ *for all* $f \in \mathbf{F}$. *Assume that* $\phi(\epsilon)/\epsilon^\gamma$ $(0 < \gamma < 1)$ *as well as* $D(\epsilon)/\epsilon$, *are non-increasing in* $\epsilon \geq \epsilon_n$. *Assume furthermore that*

$$4C_\gamma \phi(\epsilon) + 2\sqrt{2t} D(\epsilon) \leq \psi(\epsilon), \ \epsilon \geq \epsilon_n.$$

where ψ *is a function with convex increasing inverse* ψ^{-1} *having conjugate* H. *The constant* C_γ *is from Lemma 1. Let* $0 < \delta < 1/2$ *be arbitrary, and define*

$$\epsilon_{t,n} := \left[\delta H(\frac{1}{\delta\sqrt{n}}) + \frac{2t}{3n\delta} \right] \vee \epsilon_n.$$

Then for all $\epsilon \geq \frac{1}{1-2\delta}(\epsilon_{t,n} + \mathrm{E}_0(f^*))$, *we have*

$$\mathbf{P}(\mathrm{E}_0(\hat{f}) > \epsilon) \leq 2\mathrm{e}^{-t}.$$

Proof. Note first that for $\mathrm{E}_0(f) > \epsilon$, we have $\epsilon\|f - f_0\|_\infty/\mathrm{E}_0(f) \leq 1$ and $\epsilon\sigma(f - f_0)/\mathrm{E}_0(f) \leq D(\epsilon)$. By Bousquet's inequality, with probability at least $1 - \mathrm{e}^{-t}$, uniformly in $f \in \mathbf{F}$ with $\mathrm{E}_0(f) > \epsilon$,

$$|(P_n - P)(f - f_0)|$$

$$\leq \frac{\mathrm{E}_0(f)}{\epsilon} \left[4C_\gamma \frac{\phi(\epsilon)}{\sqrt{n}} + \sqrt{\frac{2t}{n}} D(\epsilon) + \frac{2t}{3n} \right].$$

Moreover

$$\mathrm{E}_0(\hat{f}) \leq |(P_n - P)(\hat{f} - f_0)|$$

$$+(P_n - P)(f^* - f_0) + \mathrm{E}_0(f^*)$$

$$\leq |(P_n - P)(\hat{f} - f_0)| + \sqrt{\frac{2t}{n}} D(\epsilon) + \frac{t}{3n} + \mathrm{E}_0(f^*),$$

with probability at least $1 - \mathrm{e}^{-t}$. Here, we used Bernstein's inequality, and the fact that $\mathrm{E}_0(f^*) \leq \epsilon$ implies $\sigma(f^* - f_0) \leq D(\epsilon)$. It follows that if $\mathrm{E}_0(\hat{f}) > \epsilon$, then with probability at least $1 - 2\mathrm{e}^{-t}$,

$$\mathrm{E}_0(\hat{f}) < \frac{\mathrm{E}_0(\hat{f})}{\epsilon} \left[4C_\gamma \frac{\phi(\epsilon)}{\sqrt{n}} + 2\sqrt{\frac{2t}{n}} D(\epsilon) + \frac{2t}{3n} \right]$$

$$+\mathrm{E}_0(f^*) + \frac{t}{3n}$$

$$\leq \frac{\psi(\mathrm{E}_0(\hat{f}))}{\sqrt{n}} + \frac{\mathrm{E}_0(\hat{f})}{\epsilon} \frac{2t}{3n} + \mathrm{E}_0(f^*) + \frac{t}{3n}$$

$$\leq \frac{\psi(\mathrm{E}_0(\hat{f}))}{\sqrt{n}} + \delta \mathrm{E}_0(\hat{f}) + \mathrm{E}_0(f^*) + \frac{t}{3n},$$

where we used that $\epsilon \geq 2t/(3n\delta)$. Now, we have for all $z > 0$,

$$\frac{\psi(z)}{\sqrt{n}} \leq \delta z + \delta H\left(\frac{1}{\delta\sqrt{n}}\right).$$

So, using the bound $1 \leq 2/\delta$, we arrive at

$$\mathrm{E}_0(\hat{f}) \leq 2\delta \mathrm{E}_0(\hat{f}) + \delta H\left(\frac{1}{\delta\sqrt{n}}\right) + \mathrm{E}_0(f^*) + \frac{2t}{3n\delta},$$

which implies

$$\mathrm{E}_0(\hat{f}) \leq \frac{1}{1-2\delta}\left(\delta H\left(\frac{1}{\delta\sqrt{n}}\right) + \mathrm{E}_0(f^*) + \frac{2t}{3n\delta}\right) \leq \frac{1}{1-2\delta}\epsilon_{n,t} < \epsilon.$$

