

# Adaptive estimation of linear functionals by model selection

Carenne Ludeña, IVIC

joint work with Beatrice Laurent and Clementine Prieur,  
INSA, Toulouse, France

IX Escuela de Probabilidad y Estadística  
Guanajuato, January 2007

# What we will talk about

Stating the problem

Gaussian Isonormal processes

Sharp estimation of linear functionals

Adaptive model selection

Gaussian Model selection: Birgé and Massart

"Lepski's" method

Main results

Applications and simulations

Multiresolution analysis: estimating a function at a fixed point

Simulations

## Stating the problem

Gaussian Isonormal processes

Sharp estimation of linear functionals

Adaptive model selection

Gaussian Model selection: Birgé and Massart

"Lepski's" method

Main results

Applications and simulations

Multiresolution analysis: estimating a function at a fixed point

Simulations

## The problem

- ▶ Consider the usual regression model

$$y_i = x_i + \zeta_i, \quad \zeta_i \sim N(0, \sigma^2), \quad i = 1, \dots, n.$$

- ▶ We have  $T$  a linear operator
- ▶ Problem: estimate  $u = Tx$ .
- ▶ Estimation error  $R(\hat{u}) = E\|\hat{u} - u\|^2$ .
- ▶ Goal: adaptive estimation. That is, construct an estimator that achieves (almost) optimal rates.

## Optimality: minimax rates

- ▶ Let  $u$  be our target and  $\hat{u}$  an estimator.
- ▶ Assume  $u \in \mathcal{F}$  a certain class of functions.
- ▶ Let  $d$  be some distance, or pseudo-distance over  $\mathcal{F}$ .
- ▶ Define the estimator's rate over  $\mathcal{F}$  as

$$r(\hat{u}, \mathcal{F}) = \sup_u E_u[d^q(u, \hat{u})].$$

- ▶ Define the minimax rate over  $\mathcal{F}$  as the infimum over the set of all estimators (sometimes we will be interested in subsets of estimators such as linear or affine estimators)

$$r_M(\mathcal{F}) = \inf_{\hat{u}} r(\hat{u}, \mathcal{F}).$$

Stating the problem

## Gaussian Isonormal processes

Sharp estimation of linear functionals

Adaptive model selection

Gaussian Model selection: Birgé and Massart

"Lepski's" method

Main results

Applications and simulations

Multiresolution analysis: estimating a function at a fixed point

Simulations

## Building up: a more general framework

- ▶ More generally, consider the following model:

$$Y(t) = \langle s, t \rangle + \frac{\sigma}{\sqrt{n}} L(t) \quad t \in \mathbb{H},$$

- ▶  $\mathbb{H}$  is an Hilbert space: with scalar product  $\langle \cdot, \cdot \rangle$
- ▶  $L$  is a centered Gaussian isonormal process ...

## Gaussian isonormal processes

- ▶ That is,  $L$  maps isometrically  $\mathbb{H}$  onto some Gaussian subspace of  $\mathbb{L}_2(\Omega)$ .
- ▶ For all  $t, u \in \mathbb{H}$ ,  $\text{Cov}(L(t), L(u)) = \langle t, u \rangle$ .

## Examples

- ▶ Finite dimensional Gaussian regression

- ▶ Observations:

$$Y_i = s_i + \varepsilon_i, i = 1, \dots, n$$

where  $(\varepsilon_1, \dots, \varepsilon_n)$  are independent standard normal variables.

- ▶ The space:  $\mathbb{H} = \mathbb{R}^n$
    - ▶ Scalar product:  $\langle x, y \rangle = \frac{1}{n} \sum_{i=1}^n x_i y_i$  and set  $s = (s_1, \dots, s_n)$ .
    - ▶ Original model: set  $t = (t_1, \dots, t_n) \in \mathbb{R}^n$ ,  $Y(t) = \frac{1}{n} \sum_{i=1}^n t_i Y_i$  and  $L(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n t_i \varepsilon_i$ .

