ADAPTIVE ESTIMATION WITH SOFT THRESHOLDING TYPE PENALTIES

By

Jean-Michel Loubes
Laboratoire de Statistique et Probabilités
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex
France
loubes@cict.fr

and

Sara van de Geer
Mathematical Institute
University of Leiden
P.O. Box 9512
2300 RA Leiden
The Netherlands
geer@math.leidenuniv.nl

Abstract We show that various robust nonparametric regression estimators, such as the least absolute deviations estimator, can be made adaptive (up to logarithmic factors), by adding a soft thresholding type penalty to the loss function. As an example, we consider the situation where the roughness of the regression function is described by a single parameter ρ. The theory is complemented with a simulation study.

1. Introduction

We study the regression model

$$Y_i = \theta_0(z_i) + W_i, \ i = 1, \ldots, n, \tag{1.1}$$

where $Y_1, \ldots, Y_n$ are independent real-valued observations, $z_1, \ldots, z_n$ are covariables with values in some space $\mathcal{Z}$, θ_0 is an unknown regression function, and $W_1, \ldots, W_n$ are measurement

errors. At the values $z_1, \ldots, z_n$, the regression function θ_0 is defined by means of a given convex loss function $\gamma : \mathbf{R} \to \mathbf{R}$. Namely, we require for each $b \in \mathbf{R}$, that the expectation $\mathbf{E}\gamma(Y_i - b)$, is finite and moreover that $b \mapsto \mathbf{E}\gamma(Y_i - b)$ has a unique minimum in $b = \theta_0(z_i)$. Thus, when $\gamma(\xi) = \xi^2$, the measurement errors $W_i = Y_i - \theta_0(z_i)$ $(i = 1, \ldots, n)$, are required to have mean zero and finite variance, and when when $\gamma(\xi) = |\xi|$, then their median is assumed to be zero, etc.

Suppose $\theta_0 \in \Theta$, where the parameter space Θ is a given subset of the set of all real-valued functions on $\mathcal{Z}$. We shall study the regression estimator

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \gamma(Y_i - \theta(z_i)) + \lambda_n^2 I_n(\theta) \right]. \tag{1.2}$$

Here, $I_n(\theta)$ is taken to be the *soft thresholding type penalty* explained below (equation (1.3)), and λ_n^2 is a (to be chosen) smoothing parameter.

Take Q_n as the empirical measure of the covariables:

$$Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{z_i}.$$

We denote the $L_2(Q_n)$-norm of a function $\theta : \mathcal{Z} \to \mathbf{R}$ as

$$\|\theta\|_{Q_n} = (\int \theta^2 dQ_n)^{1/2}.$$

Now, let $\psi_1, \ldots, \psi_n$ be an orthonormal basis in $L_2(Q_n)$. Each function θ can be written as

$$\theta = \sum_{j=1}^n \alpha_j \psi_j,$$

in $L_2(Q_n)$. Moreover,

$$\|\theta\|_{Q_n}^2 = \sum_{j=1}^n \alpha_j^2 = \|\alpha\|_n^2,$$

where $\alpha = (\alpha_1 \ldots, \alpha_n)'$, and where we denote the Euclidean norm of a vector in $\mathbf{R}^n$ by $\| \cdot \|_n$. For $\theta = \sum_{j=1}^n \alpha_j \psi_j$, the *soft thresholding type penalty* is

$$I_n(\theta) = \sum_{j=1}^n |\alpha_j|. \tag{1.3}$$

As will be explained in Section 2, in the least squares (LS) case ($\gamma(\xi) = \xi^2$), the estimator in (1.2) is in fact the standard soft thresholding estimator (see e.g. DONOHO and JOHNSTONE (1994b) and DONOHO (1995)). The least squares estimator (LSE) with soft thresholding is well studied, and very convenient from a computational point of view. It is however of interest to investigate other loss functions as well, because the theory for the least squares case depends on the assumption of existence of second moments of the errors. Moreover, the LS method requires a smoothing parameter λ_n^2 which depends on the scale on which the observations are measured. It depends on (an estimator of) the variance of the errors. Robust regression estimators can do with a choice for λ_n^2 which works for all data. This is in particular true for the least absolute deviations (LAD) estimator ($\gamma(\xi) = |\xi|$). Moreover, LAD estimation with

soft thresholding type penalty is computationally quite feasible. The minimization problem (1.2) can be solved using a standard L_1-fitting routine (see Section 6), which in turn may be based on a simplex algorithm or accelerated version thereof.

There exists a large amount of literature on penalized M-estimation. We will not give a complete overview here, but only mention some of the relevant references. SILVERMAN (1982), STONE (1990), and BARRON and SHEU (1991), study the asymptotic theory for the more classical (spline-)smoothing techniques by penalization (in a density estimation problem). VAN DE GEER (1999, 2000) studies penalized M-estimation using the entropy of parameter space. BIRGÉ and MASSART (1997, 2001), BARRON, BIRGÉ and MASSART (1999), YANG (1999) and BARAUD (2000), present general results on model selection via penalization. The literature on wavelets contains many results on estimation via penalization, see e.g. DONOHO, JOHNSTONE, KERKYACHARIAN and PICARD (1996a,b). Soft thresholding in connection with LS estimation is also studied in TIBSHIRANI (1996). The soft thresholding type penalty we consider can be viewed as an special type of L_1-penalty. Such penalties are also in MAMMEN and VAN DE GEER (1997) in connection with LS estimation, and in PORTNOY (1997) in connection with more general loss functions.

We shall study the large sample behavior ($n \geq 2$ large) of estimators of θ_0. Throughout, as n varies, θ_0 as well as Θ are allowed to vary as well. However, to avoid too many indices, we will not always express dependence on n in our notation.

The paper is organized as follows. In the next section, we re-establish a rate of convergence for the LSE with soft thresholding. We use a method of proof that does not need an explicit expression for the estimator, so that it is well tailored for transfer to other estimation methods.

In Section 3, we present the extension to robust regression. Here, we need an inequality derived from empirical process theory.