□

As an example, suppose that $|\mathbf{F}| = M < \infty$. Then for $D(\epsilon) \geq \sqrt{\log(2M)/(2n)}$, we have

$$\phi(\epsilon) \leq 2D(\epsilon)\sqrt{\frac{2\log(2M)}{n}}.$$

Assume now that $D(\epsilon)/\epsilon^\gamma \downarrow$ for some $0 < \gamma < 1$. Let $\mathbf{D} \geq D$ be a concave increasing function, with inverse $\mathbf{D}^{-1}$ having conjugate $\mathbf{H}$. We can take

$$\psi(\epsilon) = \mathbf{D}(\epsilon)\left(8C_\gamma\sqrt{2\log(2M)} + 2\sqrt{2t}\right) := \mathbf{D}(\epsilon)c_t.$$

Thus,

$$\psi^{-1}(\sigma) = \mathbf{D}^{-1}(\sigma/c_t),$$

and

$$H(\epsilon) = \mathbf{H}(\epsilon c_t).$$

We obtain

$$\epsilon_{t,n} = \delta\mathbf{H}\left(\frac{8C_\gamma\sqrt{2\log(2M)} + 2\sqrt{2t}}{\sqrt{n}\delta}\right) + \frac{2t}{3n\delta} \vee \mathbf{D}^{-1}\left(\sqrt{\frac{\log(2M)}{2n}}\right).$$

In the special case of quadratic margin behavior, we have $D(\epsilon) = \sqrt{2\epsilon}$ (say), so that we may take $\gamma = 1/2$, $C_\gamma = 4$ and $\mathbf{H}(\epsilon) = \epsilon^2/2$. This gives

$$\epsilon_{t,n} = \frac{1}{\delta}\left(\frac{32\sqrt{\log(2M)} + 2\sqrt{t}}{\sqrt{n}}\right)^2 + \frac{2t}{3n\delta}.$$

References

[1] P.L. Bartlett, O. Bousquet and S. Mendelson, *Local Rademacher complexities*, Ann. Statist. **33** (2005), 1497–1537.

[2] O. Bousquet, *A Bennett concentration inequality and its application to suprema of empirical processes*, C.R. Acad. Sci. Paris **334** (2002), 495–500.

[3] V. Koltchinskii, *2004 IMS Medallion Lecture: Local Rademacher complexities and oracle inequalities in risk minimization*, Ann. Statist. **34** (2006).

[4] P. Massart, *Some applications of concentration inequalities to statistics*, Annales de la Faculté de Toulouse **9** (2000), 245–303.

Regularization by Gaussian process priors

Aad van der Vaart
(joint work with Harry van Zanten)

Gaussian processes have been suggested as prior models for regularizing functions in nonparametric statistical estimation. (E.g. Kimeldorf and Wahba (1970), Ghosal and Roy (2006), Rasmussen et al. (2006).) For instance, a sample path of a Gaussian process:

- can be used directly as a prior model for a regression function w_0, where the observations are a sample $(x_1, Y_1), \ldots, (x_n, Y_n)$ from the regression model $Y_i = w_0(x_i) + e_i$, with mean zero errors.
- can be used after a monotonic transformation $\Psi : \mathbb{R} \to [0,1]$ to the unit interval in the setting of classification, where we observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $Y_i \in \{0,1\}$ and $\mathrm{P}(Y_i = 1 | X_i = x) = \Psi \circ w_0(x)$.
- can be used for density estimation after exponentiation and renormalization, where the observations are a sample $X_1, \ldots, X_n$ from the density $x \mapsto \exp(w_0(x)) \int \exp(w_0(t))\, dt$.

Considering the unknown function as a sample path of the Gaussian process and the observations as sampled from the model specified by this realization, we can construct the conditional distribution of the function given the observations, by Bayes' rule. This "posterior distribution" is a random measure on the parameter set, and can also be studied under the assumption that the observations are sampled from a fixed distribution given by a "true" function w_0, without adopting the Bayesian frame-work. We study whether the posterior distributions contract to the true parameter w_0, and the rate at which this happens, as the informativeness of the data increases indefinitely.

For contraction to happen it typically suffices that the true function is in the support of the Gaussian process. Van der Vaart and Van Zanten (2006) have obtained the following upper bound on the rate of contraction. For a Borel measurable Gaussian random variable W in a Banach space $(\mathbb{B}, \|\cdot\|)$, let $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ be its reproducing kernel Hilbert space (RKHS). For given $w_0 \in \mathbb{B}$ define

$$\phi_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathrm{P}\big(\|W\| < \varepsilon\big). \tag{1}$$

Then in statistical settings in which the statistically relevant distance combines correctly with the norm $\|\cdot\|$ on the Banach space $\mathbb{B}$ and where n describes the informativeness of the data in the usual way, the posterior contracts at the rate ε_n satisfying

$$\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2. \tag{2}$$

This is true for the three settings described previously, where in the case of regression this has been proved for Gaussian errors and using the L_2-distance, in the classification example for the logistic link and the L_2-distance on the functions $\Psi \circ w$, and in the density estimation problem when using the Hellinger distance.