► The Gaussian sequence model.

- Observations:

$$Y_\lambda = \beta_\lambda + \frac{1}{\sqrt{n}}\varepsilon_\lambda, \lambda \in \mathbb{N}^*,$$

where  $(\varepsilon_\lambda)_{\lambda \in \mathbb{N}^*}$  is a sequence of independent standard normal variables.

- Space:  $\mathbb{H} = l_2(\mathbb{N}^*)$
- Scalar product:  $\langle \beta, \gamma \rangle = \sum_{\lambda \in \mathbb{N}^*} \beta_\lambda \gamma_\lambda$

## Relation to the original model

- ▶ Set  $\mathbf{s} = (\beta_\lambda)_{\lambda \in \mathbb{N}^*}$
- ▶ Set for any  $\mathbf{t} = (\alpha_\lambda)_{\lambda \in \mathbb{N}^*} \in \mathbb{H}$ ,  $Y(\mathbf{t}) = \sum_{\lambda \in \mathbb{N}^*} \alpha_\lambda Y_\lambda$
- ▶ Set  $L(\mathbf{t}) = \sum_{\lambda \in \mathbb{N}^*} \alpha_\lambda \varepsilon_\lambda$

► The multivariate white noise model.

► Observations:

$$Z(x) = \int_{[0,1]^d} \mathbf{1}_{[0,x_1] \times \dots \times [0,x_d]}(u) s(u) du + \frac{1}{\sqrt{n}} W(x)$$

for all  $x = (x_1, \dots, x_d) \in [0, 1]^d$ , where  $W$  is the standard Wiener Process on  $D = [0, 1]^d$ .

- Space:  $\mathbb{H} = \mathbb{L}_2([0, 1]^d)$
- Scalar product:  $\langle t, s \rangle = \int_D t(u) s(u) du$
- Original model: set  $Y(t) = \int_D t(u) dZ(u)$  and  $L(t) = \int_D t(u) dW(u)$ .

So, coming back to what do we want to do

- ▶ Let  $T$  be a linear functional defined over  $\mathcal{S} \subset \mathbb{H}$  (for example evaluating  $s$  or its derivatives at a fixed point).
- ▶ Want to estimate  $T(s)$ , based on  $(Y(t), t \in \mathbb{H})$ .
- ▶ Our main goal will be developing procedures which adapt to the smoothness of the underlying function  $s$  in the framework of **model selection** as proposed by Barron et al.

- ▶ Before discussing the model selection framework, we develop an example. This will allow us to discuss problems related to the estimation and introduce some general results for estimation of linear functionals.

An example: an initial boundary value problem for partial differential equations (Chow et. al.)

- ▶ Wave equation for the vibration of an elastic medium in domain  $D \in \mathbb{R}^3$ :

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u + \Theta(t, x), \quad 0 < t < T, \quad x \in D$$

- ▶  $c$  is a constant
- ▶  $\Delta = \sum_{i=1}^3 \frac{\partial^2}{\partial x_i^2}$  is the Laplacian
- ▶  $\Theta$  is an unknown source function.
- ▶ The problem consists of determining the source  $\Theta$  upon observing  $u(t)$ ,  $0 < t < T$ .

- ▶ Call  $\mathcal{L}(x) = \frac{\partial^2}{\partial t^2} - c^2 \Delta$
- ▶ The problem is equivalent to estimating  $\mathcal{L}u = \Theta$  based on  $u$ .
- ▶ In general,  $\mathcal{L}$  can be any linear partial differential operator and  $u$  is assumed to satisfy certain regularity conditions
- ▶ However, we would like that the knowledge of such conditions (or those of  $\Theta$ ) need not be essential in order to construct optimal estimators.