To illustrate the regression theory, and yet minimize the amount of approximation theory, we introduce in Section 4 a space of functions governed by a *roughness* parameter ρ. It is assumed there that the true regression function $\theta_0 = \sum_{j=1}^{n} \alpha_{j,0} \psi_j$ has roughness at most ρ, in the sense that $\sum_{j=1}^{n} |\alpha_{j,0}|^\rho \leq M$. If the parameter ρ ($0 \leq \rho \leq 2$) (as well as the parameter M) is known, one may consider estimators without penalty. Rates of convergence for such estimators follow e.g. from entropy calculations. The regression estimators with soft thresholding type penalty do not require knowledge of ρ, and turn out to be rate-adaptive in ρ. We also discuss the relation with Besov spaces. In Section 5, we consider the special case of least absolute deviations (LAD). Section 6 presents a simulation study, where LS is compared to LAD, when the errors are Laplacian (i.e., double exponential) or Gaussian.

2. LEAST SQUARES ESTIMATION USING SOFT THRESHOLDING

In this section, we investigate the classical regression model (1.1), with independent errors $W_1, \ldots, W_n$ with zero expectation and finite variance σ^2. Moreover, we let $\hat{\theta}_n$ denote the LSE with soft thresholding, i.e.,

$$\hat{\theta}_n = \arg \min_{\theta = \alpha_1 \psi_1 + \ldots + \alpha_n \psi_n} \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta(z_i))^2 + 2\lambda_n^2 \sum_{j=1}^{n} |\alpha_j| \right] \qquad (2.1)$$

$$= \sum_{j=1}^{n} \hat{\alpha}_{j,n} \psi_j.$$

(We have added a factor 2 to the penalty term to simplify the expressions.) The explicit expression, which explains why the method is called soft thresholding, is the following. Let α and a be two vectors in $\mathbf{R}^n$, and consider the loss functions

$$\tau_{\text{soft}}^2(\alpha|a) = \left(\sum_{j=1}^{n} |\alpha_j - a_j|^2 + 2\lambda_n^2 \sum_{j=1}^{n} |\alpha_j| \right),$$

and

$$\tau_{\text{hard}}^2(\alpha|a) = \left(\sum_{j=1}^{n} |\alpha_j - a_j|^2 + \lambda_n^4 (\#\{\alpha_j \ : \ |\alpha_j| > 0\}) \right).$$

The solution of the minimization problem

$$(a)_{\text{soft}} = \arg\min_{\alpha} \tau_{\text{soft}}^2(\alpha|a)$$

is

$$(a_j)_{\text{soft}} = \begin{cases} a_j - \lambda_n^2, & \text{if } \tilde{a}_j > \lambda_n^2 \\ 0, & \text{if } |a_j| \leq \lambda_n^2 \ , \ j = 1, \dots, n. \\ a_j + \lambda_n^2, & \text{if } a_j < -\lambda_n^2 \end{cases}$$

Furthermore, the solution of the minimization problem

$$(a)_{\text{hard}} = \arg\min_{\alpha} \tau_{\text{hard}}^2(\alpha|a)$$

is

$$(a_j)_{\text{hard}} = \begin{cases} a_j, & \text{if } |a_j| > \lambda_n^2 \\ 0, & \text{if } |a_j| \leq \lambda_n^2 \end{cases}, \ j = 1, \dots, n.$$

Now, let us write the empirical coefficients as

$$\tilde{\alpha}_{j,n} = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_j(z_i).$$

Then the coefficients of the soft thresholding estimator are $(\tilde{\alpha}_n)_{\text{soft}}$, and those of the hard thresholding estimator are $(\tilde{\alpha}_n)_{\text{hard}}$. Thus, the penalized least squares estimator (2.1) is the soft thresholding estimator.

Consider now "true" coefficients α_0. Let

$$\alpha_* = (\alpha_0)_{\text{hard}}.$$

Then we have

$$\tau_{\text{hard}}^2(\alpha_*|\alpha_0) = \|\alpha_* - \alpha_0\|_n^2 + \lambda_n^4 N_n, \tag{2.2}$$

where $N_n = \#\{|\alpha_{j,0}| > \lambda_n^2\}$. When $\lambda_n^4 = \sigma^2/n$, this is the bias-variance decomposition for the projection estimator when the optimal choice of the subspace formed by a subset of $\psi_1, \dots, \psi_n$ is known.

We will now establish an upper bound of the form (2.2) for $\|\hat{\alpha}_n - \alpha_0\|_n$. We will not use the explicit expression $\hat{\alpha}_n = (\tilde{\alpha}_n)_{\text{soft}}$. This is of importance, because it will allow the extension to other estimation procedures, where no explicit expressions are available.

In the theorem, we put

$$V_j = \frac{1}{n} \sum_{i=1}^n \psi_j(z_i) W_i, \ \ j = 1, \ldots, n.$$

Theorem 2.1. *Let B_n be the set*

$$B_n = \{ \max_{j=1,\ldots,n} |V_j| \leq \lambda_n^2 \}.$$

Then on B_n,

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \leq 4(\|\alpha_* - \alpha_0\|_n^2 + 4\lambda_n^4 N_n). \tag{2.3}$$

Proof. Write

$$\mathcal{J}_n = \{ j : \ |\alpha_{j,0}| > \lambda_n^2 \},$$

and

$$I_N(\alpha) = \sum_{j \in \mathcal{J}_n} |\alpha_j|, \ \ I_M(\alpha) = \sum_{j \notin \mathcal{J}_n} |\alpha_j|,$$

and (identifying a function θ with its coefficients α),

$$I_n(\alpha) = I_N(\alpha) + I_M(\alpha) = \sum_{j=1}^n |\alpha_j|.$$

Clearly, by the definition of the estimator $\hat{\alpha}_n$,

$$\tau_{\text{soft}}^2(\hat{\alpha}_n | \tilde{\alpha}_n) \leq \tau_{\text{soft}}^2(\alpha_* | \tilde{\alpha}_n).$$

Rewrite this to

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n) \leq 2 \sum_{j=1}^n V_j(\hat{\alpha}_{j,n} - \alpha_{j,*}) + 2\lambda_n^2 I_n(\alpha_*) + \|\alpha_* - \alpha_0\|_n^2.$$

This gives

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n) \leq 2 \max_{j=1,\ldots,n} |V_j| I_n(\hat{\alpha}_n - \alpha_*) + 2\lambda_n^2 I_n(\alpha_*) + \|\alpha_* - \alpha_0\|_n^2.$$