The function $\varepsilon \mapsto \phi_{w_0}(\varepsilon)$ is closely related to the probability that the Gaussian variable W falls in a ball of radius ε around w_0. This probability for $w_0 = 0$ is known as the *small ball probability* and has been studied in many papers. (E.g Kuelbs and Li (1993), Li and Shao (2001).)

A common test case for nonparametric methods is the performance if the true parameter belongs to a regularity class, such as a Hölder space $C^\alpha[0,1]$ or a Besov space. Within our Bayesian context we ask which Gaussian priors will lead to the optimal rate of contraction, e.g. in the Hölder case whether the smallest ε_n satisfying the preceding display is of the order $n^{-\alpha/(2\alpha+1)}$.

We have studied several classes of priors. Two examples which (nearly) achieve this aim are based on integrated Brownian motion and rescaled stationary processes.

For $\alpha > 0$ and B a standard Brownian motion the *Riemann-Liouville process with Hurst parameter* $\alpha > 0$ is defined as

$$R_t = \int_0^t (t-s)^{\alpha-1/2}\, dB_s, \quad t \geq 0.$$

The process R is a centered Gaussian process with continuous sample paths. It can be viewed as a multiple of the $(\alpha+1/2)$-fractional integral of the "derivative dB of Brownian motion". Since Brownian motion is "regular of order $1/2$", the Rieman-Liouville process R ought to be a good model for "α-regular functions", except for the fact that through its definition as an integral from 0 its derivatives at 0 are tied down. Relaxation at 0 can be achieved by adding a Gaussian polynomial. The Gaussian process

$$W_t = \sum_{k=0}^{\lfloor \alpha \rfloor + 1} Z_k t^k + R_t, \tag{3}$$

where $Z_1, \ldots, Z_{\lfloor \alpha \rfloor + 1}$ are i.i.d. standard normal variables independent of R turns out to be an appropriate model for "α-regular functions". Indeed, it can be shown to lead to the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ if the true parameter is α-regular. If the true parameter is of a different regularity, then the posterior will still contract to the true parameter, but at a suboptimal rate: never faster than $n^{-\alpha/(2\alpha+1)}$ and slower than $n^{-\beta/(2\beta+1)}$ if the true regularity β is smaller than α. For instance, for Brownian motion itself ($\alpha = 1/2$) the rate is never faster than $n^{-1/4}$. This is a consequence of the low small ball probability of Brownian motion.

As a second example consider a sequence of mean-zero Gaussian process $W^n = (W_t^n : 0 \leq t \leq 1)$ with covariance function

$$\mathrm{E} W_s^n W_t^n = \phi\Big(\frac{s-t}{c_n}\Big),$$

for a symmetric function $\phi : \mathbb{R} \to \mathbb{R}$ and constants $c_n \downarrow 0$. This corresponds to a rescaling of a sample path on $[0, 1/c_n]$ of the stationary Gaussian process with covariance function ϕ to the interval $[0,1]$. We assume that the spectral measure μ corresponding to ϕ satisfies $\int e^{\alpha\lambda}\, d\mu(\lambda) < \infty$ for some $\alpha > 0$ and possesses a positive density relative to Lebesgue measure. Under this condition Van der Vaart

and Van Zanten (2007) show that the rate of contraction is the optimal one up to a logrithmic factor, if c_n is set equal to the usual bandwidth $n^{-1/(2\alpha+1)}$ for nonparametric smoothing. The proof is based on a new small ball probability for smooth Gaussian processes and a study of the approximation properties of the reproducing kernel Hilbert space. The exponent $\phi_0(\varepsilon)$ of the small ball probability of the process W^n relative to the uniform norm on $C[0,1]$ is bounded above by

$$\frac{1}{c_n}\Big(\log\frac{1}{\varepsilon}\Big)^2,$$

up to logarithmic factors.

Thus for every level of regularity there exist Gaussian processes that achieve the optimal posterior rate of contraction. The Bayesian approach to choose the prior from these various candidate priors is to put a prior on the regularity level α, and consider the Gaussian process priors as conditional priors given the regularity. It is known that for certain priors on α this hierarchical procedure will obtain the posterior rate for any regularity level of the true parameter. The exact class of priors on α for which is true is under study.

References

[1] S. Ghosal and A. Roy, *Posterior consistency of Gaussian process prior for nonparametric binary regression*, Preprint (2006).

[2] S. Ghosal, J.K. Ghosh and A.W. van der Vaart, *Convergence rates of posterior distributions*, Ann. Statist. **28** (2000), 500–531.

[3] G. Kimeldorf and G. Wahba, *A correspondence between Bayesian estimation on stochastic processes and smoothing splines*, Ann. Math. Statist. **41** (1970), 495–502.

[4] J. Kuelbs and W. Li, *Metric entropy and the small ball problem for Gaussian measures*, J. Functional Analysis **116** (1993), 133–157.