## Estimating $\Theta$ (Kernel based estimators, Chow et al.)

- ▶ White noise model:

$$Z_n(x) = \int_{[0,1]^d} \mathbf{1}_{[0,x_1] \times \dots \times [0,x_d]}(u) s(u) du + \frac{1}{\sqrt{n}} W(x)$$

Set  $Z_n(dx)$  the measure generated by  $Z_n$

- ▶ The Kernel:  $\phi_h(x) = \prod_{i=1}^3 \frac{1}{h_i} \phi_i(\frac{x_i}{h_i})$ , with  $\phi_i$  density functions.
- ▶ The estimator (in the simpler case  $\mathcal{L} = \partial_x^m$ ):

$$\hat{\Theta}(x) = \int \partial_y^m \phi_h(x - y) Z_n(dx)$$

- ▶ Optimal rates if  $h_i$  is chosen to the (known) regularity of  $\Theta$ .

## Doing better:sharp estimation

- ▶ Optimal Kernels (Donoho and Liu, Donoho and Low, Donoho): The kernel can be chosen as to provide optimal linear estimators. In general, these kernels depend on the linear functional  $T$  and on the (assumed known) regularity of  $\Theta$  as measured by a certain regularity functional  $\rho_\nu$  (semi-norm) depending on a regularity parameter  $\nu$  (think about the norm of the  $[\nu]$ -th derivative, for example)

## Sharp estimation of linear functionals

- ▶ The estimators: Given the optimal kernel  $K(\nu, n, T)$ , construct the estimator  $\hat{T}_\nu = \int K(\nu, n, T)(y) Y(dy)$ .
- ▶ Over a general class of functions  $\mathcal{F}$ , optimal MS rates are of order  $w^2(1/\sqrt{n})$ . Here  $w$  is the modulus of continuity of  $T$  over the class:

$$w(\nu) = \sup\{|T(f) - T(g)| : \|f - g\|_2 \leq \nu, f, g \in \mathcal{F}\}.$$

- ▶ In this case also (Donoho and Low) it is possible to construct optimal kernels.

## Adaptive estimation: classes of functions

- ▶ Consider the linear functional  $T(s) = s(x_0)$ .
- ▶ Problem: estimating  $T(s)$  when we do not previously know the regularity of function  $s$  (as measured by  $\rho_\nu$ , for example).
- ▶ Lepski showed it is necessary to include a logarithmic factor in the mean squared error when dealing simultaneously with two Lipschitz classes. That is, adaptive rates are worse than rates over a specified class.
- ▶ Cai and Low show that for two convex classes of functions  $\mathcal{F}_1, \mathcal{F}_2$  optimal rates also depend on an additional logarithmic factor.

## Adaptive estimation: the "Lepski" way

- ▶ Klemela and Tsybakov, Cai and Low (based on "Lepski's" method): Consider a class of functions  $\mathcal{F}_\nu = \{f : \rho_\nu(f) \leq L\}$ . Set  $\mathcal{F} = \{\mathcal{F}_\nu\}_\nu$ .
- ▶ Consider a discretization of the set of possible values of  $s$ .
- ▶ Choose the right  $s$  by

$$\hat{s} = \max\{s, | \hat{T}_s - \hat{T}_{s'} | \leq \eta(s'), s' < s\},$$

where  $\eta(s)$  is the optimal rate for fixed  $s$  (times a logarithmic factor in  $n$  which is the price we pay for adaptivity).

- ▶ We would like to do the same, considering a sequence of nested subspaces and model selection techniques.
- ▶ We start by giving a simple introduction to the problem of Gaussian model selection

Stating the problem

Gaussian Isonormal processes

Sharp estimation of linear functionals

**Adaptive model selection**

Gaussian Model selection: Birgé and Massart

"Lepski's" method

Main results

Applications and simulations

Multiresolution analysis: estimating a function at a fixed point

Simulations

- ▶ Let  $(S_m, m \in \mathcal{M})$  be a collection of linear subspaces of  $\mathbb{H}$ .
- ▶ Given an orthonormal basis  $(\phi_\lambda, \lambda \in \Lambda_m)$  of  $S_m$ , let

$$\hat{s}_m = \sum_{\lambda \in \Lambda_m} Y(\phi_\lambda) \phi_\lambda,$$

be the projection of  $Y$  over  $S_m$ .