On B_n, we find

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n)$$
$$\leq 2\lambda_n^2 I_n(\hat{\alpha}_n - \alpha_*) + 2\lambda_n^2 I_n(\alpha_*) + \|\alpha_* - \alpha_0\|_n^2 \tag{2.4}$$

or, since $\alpha_{j,*} = 0$ for $j \notin \mathcal{J}_n$,

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_N(\hat{\alpha}_n) + 2\lambda_n^2 I_M(\hat{\alpha}_n) \leq 2\lambda_n^2 I_N(\hat{\alpha}_n - \alpha_*) + 2\lambda_n^2 I_M(\hat{\alpha}_n)$$

$$+ 2\lambda_n^2 I_N(\alpha_*) + \|\alpha_* - \alpha_0\|_n^2.$$

But then

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \le 2\lambda_n^2 I_N(\hat{\alpha}_n - \alpha_*) + 2\lambda_n^2(I_N(\alpha_*) - I_N(\hat{\alpha}_n)) + \|\alpha_* - \alpha_0\|_n^2$$
$$\le 4\lambda_n^2 I_N(\hat{\alpha}_n - \alpha_*) + \|\alpha_* - \alpha_0\|_n^2$$
$$\le 4\lambda_n^2 \sqrt{N_n} (\sum_{j \in \mathcal{J}_n} |\hat{\alpha}_{j,n} - \alpha_{j,*}|^2)^{1/2} + \|\alpha_* - \alpha_0\|_n^2,$$

where in the last inequality we applied Cauchy-Schwarz. Now, use that

$$\sum_{j \in \mathcal{J}_n} |\hat{\alpha}_{j,n} - \alpha_{j,*}|^2 = \sum_{j \in \mathcal{J}_n} |\hat{\alpha}_{j,n} - \alpha_{j,0}|^2 \le \|\hat{\alpha}_n - \alpha_0\|_n^2. \tag{2.5}$$

We then arrive at

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \le 4\lambda_n^2 \sqrt{N_n} \|\hat{\alpha}_n - \alpha_0\|_n + \|\alpha_* - \alpha_0\|_n^2.$$

But this implies (2.3). □

Corollary 2.2. *Suppose that for some constant $K < \infty$, the errors $W_1, \ldots, W_n$ satisfy*

$$\max_{i=1,\ldots,n} \mathbf{E} \exp[W_i^2/K^2] \le K.$$

Then it follows from e.g. VAN DE GEER *(2000, Lemma 8.2), that for a constant c depending on K,*

$$\mathbf{P} \left(\max_{j=1,\ldots,n} |V_j| > c\sqrt{\frac{\log n}{n}} \right) \le c \exp[-\frac{\log n}{c^2}].$$

Thus, then we may take $\lambda_n^2 = c\sqrt{\log n/n}$, and obtain

$$\mathbf{P}(\|\hat{\alpha}_n - \alpha_0\|_n^2 > 4(\|\alpha_* - \alpha_0\|_n^2 + 4\lambda_n^4 N_n)) \le c \exp[-\frac{\log n}{c^2}].$$

In general, it is clear that the choice of λ_n^2 depends on the distribution of the errors. As a consequence, if the errors have heavy tails, the rate of convergence of the LSE with soft thresholding may be very slow. Therefore, robust methods may provide a welcome alternative to the LS method.

3. ROBUST ADAPTIVE ESTIMATORS

The estimator with soft thresholding type penalty, based on the convex loss function γ, is defined as

$$\hat{\theta}_n = \min_{\theta \in \Theta, \ \theta = \sum_{j=1}^n \alpha_j \psi_j} \left[\frac{1}{n} \sum_{i=1}^n \gamma(Y_i - \theta(z_i)) + \lambda_n^2 \sum_{j=1}^n |\alpha_j| \right]$$
$$= \sum_{j=1}^n \hat{\alpha}_{j,n} \psi_j. \tag{3.1}$$

The case $\gamma(\xi) = \xi^2$ was studied in the previous section. In this section, we examine the robust case, where γ satisfies

$$|\gamma(\xi) - \gamma(\tilde{\xi})| \le |\xi - \tilde{\xi}|, \ \xi, \tilde{\xi} \in \mathbf{R}. \tag{3.2}$$

The following notation is convenient. Let for $i = 1, \ldots, n$, $X_i = (Y_i, z_i)$, let $P^{(i)}$ be the distribution of X_i and $\gamma_\theta(X_i) = \gamma(Y_i - \theta(z_i))$. Let $\bar{P} = \sum_{i=1}^n P^{(i)}/n$ and denote the empirical distribution based on $X_1, \ldots, X_n$ by P_n.

We assume that $\mathbf{E}\gamma(Y_i - b)$ has a unique minimum in $b = \theta_0(z_i)$. This implies

$$\theta_0 = \arg\min_\theta \int \gamma_\theta d\bar{P}.$$

We assume moreover that $\theta_0 \in \Theta$, where Θ is a given class of regression functions.

Now, $0 < \epsilon \leq 1$ and let $\theta_* = \sum_{j=1}^n \alpha_{j,*}\psi_j \in \Theta$ satisfy

$$\int (\gamma_{\theta_*} - \gamma_{\theta_0})d\bar{P} \leq \|\theta_* - \theta_0\|_{Q_n}^2/\epsilon. \tag{3.3}$$

In view of Section 2, a natural choice for θ_* would be the counterpart in the robust setting of a hard thresholding version of θ_0, that is

$$\theta_* = \arg\min_{\theta \in \Theta, \ \theta = \sum_{j=1}^n \alpha_j\psi_j} \left\{ \int \gamma_\theta d\bar{P} + \lambda_n^4(\#\{\alpha_j \ : \ |\alpha_j| > 0\}) \right\}.$$

However, we do not insist on this because (3.3) may not be true for this choice. It should be noted that we may also choose $\theta_* = \theta_0$, in which case (3.3) is automatically true.