[5] W. Li and Q.-M. Shao, *Gaussian processes: inequalities, small ball probabilities and applications*, Stochastic processes: theory and methods, Handbook of Statistics **19** (2001), 533–597, North Hollard, Amsterdam.

[6] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, MIT Press, Boston (2006).

[7] A.W. van der Vaart and J.H. van Zanten, (2006), *Rates of contraction of posterior distributions based on Gaussian process priors*, Preprint (2006).

[8] A.W. van der Vaart and J.H. van Zanten, *Small ball probabilities of smooth Gaussian processes and Bayesian inference with rescaled Gaussian process priors*, Preprint (2007).

A Goodness-of-fit Test for Semiparametric Models and Models with Shape Constraints in Multiresponse Regression

Ingrid van Keilegom

(joint work with Song Xi Chen)

We propose an empirical likelihood test that is able to test the goodness-of-fit of a class of semiparametric regression models. The class includes as special cases fully parametric models, semiparametric models, like the multi-index and the partially linear models, and models with shape constraints, like monotone regression

models. Another feature of the test is that it allows both the response variable and the covariate be multivariate which means that multiple regression curves can be tested simultaneously. The test also allows the presence of infinite dimensional nuisance parameters in the model to be tested. It is shown that the empirical likelihood test statistic is asymptotically normally distributed under certain mild conditions and permits a wild bootstrap calibration. Despite the fact that the class of models which can be detected consistently by the proposed test is very large, the empirical likelihood test enjoys good power properties against departures from a hypothesized model within the class.

References

[1] S.X. Chen and I. van Keilegom *A goodness-of-fit test for parametric and semiparametric models in multiresponse regression*, Discussion paper DP0616, Institute of Statistics, Université catholique de Louvain (http://www.stat.ucl.ac.be/ISpub/ISdp.html).

Forward Stagewise Regression and the Lasso

Günther Walther

(joint work with Trevor Hastie, Jonathan Taylor, Rob Tibshirani)

The Lasso [1] is a method for regularizing a least squares regression. The Lasso fits a linear model

$$\beta(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

through the criterion

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le s.$$

We consider forward stagewise algorithms for solving least squares regression problems. The iterative nature of these algorithms makes them difficult to analyze. We show that forward stagewise (in the limit for infinitesimal step sizes) can be characterized as a *monotone* version of the Lasso: If we expand the predictor set $\mathbf{X}$ to $\tilde{\mathbf{X}} = [\mathbf{X} : -\mathbf{X}]$, then Lasso solves

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - [\sum_j x_{ij}\beta_j^+ - \sum_j x_{ij}\beta_j^-])^2 \quad \text{subject to } \beta_j^+, \beta_j^- \ge 0, \sum_{j=1}^{p} |\beta_j| \le s.$$

If one requires additionally that the paths $\beta_j^+(s), \beta_j^-(s)$ be non-decreasing in s for all j, then the β-paths (collapsed to the original predictor set $\mathbf{X}$) of this monotone Lasso coincide with the paths of forward stagewise regression.

We also study a condition under which the coefficient paths of the Lasso are monotone, in which case it can be shown that the different algorithms coincide. Finally, we compare the Lasso and forward stagewise procedures in a simulation study involving a large number of correlated predictors.

References

[1] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc. B **58** (1996), 267–288.

Exploiting Non-Data Information

Roger Wets

The major objective of this analysis is to lay the groundwork to include in the formulation of statistical estimation problems information beyond that provided by the observed samples. Of course, this is by no means the first article dealing with related issues! To begin with, every parametric estimation problem includes in its formulation significant non-data information. Even, in the formulation of non-parametric problems, there a large number of papers that deal with various ways to include non-parametric information. For example, simply assuming that the distribution of the random phenomena can be described by a density function implicitly includes non-data information in the formulation of the estimation problem. But, there is a large literature that goes much beyond that. As a few examples, one could refer to the work of Groenenboom, Jongbloed and Wellner on how to include shape information, Thompson and Tapia and Wahba on to include smoothness information about the density function, and so on. And then there is the extensive litterature that deals with Bayesian statistics there is presumed to be probabilistic information about the neighborhood of the (true) distribution.

What's different here is that we introduce a general framework that applies to a wide variety, if not all, situations when there is non-data information availale, and that moreover, leads us to numerical procedures that can take advantage of such additional information; and this applies to the parametric as well as the non-parametric case. Essentially, this approach allows us to include in the formulation of the estimation problem *any* information one might have about the stochastic phenomena.

In this work we present a strong law of large numbers for nonparametric constrained maximum likelihood. The consistency result relies on convergence of the estimation problems to a limit problem, whose solution is the true density, with repect to the hypo-distance topology τ_{aw}, also called the Attouch-Wets topology. Associated with this topology are distance estimates which ultimately should allow us allow us to quantify the convergence rate for this estimation approach.