- ▶ Remark that since

$$\hat{s}_m = \operatorname{argmin}_{v \in S_m} \left( \|v\|^2 - 2Y(v) \right),$$

$\hat{s}_m$  is independent on the chosen basis.

- ▶ Let  $s_m$  be the projection of  $s$  over  $S_m$ .

- ▶ Calculating expectations, we obtain

$$E\|\hat{s}_m - s\|^2 = \|s_m - s\|^2 + \frac{\sigma^2|m|}{n},$$

where  $|m|$  is the cardinality of  $S_m$ .

- ▶ If  $s$  is assumed to belong to a given class of functions, we could then calculate  $\|s_m - s\|^2$  for each  $m \in \mathcal{M}$  and choose a model  $\hat{m}$  which attains optimal rates. Then select  $\hat{s} = \hat{s}_{\hat{m}}$

- ▶ However, since we do not have much information about  $s$  we would like to develop procedures of selecting  $m$  in order to obtain optimal adaptive MS rates (we won't discuss the issue of sharp estimation):

$$E\|\hat{s}_{\hat{m}} - s\|^2 \leq C \inf_m \{E\|\hat{s}_m - s\|^2\},$$

with  $C > 1$  depending on the family of models, but independent of  $s$ .

- ▶ The least squares estimator of  $s_m$  (or the least squares estimator of  $s$  over  $S_m$ ) is the minimizer over  $t \in S_m$  of

$$\gamma(t) = \|t\|^2 - 2Y(t).$$

- ▶ Long known procedure for selecting  $m$ : penalized criterion, that is, minimize over  $m \in \mathcal{M}$

$$\gamma(\hat{s}_m) + \text{pen}(m) = -\|\hat{s}_m\|^2 + \text{pen}(m).$$

- ▶ First examples: Mallows'  $C_p$  ( $pen(m) = 2\frac{\sigma^2}{n}|m|$ ), Akaike's AIC (maximizing the maximal log-likelihood minus the number of parameters).
- ▶ Basic idea: construct an unbiased estimator of the risk and minimize.
- ▶ More developed idea (as in Barron et al., Birgé and Massart) choose a proper  $pen(m)$ : that is, in order to obtain

$$E\|\hat{s}_{\hat{m}} - s\|^2 \leq C \inf_m \{E\|\hat{s}_m - s\|^2\},$$

with adequate values for the constant.

## Introducing a non asymptotic point of view

- ▶ Would like to avoid requiring that  $s \in S_m$  or exclude this possibility (as is required by Mallows and Akaike)
- ▶ Would like the list of possible models to depend on the noise error  $\sigma^2/n$ : in general for bigger  $n$  it is natural to consider more complex models.

- ▶ Basic approach (Birgé and Massart, 2001)  
 $pen(m) = \frac{\sigma^2}{n} K |m| (1 + \sqrt{2L_m})^2$ , where  $K > 1$  and  $L_m > 0$  are a sequence of weights.
- ▶ Kraft inequality: the sequence  $\{L_m\}_m$  should control the complexity of the set  $\mathcal{M}$ . That is, it should be such that

$$\sum_{m \in \mathcal{M} \mid |m| > 0} e^{-|m|L_m} < 1.$$

## "Lepski's" method (following Birgé)

- ▶ Assume we have a family of classes  $\{\mathcal{F}_\nu\}_\nu$ , where  $\nu \in \Upsilon$  a bounded subset of  $\mathbb{R}^+$ , which is nondecreasing with respect to  $\nu$ .
- ▶ The minimax rates  $r_M(\nu) = r_M(\mathcal{F}_\nu)$  are continuous with respect to  $\nu$
- ▶ For each  $\nu$  we have an asymptotically minimax estimator  $\hat{u}_\nu$  on  $\mathcal{F}_\nu$ .
- ▶ For big  $n$  and each  $\nu$ ,  $d^q(u, \hat{u}_\nu)$  is close to its expectation.