We also need the following condition.
For $0 \leq t \leq 1$, write $\theta_t = t\theta + (1-t)\theta_*$. (We do not require $\theta_t \in \Theta$.) Assume that for $t_0 \leq 1/(2K)$ sufficiently small and $\theta \in \Theta$, we have for all $0 \leq t \leq t_0$,

$$\int (\gamma_{\theta_t} - \gamma_{\theta_0})d\bar{P} \geq \epsilon\|\theta_t - \theta_0\|_{Q_n}^2. \tag{3.4}$$

Since $\int \gamma_\theta d\bar{P}$ is minimized at θ_0, condition (3.4) will be a reasonable condition under the following condition (3.5). We denote the supremum norm by

$$\|\theta\|_\infty = \sup_{z \in \mathcal{Z}} |\theta(z)|.$$

Suppose that for some constant K,

$$\sup_{\theta \in \Theta} \|\theta\|_\infty \leq K. \tag{3.5}$$

Condition (3.5) is of course an awkward condition. The choice for the smoothing parameter in Theorem 3.1 is based on the empirical process inequality of Lemma 3.4. It depends on a universal constant c.

Theorem 3.1. *Suppose that (3.3) and (3.4) hold. Let $\mathcal{J}_n$ be any subset of $\{1, \ldots, n\}$, and define*

$$N_n = |\mathcal{J}_n|, \ M_n = \sum_{j \notin \mathcal{J}_n} |\alpha_{j,*}|.$$

Suppose that $\|\theta_* - \theta_0\|_{Q_n}^2 + \lambda_n^4 N_n + \lambda_n^2 M_n \leq \eta^2$, where $\eta = \epsilon t_0/32$. Then for $\lambda_n^2 \geq c\sqrt{\log n/n}$, we have

$$\mathbf{P}\left(\|\hat{\theta}_n - \theta_0\|_{Q_n}^2 \geq \frac{1}{\eta^2}(\|\theta_* - \theta_0\|_{Q_n}^2 + \lambda_n^4 N_n + \lambda_n^2 M_n + \frac{1}{n}) \right)$$

$$\leq c\exp[-\frac{\log n}{c^2}].$$

Proof. The proof follows the line of reasoning of Theorem 2.1 on the LSE.

Define $t = t_0/(1 + \|\hat{\theta}_n - \theta_*\|_{Q_n})$. Consider the convex combination

$$\hat{\theta}_{t,n} = t\hat{\theta}_n + (1-t)\theta_*.$$

Using the convexity of the loss function γ, and of the soft thresholding type penalty, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\gamma(Y_i - \hat{\theta}_{t,n}(z_i)) + \lambda_n^2 I_n(\hat{\theta}_{t,n})$$

$$\leq t\left\{\frac{1}{n}\sum_{i=1}^{n}\gamma(Y_i - \hat{\theta}_n(z_i)) + \lambda_n^2 I_n(\hat{\theta}_n)\right\}$$

$$+(1-t)\left\{\frac{1}{n}\sum_{i=1}^{n}\gamma(Y_i - \theta_*(z_i)) + \lambda_n^2 I_n(\theta_*)\right\}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\gamma(Y_i - \theta_*(z_i)) + \lambda_n^2 I_n(\theta_*),$$

where in the second inequality, we used that $\hat{\theta}_n$ minimizes the penalized loss function over Θ and that $\theta_* \in \Theta$. We rewrite this in a convenient form, namely

$$\int(\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_0})d\bar{P} + \lambda_n^2 I_n(\hat{\theta}_{t,n}) \leq -\int(\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_*})d(P_n - \bar{P}) + \lambda_n^2 I_n(\theta_*)$$

$$+\int(\gamma_{\theta_*} - \gamma_{\theta_0})d\bar{P}.$$

By assumption (3.3),

$$\int(\gamma_{\theta_*} - \gamma_{\theta_0})d\bar{P} \leq \|\theta_* - \theta_0\|_{Q_n}^2/\epsilon.$$

Moreover, by assumption (3.4)

$$\int(\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_0})d\bar{P} \geq \epsilon\|\hat{\theta}_{t,n} - \theta_0\|_{Q_n}^2.$$

So we find

$$\epsilon\|\hat{\theta}_n - \theta_0\|_{Q_n}^2 + \lambda_n^2 I_n(\hat{\theta}_{t,n}) \leq -\int(\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_*})d(P_n - \bar{P}) + \lambda_n^2 I_n(\theta_*)$$

$$+\|\theta_* - \theta_0\|_{Q_n}^2/\epsilon.$$

We will now apply Lemma 3.4. Fot this purpose, we remark that if $I_n(\hat{\theta}_{t,n} - \theta_*) \leq n^{-\frac{1}{2}}$, it follows immediately that also
$\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n} \leq n^{-\frac{1}{2}}$. One can show that then also $\|\hat{\theta}_n - \theta_*\|_{Q_n} \leq \frac{1}{\eta}n^{-\frac{1}{2}}$ (this follows from the same arguments as those used at the end of this proof). We may therefore assume that $I_n(\hat{\theta}_{t,n} - \theta^*) > n^{-\frac{1}{2}}$. Observe also that $\|\hat{\theta}_{t,n} - \theta_*\|_\infty \leq 1$.

Let B_n be the set

$$|\int (\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_*}) d(P_n - \bar{P})| \leq c\sqrt{\frac{\log n}{n}} I_n(\hat{\theta}_{t,n} - \theta_*).$$

We show in Lemma 3.4 that

$$\mathbf{P}(B_n) \geq 1 - c\exp[-\frac{\log n}{c^2}].$$

On B_n, we find

$$\epsilon\|\hat{\theta}_{t,n} - \theta_0\|_{Q_n}^2 + \lambda_n^2 I_n(\hat{\theta}_{t,n})$$

$$\leq \lambda_n^2 I_n(\hat{\theta}_{t,n} - \theta_*) + \lambda_n^2 I_n(\theta_*) + \|\theta_* - \theta_0\|_{Q_n}^2/\epsilon. \tag{3.5}$$

This inequality is similar to (2.4) in Theorem 2.1, and we may proceed as there. Then

$$\epsilon\|\hat{\theta}_{t,n} - \theta_0\|_{Q_n}^2 \leq 2\lambda_n^2\sqrt{N_n}\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n} + \|\theta_* - \theta_0\|_{Q_n}^2/\epsilon + 2\lambda_n^2 M_n,$$

where we now used that for $\hat{\theta}_{t,n} = \sum_{j=1}^n \hat{\alpha}_{t,j,n}\psi_j$,

$$\sum_{j\in\mathcal{J}_n} (\hat{\alpha}_{t,j,n} - \alpha_{j,*})^2 \leq \|\hat{\theta}_{t,n} - \theta_*\|_{Q_n}^2,$$

instead of inequality (2.5). The additional term $2\lambda_n^2 M_n$ is due to the fact that $\alpha_{j,*}$ may not be zero for $j \notin \mathcal{J}_n$.