We begin by illustrating the approach via some examples, basically how the method would be used in generating density estimates given a small collection of samples coming from an exponential distribution. A separate report will deal extensively with a specific implementation of the strategy laid out here; in fact, an alternative to the one based on using basis functions or wavelets. The description of the general framework of the problem that we consider as well as the strategy that will be followed to obtain consistency is dealt with form a very comprehensive viewpoint. The methodology relies on approximation theory for optimization

problems and some compactification argument via the Hilbert-Schmidt embedding.

Estimating Dark Matter Distributions

Michael Woodroofe
(joint work with Xiao Wang, Jayanta Palm, Matthew Walker, Mario Mateo)

In addition to the mass that astronomers can see (mostly in the form of stars), there is matter that they cannot see but which has to be there. The dwarf spheroidal galxies in the neigborhood of the Milky Way provide a good illustration. They are very dim with about $10^6 - 10^7$ stars, but spread over a fairly large area, roughly $2 - 6$ kpc: If all matter present were visible, then there would not be enough mass to hold the system together (and, so, the dwarf spheroidals would not exist). The missing mass is called dark matter. Most of the mass in the dwarf spheroidals is thought to be dark matter.

The talk explains how estimate the distribution of total mass (including dark matter) from a sample of stars and, especially, the role of shape resricted estimation in the estimation process. Statistically, this is an inverse problem with missing data. The available data consists of velocity measurments in the line of sight from earth and positions of the stars projected on the plane orthognal to the line of sight. Thus, one position coordinate and two velocity coordinates are missing. If the mass distribution were known, then the positions and velocities of stars would be determined from the equations of motion in classical mechanics. The inverse problem is to recover the distribution of mass from a sample of observed (projected position) and (line of sight) velocities.

The inverse problem is solved by Jean's Equations (also from classical mechanics) which expresses the distribution of (total) mass in terms of the marginal density of position and the velocity dispersion, the conditional expeptation of squared velocity given position. Unsurprisingly, these equations simply in the presence of spherical symmetry and isotropy. That leaves the missing data. The marginal density and velocity dispersion can be expressed as an Abel transformation of the density of observed position, and the conditional expection of observed velocity given observed position. The Abel transformation is easily inverted and, so, the distribution of mass can be expressed in terms of the distrrbution of observables.

The inverse Abel transformation has two important properties. From Jean's Equation, it is a convex function of its argument; and it can be expressed as an expectation and, so, admits an unbiased estimator. Viewed as function, the unbiased is highly irregular with lots of infinite discontinuities, but becomes regular when when the convexity requirement is imposed. The regularized estimator of the inverse Abel Transformation can then be used to construct estimators of velocity dispersions and the mass distributions. The process is illustrated by simulations and application to data from Fornax, one of the dwarf spheroidal galaxies.

The non-parametric approach described above is made possible by recent advances in instrumentation which will make samples of 1500 or more stars available from each of the dwarf speroidals. Early researchers had only samples of size 50 and less and, so, were forced to used parametric models. While formal goodness of fit tests are still to be developed, our non-parametric estimates to not appear to be consistent with the parametric ones (the traumatic intraoccular test), and our estimates of the total mass of Fornax is about 3 times larger than earlier ones.

Rate consistency of the LASSO in model dimension and bias

Cun-Hui Zhang
(joint work with Jian Huang)

The linear regression model can be written as

$$\mathbf{y} = \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \varepsilon = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{x}_j$ are the columns of the design matrix $\mathbf{X} \equiv (x_{ij})_{n\times p}$, $\beta \equiv (\beta_1, \ldots, \beta_p)'$ is the vector of unknown regression coefficients, and ε is the error vector. For a given penalty level $\lambda \geq 0$, the LASSO (Tibshirani, 1996) estimator of β is

$$\widehat{\beta} \equiv \widehat{\beta}(\lambda) \equiv \operatorname*{argmin}_{\beta} \Big\{ \|\mathbf{y} - \mathbf{X}\beta\|^2/2 + \lambda\|\beta\|_1 \Big\}, \tag{2}$$

where $\|\cdot\|$ is the Euclidean distance and $\|\beta\|_1 = \sum_j |\beta_j|$ is the ℓ_1 norm. We consider

$$\widehat{A} \equiv \widehat{A}(\lambda) \equiv \Big\{ j \leq p : \widehat{\beta}_j \neq 0 \Big\}, \quad \widehat{q} \equiv \widehat{q}(\lambda) \equiv \big|\widehat{A}\big|, \tag{3}$$

as the model selected by the LASSO and its dimension. Compared to the classical variable selection methods such as subset selection, the LASSO has two advantages. First, the selection process in LASSO is based on continuous trajectories of regression coefficients as functions of the penalty level and hence more stable than subset selection methods. Second, the LASSO is computationally feasible for high-dimensional data (Osborne, Presnell and Turlach, 2000; Efron, Hastie, Johnstone and Tibshirani, 2004). In contrast, computation in subsect selection is combinatorial and not feasible when p is large.