## "Lepski's" method

- ▶ Assume for each  $n$  we have a finite discretization  $\nu_1 < \dots < \nu_{m(n)}$ .
- ▶ Choose  $\hat{u}_{\nu_j}$  as the adaptive estimator where

$$\hat{j} = \inf\{j \leq m(n) \mid d^q(\hat{u}_{\nu_j}, \hat{u}_{\nu_k}) \leq K r(\hat{u}_{\nu_k}, \nu_k), \text{ for all } j < k \leq m(n)\},$$

for some big enough  $K$ .

- ▶ Lepski shows that  $\hat{u}_{\nu_{\hat{j}}}$  achieves the minimax rates asymptotically over all sets  $\mathcal{F}$ .

## An alternative interpretation (Birgé)

- ▶ If  $u \in \mathcal{F} = \cup_{\nu} \mathcal{F}_{\nu}$ , there exists a first  $\nu(u)$ , such that  $u \in \mathcal{F}_{\nu(u)}$ .
- ▶ Hence,  $\hat{u}_{\nu(u)}$  should be the best estimator of  $u$ .
- ▶ Given a known value of  $n$  (and  $\sigma^2$ ) it should be possible to select an almost best  $\nu(u)$ .
- ▶ Birgé (in an  $L^2$  setting with projection estimators):

$$\hat{j} = \inf \{j \leq m(n) \mid \|\hat{u}_j - \hat{u}_m\|^2 \leq \frac{\sigma^2}{n} (1 + \gamma)(|m| - |j|) + 2K_{m,j}\},$$

for all  $j < m \leq m(n)$ ,

where  $C > 1$ ,  $K_{m,j} \geq [(1 + 2\gamma)\lambda_m(|m| - |j|)]^{1/2} + \lambda_m$  and  $\sum_m e^{-\lambda_m} < \infty$ .

- ▶ Based on these ideas we turn back to our original problem.

Stating the problem

Gaussian Isonormal processes

Sharp estimation of linear functionals

Adaptive model selection

Gaussian Model selection: Birgé and Massart

"Lepski's" method

**Main results**

Applications and simulations

Multiresolution analysis: estimating a function at a fixed point

Simulations

- ▶ Since  $T$  is linear  $E(T(\hat{s}_m)) = T(s_m)$ .
- ▶ Natural idea: estimate  $T(s)$  using  $T(\hat{s}_m)$ , select  $m$ .
- ▶ MS error:

$$E \left[ (T(\hat{s}_m) - T(s))^2 \right] = (T(s_m) - T(s))^2 + E \left[ (T(\hat{s}_m) - T(s_m))^2 \right].$$

- ▶ The variance term is

$$E \left[ (T(\hat{s}_m) - T(s_m))^2 \right] = \frac{\sigma^2}{n} \sum_{\lambda \in \Lambda_m} T^2(\phi_\lambda).$$

- ▶ The bias term  $(T(s_m) - T(s))^2$  depends on  $s$ !

- ▶ Bias problem is an important issue!
- ▶ Recall the regression setting:

$$E\|\hat{s}_m - s\|^2 = \|s_m - s\|^2 + \frac{\sigma^2|m|}{n}$$

- ▶ By Pythagoras we have  $\|s_m - s\|^2 = \|s\|^2 - \|s_m\|^2$ . So an unbiased estimator of the risk depends only on  $\|\hat{s}_m\|^2$ .
- ▶ In the case of estimating  $T(s)$ , this is no longer possible.