Invoking $\|\hat{\theta}_{t,n} - \theta_0\|_{Q_n}^2 \geq \frac{1}{2}\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n}^2 - \|\theta_* - \theta_0\|_{Q_n}^2$, and $\epsilon \leq 1$, we find

$$\frac{\epsilon}{2}\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n}^2 \leq 2\lambda_n^2\sqrt{N_n}\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n} + \frac{2}{\epsilon}\|\theta_* - \theta_0\|_{Q_n}^2 + 2\lambda_n^2 M_n.$$

But then

$$\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n} \leq \frac{8}{\epsilon}\max(\|\theta_* - \theta_0\|_{Q_n}, \lambda_n^2\sqrt{N_n}, \lambda_n\sqrt{M_n}). \tag{3.6}$$

Here, we used twice the inequality $a + b \leq 2\max(a, b)$ for positive a and b.

The right hand side of (3.6) is less than $8\eta/\epsilon = t_0/4 \leq t_0/2$. Since

$$\|\hat{\theta}_{t,n} - \theta_*\|_{Q_n} = t_0\frac{\|\hat{\theta}_n - \theta_*\|_{Q_n}}{1 + \|\hat{\theta}_n - \theta_*\|_{Q_n}},$$

we find

$$\frac{\|\hat{\theta}_n - \theta_*\|_{Q_n}}{1 + \|\hat{\theta}_n - \theta_*\|_{Q_n}} \leq \frac{1}{2},$$

so $\|\hat{\theta}_n - \theta_*\|_{Q_n} \leq 1$, and hence $t = t_0/(1 + \|\hat{\theta}_n - \theta_*\|_{Q_n}) \geq t_0/2$. Thus (3.6) gives

$$\|\hat{\theta}_n - \theta_*\|_{Q_n} = \|\hat{\theta}_{t,n} - \theta_*\|_{Q_n}/t \leq \frac{16}{\epsilon t_0}\max(\|\theta_* - \theta_0\|_{Q_n}, \lambda_n^2\sqrt{N_n}, \lambda_n\sqrt{M_n}).$$

Hence

$$\|\hat{\theta}_n - \theta_0\|_{Q_n}^2 \leq 2\|\hat{\theta}_n - \theta_*\|_{Q_n}^2 + 2\|\theta_* - \theta_0\|_{Q_n}^2$$
$$\leq 2(\frac{16}{\epsilon t_0})^2 \max(\|\theta_* - \theta_0\|_{Q_n}^2, \lambda_n^4 N_n, \lambda_n^2 M_n) + 2\|\theta_* - \theta_0\|_{Q_n}^2$$
$$\leq \frac{1}{\eta^2}(\|\theta_* - \theta_0\|_{Q_n}^2 + \lambda_n^4 N_n + \lambda_n^2 M_n).$$

$\square$

Corollary 3.2. *Let*

$$\theta_* = \arg \min_{\theta \in \Theta, \ \theta = \sum_{j=1}^n \alpha_j \psi_j} \left\{ \int \gamma_\theta d\bar{P} + \lambda_n^4(\#\{\alpha_j \ : \ |\alpha_j| > 0\}) \right\}$$
$$= \sum_{j=1}^n \alpha_{j,*} \psi_j.$$

Take $\mathcal{J}_n = \{j : |\alpha_{j,*}| > 0\}$, *to find from Theorem 3.1 that under the conditions stated there*

$$\mathbf{P}\left(\|\hat{\theta}_n - \theta_0\|_{Q_n}^2 \geq \frac{1}{\eta^2}(\|\theta_* - \theta_0\|_{Q_n}^2 + \lambda_n^4 N_n + \frac{1}{n}) \right) \leq c \exp[-\frac{\log n}{c^2}].$$

Theorem 3.1 and its corollary show that for the robust penalized regression estimator, one has a similar result as for the penalized LSE. Moreover, the robustness condition (3.2) implies that there exists a universal value for the smoothing parameter, that works well for all error distributions. The optimal value for the smoothing parameter is as yet not clear. The empirical process inequality of Lemma 3.4 gives an upper bound $\lambda_n^2 \geq c\sqrt{\log n / n}$.

The remainder of this section is on the empirical process inequality of Lemma 3.4. Before stating this lemma, we need the following auxiliary result, where we use the notation $a \vee b = \max(a, b)$ $(a \wedge b = \min(a, b))$.

Lemma 3.3. *For all $M > 0$, the following upper bound holds*

$$\mathbf{P}\left(\sup_{I_n(\theta) \leq M} | \int (\gamma_\theta - \gamma_0) d(P_n - \bar{P})| \geq 16M\sqrt{\frac{\log n}{n}} \right)$$
$$\leq \exp[-\frac{(M^2 \vee 1)\log n}{2}].$$

Proof. Define the following empirical process

$$Z = \sup_{I_n(\theta) \leq M} \left| \int (\gamma_\theta - \gamma_0) \right| d(P_n - \bar{P}).$$

Set

$$U_i(\theta) = \gamma(Y_i - \theta(z_i)) - \gamma(Y_i), \ i = 1, \dots, n.$$

An Hoeffding type inequality, proved by MASSART (2000), says that for all $u \geq 0$ we have

$$\mathbf{P}\left(Z \geq \mathbf{E}(Z) + u \right) \leq \exp\left[-\frac{n^2 u^2}{8b_n^2} \right]$$

where b_n satisfies

$$\sup_{I_n(\theta) \leq M} \sum_{i=1}^{n} \|U_i(\theta) - \mathbf{E}(U_i(\theta))\|_{\infty}^2 \leq b_n^2.$$

Here, $\|U\|_{\infty}$ denotes the sup-norm of the random variable U.