Meinshausen and Buhlmann (2006) showed that, for neighborhood selection in Gaussian graphic models, under a neighborhood stability condition, the LASSO is consistent even when the number of variables is of greater order than the sample size. Zhao and Yu (2006) formalized the neighborhood stability condition in the context of linear regression as a strong irrepresentable condition. They showed that under this condition, the LASSO selects exactly the set of non-zero regression coefficients, provided that these coefficients are bounded away from zero at certain rate. In this paper, the regression coefficients outside an ideal model are assumed to be small but not necessarily zero. Under a partial Riesz condition on the

correlation of design variables, we prove that the LASSO selects a model of the right order of dimensionality and controls the bias of the selected model at a level determined by the contributions of small regression coefficients and threshold bias. Consequently, the LASSO selects all coefficients greater than a threshold level determined by the controlled bias of the selected model. An interesting aspect of our results is that the logarithm of the number of variables can be of the same order as the sample size for certain random dependent designs.

We study model selection properties of the LASSO under a sparsity condition on β and a partial Riesz condition (PRC) on $\mathbf{X}$. The sparsity condition asserts

$$\sum_{j=q+1}^{p} \big|\beta_{(j)}\big| \le \eta_1, \quad \text{where } |\beta_{(1)}| \ge \cdots \ge |\beta_{(p)}|. \tag{4}$$

Compared with the typical sparsity assumption for variable selection with $\eta_1 = 0$ and $|\beta_{(q)}|$ bounded away from 0 at a specified rate, (4) is mathematically weaker and much more realistic as it specifies a connected set in the parameter space $\mathbb{R}^p$ of β. Under (4), a sensible goal is to select a sparse model which fits the mean vector $\mathbf{X}\beta$ well. The natural definition of the sparsity of the selected model is $\widehat{q} = O(q)$. We say that the selected model fits the mean $\mathbf{X}\beta$ well if $\widetilde{B} \equiv \big\|(\mathbf{I} - \widehat{\mathbf{P}})\mathbf{X}\beta\big\|$ is small, where $\widehat{\mathbf{P}}$ is the projection from $\mathbb{R}^n$ to the linear span of the set of the selected variables $\mathbf{x}_j$. The case of orthonormal design (Donoho and Johnstone, 1994) indicates that under the conditions we impose, the maximum of the following three quantities is a reasonable benchmark for $\widetilde{B}^2$:

$$\lambda\eta_1,\ \eta_2^2,\ \frac{q\lambda^2}{n}, \quad \text{where } \eta_2 \equiv \max_{A\subset A_0} \Big\| \sum_{j\in A} \beta_j \mathbf{x}_j \Big\| \le \max_{j\le p} \big\|\mathbf{x}_j\big\|\eta_1. \tag{5}$$

Thus, we say that the LASSO is rate consistent in model selection if $\widehat{q} = O(q)$ and $\widetilde{B}^2 = O(1)\max\big(\eta_1\lambda, \eta_2^2, q\lambda^2/n\big)$. We prove the rate consistency of the LASSO under (4) and the PRC which controls the range of eigenvalues of covariate matrices of subsets of a fixed number of design variables. Define $\mathbf{X}_A = (\mathbf{x}_j, j \in A)$ and $\mathbf{\Sigma}_A = \mathbf{X}_A'\mathbf{X}_A/n$. The design matrix $\mathbf{X}$ satisfies the PRC with rank q^* and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \le \frac{\|\mathbf{X}_A\mathbf{v}\|^2}{n\|\mathbf{v}\|^2} \le c^*, \quad \forall\ A \text{ with } |A| = q^* \text{ and } \mathbf{v} \in \mathbb{R}^{q^*}. \tag{6}$$

Since $\|\mathbf{X}_A\mathbf{v}\|^2/n = \mathbf{v}'\mathbf{\Sigma}_A\mathbf{v}$, all the eigenvalues of $\mathbf{\Sigma}_A$ are inside the interval $[c_*, c^*]$ under (6) when the size of A is no greater than q^*.

In terms of the scale invariant ratios $r_1^2 \equiv c^*\eta_1 n/(q\lambda)$ and $r_2^2 \equiv c^*\eta_2^2 n/(q\lambda^2)$ of the quantities in (5) and $C \equiv c^*/c_*$ of the upper and lower bounds in (6), we explicitly express in our theorem the $O(1)$ for rate consistency in variable selection as

$$\begin{aligned} M_1^* &\equiv M_1^*(\lambda) \equiv 2 + 4r_1^2 + 4\sqrt{C}r_2 + 4C, \\ M_2^* &\equiv M_2^*(\lambda) \equiv \frac{8}{3}\Big\{\frac{1}{4} + r_1^2 + r_2\sqrt{C}\big(\sqrt{2} + \sqrt{2C}\big) + C\Big(\frac{1}{2} + \frac{4}{3}C\Big)\Big\}, \end{aligned}$$

$$M_3^* \equiv M_3^*(\lambda) \equiv \frac{8}{3}\Big\{\frac{1}{4} + r_1^2 + r_2\sqrt{C}(1+2\sqrt{1+C}) + \frac{3r_2^2}{4} + C\Big(\frac{5}{6}+\frac{2}{3}C\Big)\Big\}.$$