## Estimating the bias

- ▶ For each  $m \in \mathcal{M}$ , let

$$\sigma_m^2 = \frac{\sigma^2}{n} \sum_{\lambda \in \Lambda_m} T^2(\phi_\lambda)$$

and

$$\text{pen}(m) = 2x_m \sigma_m^2,$$

- ▶ Set, for each  $j, m \in \mathcal{M}$ ,

$$\sigma_{j,m}^2 = \frac{\sigma^2}{n} \mathbb{E} \left[ \left( \sum_{\lambda \in \Lambda_m} T(\phi_\lambda) L(\phi_\lambda) - \sum_{\lambda \in \Lambda_j} T(\phi_\lambda) L(\phi_\lambda) \right)^2 \right]$$

and

$$H(j, m) = 3x_{j,m}\sigma_{j,m}^2$$

- ▶ where  $(x_{j,m}, (j, m) \in \mathcal{M}^2)$  and  $(x_m, m \in \mathcal{M})$  are sequences of non negative integers.

## Establishing the objective function

- ▶ For each  $m \in \mathcal{M}$  define

$$\widehat{\text{Crit}}(m) = \sup_{j \geq m, j \in \mathcal{M}} \left[ (T(\hat{\mathbf{s}}_m) - T(\hat{\mathbf{s}}_j))^2 - H(j, m) \right] + \text{pen}(m),$$

- ▶ let for all  $m \in \mathcal{M}$ ,

$$\text{Crit}(m) = \sup_{j \geq m} (T(\mathbf{s}_j) - T(\mathbf{s}_m))^2 + \text{pen}(m)$$

and

$$\Gamma(m) = (T(\mathbf{s}_m) - T(\mathbf{s}))^2 + \sigma_m^2 + \sup_{j \leq m} x_{m,j} \sigma_{m,j}^2.$$

- ▶ Define

$$\hat{m} = \inf\{m \in \mathcal{M}, \widehat{\text{Crit}}(m) \leq \inf_{j \in \mathcal{M}} \widehat{\text{Crit}}(j) + \frac{1}{n}\}.$$

- ▶ Define  $m_{\text{opt}}$

$$m_{\text{opt}} = \inf\{m \in \mathcal{M} / \text{Crit}(m) \leq \inf_{l \in \mathcal{M}} \text{Crit}(l) + \frac{1}{n}\}.$$

# Main result

## ► Theorem

*There exists a constant  $\kappa > 0$  such that*

$$E \left[ (T(\hat{s}_{\hat{m}}) - T(s))^2 \right] \leq \kappa (\text{Crit}(m_{opt}) + \Gamma(m_{opt})) \\ + \kappa \left( \sum_{m \in \mathcal{M}} e^{-\chi_m} \sigma_m^2 + 2 \sum_{j \geq m_{opt}} e^{-\chi_{j, m_{opt}}} \sigma_{j, m_{opt}}^2 + \frac{1}{n} \right).$$

## Main tool: a $\chi^2$ inequality

► Lemma

For all  $m \in \mathcal{M}$ , for all  $x > 0$ ,

$$\mathbb{P}(\widehat{\text{Crit}}(m) > 3(\text{Crit}(m) + x)) \leq \sum_{j \geq m, j \in \mathcal{M}} e^{-x_{j,m}} e^{-x/\sigma_{j,m}^2}.$$

Stating the problem

Gaussian Isonormal processes

Sharp estimation of linear functionals

Adaptive model selection

Gaussian Model selection: Birgé and Massart

"Lepski's" method

Main results

**Applications and simulations**

Multiresolution analysis: estimating a function at a fixed point

Simulations

## Estimating a function (or its derivatives) at a fixed point

- ▶ Set  $\mathbb{H} = L_2([0, 1])$ ,  $j_0 \in \mathbb{N}$  and  $\{\mathcal{S}_j, j \geq j_0\}$  a multiresolution analysis with father wavelet  $\varphi$  and mother wavelet  $\psi$ .
- ▶ Let  $d_n \geq j_0$  and  $\mathcal{M} = \mathbb{N} \cap [j_0, d_n]$ . Define

$$\varphi_{j,k}(x) := 2^{j/2} \varphi(2^j x - k), \quad x \in [0, 1], \quad j \geq j_0 \text{ y } k \in \mathbb{Z};$$

$$\psi_{j,k}(x) := 2^{j/2} \psi(2^j x - k), \quad x \in [0, 1], \quad j \geq j_0 \text{ y } k \in \mathbb{Z}.$$