Since γ is $1-$Lipschitz (condition (3.2)), and $\|\theta\|_{Q_n} \leq I_n(\theta)$, we have that

$$\sup_{I_n(\theta) \leq M} \sum_{i=1}^{n} \|U_i(\theta) - \mathbf{E}(U_i(\theta))\|_{\infty}^2$$

$$\leq 4n \sup_{I_n(\theta) \leq M} \|\theta\|_{Q_n}^2 \leq 4n(M^2 \wedge 1).$$

As a result, we find the upper bound $b_n^2 \leq 4n(M^2 \wedge 1)$. A symmetrization procedure (see e.g., LEDOUX and TALAGRAND (1991) or VAN DER VAART and WELLNER (1996)) implies the following bound:

$$\mathbf{E}(Z) = \mathbf{E}\left[\sup_{I_n(\theta) \leq M} \frac{1}{n} \left| \sum_{i=1}^{n} [U_i(\theta) - \mathbf{E}(U_i(\theta))] \right| \right]$$

$$\leq 2\mathbf{E}\left[\sup_{I_n(\theta) \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i U_i(\theta) \right| \right]$$

where the ϵ_i's are Rademacher random variables. Using the fact that γ is 1-Lipschitz, the contraction principle in LEDOUX and TALAGRAND (1991, Theorem 4.12) gives the following bound:

$$\mathbf{E}\left[\sup_{I_n(\theta) \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i U_i(\theta) \right| \right]$$

$$\leq 2\mathbf{E}\left[\sup_{I_n(\theta) \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \theta(z_i)) \right| \right]$$

$$= 2\mathbf{E}\left[\sup_{I_n(\theta) \leq M} \left| \frac{1}{n} \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{n} \epsilon_i \psi_j(z_i) \right| \right]$$

$$\leq 2M\mathbf{E}\left[\max_{j=1,\ldots,n} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi_j(z_i) \right| \right].$$

By Hoeffding's inequality (Hoeffding (1963)) and using results in VAN DER VAART and WELLNER (1996, Chapter 2.2), we find the bound $6M\sqrt{\frac{\log n}{n}}$ for last quantity. As a consequence we get

$$\mathbf{P}[Z \leq 12M\sqrt{\frac{\log n}{n}} + u] \geq 1 - \exp\left(-\frac{nu^2}{32(M^2 \wedge 1)}\right).$$

Taking $u = 4\sqrt{\frac{\log n}{n}}M$ completes the proof of the lemma. $\qquad \square$

Lemma 3.3 is used in the next lemma, which is in turn a basic ingredient of Theorem 3.1.

Lemma 3.4. *There exists a constant c such that*

$$\mathbf{P}\left(\sup_{\|\theta-\theta_*\|_\infty\leq 1,\ I_n(\theta-\theta_*)>n^{-\frac{1}{2}}}\frac{|\int(\gamma_\theta-\gamma_{\theta_*})d(P_n-\bar{P})|}{I_n(\theta-\theta_*)}\geq c\sqrt{\frac{\log n}{n}}\right)$$

$$\leq c\exp[-\frac{\log n}{c^2}].$$

Proof. Without loss of generality, we may assume $\theta_*\equiv 0$. We show in the previous lemma that we have the concentration inequality

$$\mathbf{P}\left(\sup_{I_n(\theta)\leq M}|\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})|\geq 16M\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \exp[-\frac{(M^2\vee 1)\log n}{2}].$$

Take j_0 as the smallest integer such that $j_0+1>\log_2\sqrt{n}$. Then we find

$$\mathbf{P}\left(\sup_{n^{-\frac{1}{2}}<I_n(\theta)\leq 1}\frac{|\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})|}{I_n(\theta)}\geq 32\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \sum_{j=0}^{j_0}\mathbf{P}\left(\sup_{I_n(\theta)\leq 2^{-j}}|\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})|\geq 16\sqrt{\frac{\log n}{n}}2^{-j}\right)$$

$$\leq (\log_2\sqrt{n}+1)\exp[-\frac{\log n}{2}]$$

$$\leq C_0\exp[-\frac{\log n}{c^2}].$$

Moreover,

$$\mathbf{P}\left(\sup_{I_n(\theta)>1}\frac{|\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})|}{I_n(\theta)}\geq 32\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \sum_{j=0}^{\infty}\mathbf{P}\left(\sup_{I_n(\theta)\leq 2^{j+1}}|\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})|\geq 16\sqrt{\frac{\log n}{n}}2^{j+1}\right)$$

$$\leq \sum_{j=0}^{\infty}\exp[-\frac{(\log n)2^{2(j+1)}}{2}]$$

$$\leq C_1\exp[-\frac{\log n}{c^2}].$$

$\square$

4. An illustration with roughness parameter ρ

In this section, we illustrate the consequences of Theorem 2.1 and Theorem 3.1 for a special case. Consider the set of functions

$$\Theta_\rho = \{\theta = \sum_{j=1}^n \alpha_j \psi_j, \ \sum_{j=1}^n |\alpha_j|^\rho \leq 1\}, \tag{4.1}$$

where $0 \leq \rho \leq 2$. Here, for $\rho = 0$, we use the convention $x^0 = 1$ if x is non zero and $0^0 = 0$. Since the sets Θ_ρ are increasing in size as ρ increases, we may think of ρ as a *roughness* parameter. Thus, the smaller ρ the "smoother" the functions in Θ_ρ will be. At the extremes, $\rho = 0$ implies that there is at most 1 non-zero coefficient, whereas $\rho = 2$ only requires that θ is within the n-dimensional unit ball.

We will first consider in Subsection 4.1, the case where it is known that for a given ρ, $\theta_0 \in \Theta_\rho$. In that situation, one may consider an estimation method without penalty. The rate of convergence can then be derived from the entropy of Θ_ρ. Subsection 4.2 computes this entropy. In Subsection 4.3, we show that the estimators with soft thresholding type penalty (which do not require knowledge of ρ) converge with the same rate as the one found in Subsection 4.1. Subsection 4.4 discusses the relation with Besov spaces.