We define a lower bound for the penalty level as $\lambda_* \equiv \min\big\{\lambda : M_1^*(\lambda)q+1 \le q^*\big\}$. With this λ_* and the c^* of the PRC, we consider the LASSO path in the interval

$$\max\Big(\lambda_*, 2\sqrt{2(1+c_0)c^* n\log(p\vee a_n)}\Big) \le \lambda \le \lambda^* \equiv \max_{j\le p}\big|\mathbf{x}_j'\mathbf{y}\big| \tag{7}$$

with $c_0 \ge 0$ and large a_n satisfying $p/(p\vee a_n)^{1+c_0} \approx 0$.

Theorem 1. *Suppose $\epsilon \sim N(0,\sigma^2\mathbf{I})$ and (4) and (6) hold. Then, there exists a set Ω_1 in the sample space of $(\mathbf{X},\epsilon)$ such that*

$$P\Big\{(\mathbf{X},\epsilon)\in\Omega_1\Big\} \ge 2-\exp\Big(\frac{2p}{(p\vee a_n)^{1+c_0}}\Big) \approx 1-\frac{2p}{(p\vee a_n)^{1+c_0}}$$

and the following assertions hold in the event $(\mathbf{X},\epsilon)\in\Omega_1$ for all λ satisfying (7):

$$\begin{aligned} \widehat{q}(\lambda) &\le M_1^*(\lambda)q, \\ \widetilde{B}^2(\lambda) &\le M_2^*(\lambda)q\lambda^2/\{c^*n\}, \\ \sum_{j\notin A_0}\big|\beta_j\big|^2 I\{\widehat{\beta}_j(\lambda)=0\} &\le M_3^*(\lambda)q\lambda^2/\{c^*c_*n^2\}. \end{aligned}$$

The partial Riesz condition may hold for large random design matrices with p as large as e^{an} for certain $a>0$. Suppose $(y_i, x_{ij}, j\le p)$ are iid copies of a Gaussian vector $(Y,\xi_{k_j}, j\le p)$ and that the Riesz condition holds for the infinite sequence of random variables ξ_k: $\rho_*\|\mathbf{b}\|_2^2 \le E\Big|\sum_{j=1}^\infty b_j\xi_j\Big|^2 \le \rho^*\|\mathbf{b}\|_2^2$, where $\mathbf{b}=\{b_j, j\ge 1\}$.

Proposition 1. *Suppose that the n rows of a random matrix $X_{n\times p}$ are iid copies of a sub-vector of a zero-mean Gaussian sequence $\{\xi_k\}$ satisfying the above Riesz condition. Let ϵ_k, $k=1,\ldots,4$, be positive constants in $(0,1)$ satisfying $2\epsilon_1+3\epsilon_2\le 1$ and $\epsilon_3+\epsilon_4 = \{\epsilon_2-\log(1-\epsilon_2)\}/2$. Then, for $c_*=\tau_*\rho_*$, $c^*=\tau^*\rho^*$ and $q^*\le\min(p,\epsilon_1 n)$ satisfying $\log\{\binom{p}{q^*}(2q^*-1)\}\le\epsilon_3 n$, the partial Riesz condition (6) holds with probability greater than $1-2e^{-n\epsilon_4}$, where $\tau_*=1+\epsilon_1-\sqrt{\epsilon_1+\epsilon_2}(\sqrt{1+\epsilon_2}+\sqrt{1-\epsilon_2})$ and $\tau^*=(\sqrt{1+\epsilon_2}+\sqrt{\epsilon_1+\epsilon_2})^2$.*

Acknowledgment: Cun-Hui Zhang's research is partially supported by the National Science Foundation and National Security Agency.

References

[1] D.L. Donoho and I. Johnstone, *Minimax risk over ℓ_p-balls for ℓ_q-error*, Probab. Theory Rel. Fields **99** (1994), 277–303.

[2] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), 407–499.

[3] N. Meinshausen and P. Buhlmann, *High dimensional graphs and variable selection with the Lasso*, Ann. Statist. **34** (2006), 1436–1462.

[4] M. Osborne, B. Presnell and B. Turlach, *On the lasso and its dual*, J. Comp. Graph. Statist. **9** (2000) (2), 319–337.

[5] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. B **58** (1996), 267–288.

[6] P. Zhao and B. Yu, *On model selection consistency of LASSO*, Technical report No. 702. Dept. of Statistics, UC Berkeley.