- ▶ For  $J \geq j_0$ ,

$$s_J(x) = \sum_{k \in \mathbb{Z}} \langle s, \varphi_{J,k} \rangle \varphi_{J,k}(x).$$

## Multiresolution analysis: estimating a function at a fixed point

- ▶ Assume that  $\varphi$   $\Psi$  satisfy the following assumptions :
  - (i)  $\exists M \geq 0$  such that  $\text{supp}(\varphi)$  and  $\text{supp}(\Psi)$  belong to  $[-M, M]$ .
  - (ii)  $\exists K \geq 0$  such that  $\|\varphi\|_\infty \vee \|\Psi\|_\infty \leq K$ .
  - (iii)  $\exists N \geq 0$  such that  $\int x^n \Psi(x) dx = 0$  para  $n = 0, \dots, N$ .
  - (iv) We assume that  $\varphi^{(r)}$  exists and is bounded on  $\text{supp}(\varphi)$  by  $K_r$ .

Let  $T(s) = s(x_0)$  for some  $x_0$  en  $[0, 1]$ . Let  $1 \leq q \leq \infty$  and  $0 < \alpha < N + 1$ .

- ▶ Let  $B_\infty^{\alpha, q}([0, 1])$  be a Besov space equipped with the norm  $\|\cdot\|_{\alpha, \infty, q}$ .

## ► Corollary

Let  $d_n \geq (\ln(n)/\ln(2))$  and let  $\mathcal{M} = \{0, \dots, d_n\}$ . Let

$$x_m = (\ln(2) + \varepsilon)(1 + 2r)(m \vee 1), \quad x_{j,m} = (\ln(2) + \varepsilon)(1 + 2r)(j \vee 1).$$

Let  $\hat{m}$  be defined in Theorem 1. There exists some constant  $C$  depending on  $\delta$ ,  $\varepsilon$  and  $\alpha$  such that if  $r < \alpha \leq r + 1$ , then

$$\sup_{s \in B_{\infty}^{\alpha, q}([0, 1])} \mathbb{E} \left[ \left( \hat{s}_{\hat{m}}^{(r)}(x_0) - s^{(r)}(x_0) \right)^2 \right] \leq CL^{\frac{2+4r}{1+2\alpha}} \left( \frac{\ln n}{n} \right)^{\frac{2(\alpha-r)}{2\alpha+1}}$$

- It follows from the results given in Lepski that these rates are optimal.

## Some simulations

- ▶ We include three examples for the finite dimensional regression model. Wavelet based estimators were considered.
  1.  $s_1(x) = 2x \mathbf{1}_{[0, 1]}$  ( Haar basis)
  2.  $s_2(x) = \exp(-30 * |x - 0.75|) + \exp(-30 * |x - 0.25|)$  for  $x \in [0, 1]$  ( Haar basis).
  3.  $s_3(x) = \sin(8\pi x) \mathbf{1}_{0 < x \leq 1/2} + \sin(32\pi x) \mathbf{1}_{1/2 < x \leq 1}$ , ( Daubechies 20 basis).

- ▶ Estimators were constructed based on the simulated values

$$y_i = s_j(i/n) + \sigma \varepsilon_i \quad i = 1, \dots, n \quad j = 1, 2, 3$$

with  $\varepsilon_i \quad i = 1, \dots, n \sim N(0, 1)$  independent, random variables with  $\sigma = 0.2$  and  $n = 256$ . We set  $d_n = 8$ ,  $j_0 = 1$ , and for all  $1 \leq m \leq j \leq d_n$ ,

$$H(j, m) = \sigma^2 j(2^j - 2^m)/n,$$

$$\text{pen}(m) = m2^m \sigma^2 / n.$$

- ▶ To compare our results we used the (global) estimator  $\tilde{s}$  introduced by Baraud, defined for all  $t$  by

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^n (y_i - t(\frac{i}{n}))^2,$$

$\tilde{s} = \operatorname{argmin}_{m \in \mathcal{M}} (\gamma_n(\hat{s}_m) + \operatorname{pen}'(m))$ , with  
 $\operatorname{pen}'(m) = 2.2^m \sigma^2 / n$ .