4.1. **The case ρ known.**

Lemma 4.1. *Suppose $\Theta \subset \Theta_\rho$, where $0 < \rho < 1$. Let*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \gamma(Y_i - \theta(z_i))$$

be the regression estimator without penalty. When $\gamma(\xi) = \xi^2$ (the least squares case), assume that for some $K < \infty$,

$$\max_{i=1,\ldots,n} \mathbf{E}\exp[W_i^2/K^2] \leq K,$$

and when the Lipschitz condition (3.2) holds (the robust case), assume (3.4) for $\theta_* = \theta_0$. Then there exist a constant C, depending on K in the LS case, and on ρ, such that for all $T \geq C$,

$$\mathbf{P}(\|\hat{\theta}_n - \theta_0\|_{Q_n} \geq T(\frac{\log n}{n})^{\frac{2-\rho}{4}}) \leq C\exp[-\frac{n^{\frac{\rho}{2}}(\log n)^{\frac{2-\rho}{2}}}{C^2}].$$

Proof. This follows from general results in VAN DE GEER (2000, Theorem 9.1 and Theorem 12.3), using the entropy given in Lemma 4.3 below. $\qquad\square$

The rate of convergence $(\log n/n)^{\frac{2-\rho}{4}}$ corresponds to the minimax rate over Θ_ρ, when the errors are i.i.d. Gaussian random variables (see DONOHO and JOHNSTONE (1994a)).

The results of VAN DE GEER (2000, Theorem 9.1 and 12.3) can be also used for the case $1 \leq \rho < 2$, and the resulting rates will then be sub-optimal (i.e. slower than the minimax rate over Θ_ρ).

4.2. **The entropy of Θ_ρ.**

Definition 4.2. *Let T be a (subset) of a metric space. The δ-covering number $N(\delta, T)$ is the minimal number of balls with radius $\delta > 0$ necessary to cover T. The δ-entropy is $H(\delta, T) = \log N(\delta, T)$.*

In our situation, we need entropies of subsets of $L_2(Q_n)$. Note that $L_2(Q_n)$ is essentially the n-dimensional Euclidean space $\mathbf{R}^n$.

Of course, the smaller ρ, the smaller the entropy. Bounds are given in the next lemma (where we omit the two extreme but trivial cases $\rho = 0$ and $\rho = 2$).

Lemma 4.3. *Consider the following subset of $\mathbf{R}^n$:*
$\mathcal{A}_n = \{\alpha = (\alpha_1, \ldots, \alpha_n)' : \sum_{j=1}^n |\alpha_j|^\rho \le 1\}$, *with $0 < \rho < 2$. We have for some constant A, depending only on ρ,*

$$H(\delta, \mathcal{A}_n) \le A\delta^{-\frac{2\rho}{2-\rho}} \left(\log n + \log \frac{1}{\delta} \right), \quad \delta > 0. \tag{4.2}$$

Proof. Let $\epsilon = \delta^{\frac{2}{2-\rho}}$. Define for $\alpha \in \mathcal{A}_n$,

$$N_\alpha(\epsilon) = \#\{\alpha_j : |\alpha_j| > \epsilon\}.$$

Moreover, let

$$N(\epsilon) = \lfloor \epsilon^{-\rho} \rfloor,$$

where $\lfloor x \rfloor$ denotes the integer part of $x > 0$.

We have

$$\max_{\alpha \in \mathcal{A}_n} N_\alpha(\epsilon) \le N(\epsilon).$$

It suffices to have a δ-approximation of the coefficients larger than ϵ, neglecting the other coefficients. That is, let $\alpha \in \mathcal{A}_n$, and suppose that for some $\bar{\alpha}$,

$$\sum_{|\alpha_j| > \epsilon} |\alpha_j - \bar{\alpha}_j|^2 \le \delta^2.$$

In addition, suppose that $\bar{\alpha}_j = 0$ for all $|\alpha_j| \le \epsilon$. Then we have

$$\|\alpha - \bar{\alpha}\|_n^2 \le \delta^2 + \sum_{|\alpha_j| \le \epsilon} |\alpha_j|^2 \le 2\delta^2,$$

provided

$$\sum_{|\alpha_j| \le \epsilon} |\alpha_j|^2 \le \epsilon^{2-\rho}.$$

But this follows from

$$\sum_{|\alpha_j| \le \epsilon} |\alpha_j|^2 = \sum_{|\alpha_j| \le \epsilon} |\alpha_j|^{2-\rho+\rho}$$
$$\le \sum_{|\alpha_j| \le \epsilon} |\alpha_j|^\rho \epsilon^{2-\rho}$$
$$\le \epsilon^{2-\rho}.$$

The number of ways to choose $N(\epsilon) \leq n$ coefficients out of n is

$$\binom{n}{N(\epsilon)} \leq n^{N(\epsilon)}.$$

Moreover, the δ-entropy of a unit ball in Euclidean space with dimension $N(\epsilon)$ is at most $5N(\epsilon) \log \frac{1}{\delta}$ (see e.g. VAN DE GEER (2000)). So, we arrive at

$$H(\sqrt{2}\delta, \mathcal{A}_n) \leq N(\epsilon) \left(5 \log \frac{1}{\delta} + \log n \right),$$

where $\epsilon = \delta^{\frac{2}{2-\rho}}$, and $N(\epsilon) \leq \epsilon^{-\rho}$. $\qquad\square$

4.3. The case ρ unknown: adaptation.

Lemma 4.4. *Let*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i - \theta(z_i)) + \lambda_n^2 I_n(\theta) \right]$$

where $I_n(\theta)$ is the soft thresholding type penalty (1.3). When $\gamma(\xi) = \xi^2$ (the least squares case), assume that for some $K < \infty$,

$$\max_{i=1,\ldots,n} \mathbf{E} \exp[W_i^2/K^2] \leq K,$$

and when the Lipschitz condition (3.2) holds (the robust case), assume (3.4) for $\theta_* = \theta_0$. Then there exists a constant c, which depends on K in the LS case, and which is universal in the robust case, such that for $\lambda_n = c\sqrt{\log n/n}$, we have

$$\mathbf{P}\left(\|\hat{\theta}_n - \theta_0\|_{Q_n} \geq c(\frac{\log n}{n})^{\frac{2-\rho}{4}} \right) \leq c\exp[-\frac{\log n}{c^2}].$$