Reporter: Lutz Dümbgen

Participants

Prof. Dr. Dragi Anevski
Department of Mathematics
Chalmers University of Technology
S-412 96 Göteborg

Dr. Fadoua Balabdaoui
Centre de Recherche de
Mathematiques de la Decision
CEREMADE
Universite Paris IX
F-75016 Paris

Prof. Dr. Moulinath Banerjee
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
USA

Prof. Dr. Misha Belkin
Dept. of Computer and Information
Science
Ohio State University
2015 Neil Ave.
Columbus, OH 43210-1277
USA

Prof. Dr. Rudolf J. Beran
Department of Statistics
University of California
Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Lucien Birge
Laboratoire de Probabilites-Tour 56
Universite P. et M. Curie
4, Place Jussieu
F-75252 Paris Cedex 05

Melanie Birke
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Dr. Nicolai Bissantz
Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Dr. Peter Bühlmann
Seminar für Statistik
ETH-Zürich
LEO C 17
CH-8092 Zürich

Prof. Dr. T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
USA

Prof. Dr. Eric A. Cator
Delft University of Technology
Faculty of Information Technology
and Systems / Dept. of CROSS
Mekelweg 4
NL-2628 CD Delft

Prof. Dr. P. Laurie Davies
Fachbereich Mathematik
Universität Duisburg-Essen
45117 Essen

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Prof. Dr. Lutz Dümbgen
Mathematische Statistik
und Versicherungslehre
Universität Bern
Sidlerstraße 5
CH-3012 Bern

Prof. Dr. Michael Eichler
Dept. of Quantitative Economics
University of Maastricht
Postbus 616
NL-6200 MD Maastricht

Prof. Dr. Ursula Gather
Fachbereich Statistik
Universität Dortmund
44221 Dortmund

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH-Zentrum Zürich
LEO D2
Leonhardstr. 27
CH-8092 Zürich

Prof. Dr. Piet Groeneboom
Department of Mathematics
Vrije University
De Boelelaan 1083 a
NL-1081 HV Amsterdam

Prof. Dr. Arnold Janssen
Mathematisches Institut
Heinrich-Heine-Universität
Universitätsstr. 1
40225 Düsseldorf

Dr. Geurt Jongbloed
Faculteit Wiskunde en Informatica
Vrije Universiteit Amsterdam
De Boelelaan 1081
NL-1081 HV Amsterdam

Prof. Dr. Roger W. Koenker
Dept. of Economics
University of Illinois
330 Commerce Bldg. (West)
Champaign, IL 61820
USA

Dr. Kirill Kopotun
Department of Mathematics
University of Manitoba
422 Machray Hall
Winnipeg MB, R3T 2N2
CANADA

Prof. Dr. Arne Kovac
School of Mathematics
University of Bristol
University Walk
GB-Bristol BS8 1TW

Dr. Vladimir N. Kulikov
Ing Financial Markets
TR 00.21
Foppinga Dreft 7
NL-1102 BD Amsterdam

Dr. Rik Lopuhaä
Delft Institute of Applied
Mathematics
Delft University of Technology
Mekelweg 4
NL-2628 CD Delft

Prof. Dr. Mark Low
University of Pennsylvania
Department of Statistics
The Wharton School
Philadelphia PA 19104-6302
USA

Marloes Maathuis
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
USA

Prof. Dr. Enno Mammen
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68131 Mannheim

Dr. Nicolai Meinshausen
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley, CA 94720-3860
USA

Prof. Dr. Mary Meyer
Department of Statistics
University of Georgia
Athens, GA 30602
USA

Prof. Dr. Ivan Mizera
Department of Mathematical and
Statistical Sciences
University of Alberta
632 CAB
Edmonton AB T6G 2G1
Canada

Prof. Dr. Klaus-Robert Müller
Fraunhofer Institut FIRST
Intelligent Data Analysis Group
(IDA)
Kekulestr. 7
12489 Berlin

Prof. Dr. Axel Munk
Institut f. Mathemat. Stochastik
Georg-August-Universität Göttingen
Maschmühlenweg 8-10
37073 Göttingen

Prof. Dr. Wolfgang Polonik
Department of Statistics
University of California
Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Yaacov Ritov
Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905
ISRAEL

Angelika Rohde
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg

Dr. Kaspar Rufibach
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
USA

Regine Scheder
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Prof. Dr. Jiayang Sun
Dept. of Mathematics and Statistics
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106-7058
USA

Prof. Dr. Alexander B. Tsybakov
Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Aad W. van der Vaart
Faculteit Wiskunde en Informatica
Vrije Universiteit Amsterdam
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Prof. Dr. Ingrid Van Keilegom
Institut de Statistique
Universite Catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve

Prof. Dr. Günther Walther
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
USA

Prof. Dr. Jon A. Wellner
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
USA

Prof. Dr. Roger Wets
Department of Mathematics
University of California
1, Shields Avenue
Davis, CA 95616-8633
USA

Prof. Dr. Michael B. Woodroofe
Department of Statistics
The University of Michigan
1447 Mason Hall
Ann Arbor, MI 48109-1027
USA

Prof. Dr. Cun-Hui Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen rd.
Piscataway, NJ 08854
USA