## Simulations

- ▶ We estimated the quadratic risk of the estimators of  $s_j(i/n)$ ,  $1 \leq i \leq n$ , as in (P 1) and as in Baraud (P 2).
- ▶ Estimation of QR is based on  $N = 5000$  simulations and is defined by

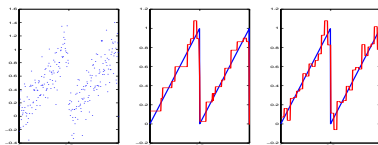
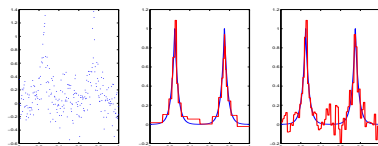
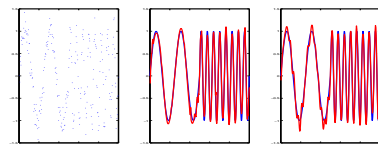
$$\hat{r}^2\left(\frac{i}{n}\right) = \frac{1}{N} \sum_{l=1}^N \left(\hat{s}^{(l)}\left(\frac{i}{n}\right) - s\left(\frac{i}{n}\right)\right)^2$$

with  $s = s_1, s_2$  and  $s_3$  and  $\hat{s}^{(l)}$  the estimator of  $s$  based on the  $l$ -th simulated sample.

## Results

| $s$   | $\max_{1 \leq i \leq n} \hat{r}(i/n)$ |           | $(\frac{1}{n} \sum_{i=1}^n \hat{r}^2(i/n))^{1/2}$ |           |
|-------|---------------------------------------|-----------|---|-----------|
|       | <b>P1</b>                             | <b>P2</b> | <b>P1</b>   | <b>P2</b> |
| $s_1$ | 0.4188                                | 0.8745    | 0.0910  | 0.1035    |
| $s_2$ | 0.2573                                | 0.2972    | 0.0875  | 0.1000    |
| $s_3$ | 0.2406                                | 0.3543    | 0.0931  | 0.1057    |

## Simulations

Figure 1: Estimating function  $s_1$ ,  $n = 2^8$ ,  $\sigma = 0.2$ Figure 2: Estimating function  $s_2$ ,  $n = 2^8$ ,  $\sigma = 0.2$ Figure 3: Estimating function  $s_3$ ,  $n = 2^8$ ,  $\sigma = 0.2$

## Simulations

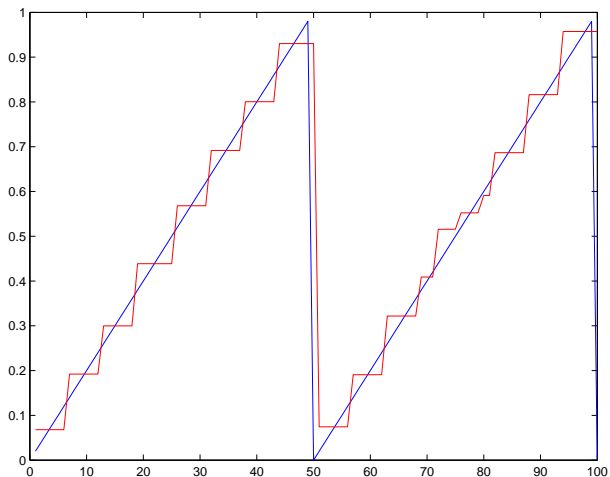


Figure: Function with varying regularity