Proof. As in the proof of Lemma 4.3 we employ the fact that if

$$\sum_{j=1}^{n} |\alpha_j|^\rho \leq 1,$$

then

$$N_\alpha(\epsilon) = \#\{\alpha_j : |\alpha_j| > \epsilon\} \leq \epsilon^{-\rho}.$$

It is also easy to see that in that case, for $\rho \leq 1$,

$$M_\alpha(\epsilon) = \sum_{|\alpha_j| \leq \epsilon} |\alpha_j| = \sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^{1-\rho+\rho} \leq \epsilon^{1-\rho}.$$

Thus, in Theorem 2.1, $N_n \leq \lambda_n^{-2\rho}$, which gives by Corollary 2.2 the result for the LS case. Moreover, taking in Theorem 3.1, $\mathcal{J}_n = \{j : |\alpha_{j,0}| > \lambda_n^2\}$, there gives $N_n \leq \lambda_n^{-2\rho}$ and $M_n \leq \lambda_n^{2(1-\rho)}$. This yields the result for the robust case. $\qquad\square$

4.4. **Relation with Besov spaces.** In the literature on adaptive estimation, one often considers so-called Besov spaces $B_{\sigma,p,q}([0,1]^d)$. Such spaces are intrinsically connected to the analysis of curves since the scale of Besov spaces yields the opportunity to describe the regularity of functions, with more accuracy than the classical Hölder scale. General references about Besov spaces are BERGH and LÖFSTRÖM (1976), BESOV, IL'IN and NIKOL'SKII (1978), EDMUND and TRIEBEL (1992) and DEVORE and LORENTZ (1993). This subsection discusses the link with our roughness parameter ρ. The notation $B_{\sigma,p,q}([0,1]^d)$ refers to the case of functions on the d-dimensional unit cube, with "smoothness" σ, and where p and q refer to L_p- and L_q-norms with respect to Lebesgue measure. We will not go into the details, but mainly want to show that, apart from logarithmic factors, such Besov spaces correspond to a roughness parameter ρ equal to $\rho = 2/(2s+1)$, where s is the "effective" smoothness σ/d, and where we assume $\rho \leq p$ (see Lemma 4.5). Similar observations can be found in DONOHO and JOHNSTONE (1996). The application of Lemma 4.3 then yields a bound for the entropy. However, in Besov spaces, the coefficients at higher *levels* tend to be smaller, i.e., there is more structure than as can be described by our roughness parameter ρ. As a result, it turns out that Besov spaces have entropies without logarithmic factors (see Lemma 4.6).

Consider a wavelet basis $\psi_{j,k}$ of $L^2(Q_n)$ with regularity r such that $r \geq s$. We recall that a wavelet regularity is expressed through its number of vanishing moments, see e.g. MEYER (1987) or MALLAT (1998). Then a Besov norm is equivalent to an appropriate norm in the sequence space, that is, the space of the wavelet coefficients, see DEVORE and LORENTZ (1993) or DONOHO, JOHNSTONE, KERKYACHARYAN and PICARD (1996a,b) We take a sequence space as a starting point (and consider for simplicity the case corresponding to $d = 1$ ($\sigma = s$) in the Besov interpretation). The coefficients α are now indexed by two integers: $\alpha = \{\alpha_{j,k}\}$, where k runs from 1 to 2^j, and where $j \in \{1, 2, \ldots, J\}$ can be seen as a zoom-level.

Let $\mathcal{B}_{s,p,q}$ be the set of coefficients $\{\alpha_{j,k}\}$ that satisfy

$$\left(\sum_{j=1}^{J} 2^{j((2s+1)\frac{p}{2}-1)\frac{q}{p}} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \leq 1. \tag{4.3}$$

When the $\alpha_{j,k}$ are the coefficients of the appropriate Besov basis, this quantity is equivalent to the Besov semi-norm. Throughout, we assume $s \geq 0$, $p \geq 1$, and $q \geq 1$. In the Besov space interpretation, $\mathcal{B}_{s,p,q}$ (with $J = \infty$) corresponds (in the sense of norm equivalence) to a Besov ball in the space $B_{s,p,q}([0,1])$.

Lemma 4.5. *Suppose that $\alpha = \{\alpha_{j,k}\}$ satisfies (4.3), with $\rho = 2/(2s+1) \leq \min(p,q)$, and $J < \infty$. Then*

$$\sum_{j=1}^{J} \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \leq J^{\frac{q-\rho}{q}}. \tag{4.4}$$

Proof. By Hölder's inequality, for a sequence $a_1, \ldots, a_L$, and for $t \geq 1$,

$$\sum_{l=1}^{L} |a_l| \leq L^{\frac{t-1}{t}} \left(\sum_{l=1}^{L} |a_l|^t \right)^{\frac{1}{t}}. \tag{4.5}$$

Apply this first with $L = J$, $|a_j| = \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho$, and $t = q/\rho$. Then we find

$$\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\} \leq J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}. \tag{4.6}$$

Next, apply (4.5) with $L = 2^j$, $|a_{j,k}| = |\alpha_{j,k}|^\rho$, and $t = p/\rho$. This yields

$$\left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\} \leq \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\alpha_{j,k}|^p)^{\frac{\rho}{p}} \right\}.$$

Do this for each $j = 1, \ldots J$, and insert the result in (4.6):

$$J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$\leq J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\alpha_{j,k}|^p)^{\frac{\rho}{p}} \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$= J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} 2^{j(\frac{p-\rho}{p})\frac{q}{\rho}} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{\rho}{q}}$$

$$\leq J^{\frac{q-\rho}{q}},$$

since

$$(\frac{p-\rho}{p})\frac{q}{\rho} = ((2s+1)\frac{p}{2} - 1)\frac{q}{p}.$$

$\square$

Remark 4.1. One can also define spaces $\mathcal{B}_{s,p,q}$ with $p = \infty$ and/or $q = \infty$. Condition (4.3) is then to be understood with the usual adjustments. Note that $\mathcal{B}_{s,p,q} \subset \mathcal{B}_{s,p,\infty}$.

In our applications, the number of levels J is logarithmic in n. The entropy of the spaces $\mathcal{B}_{s,p,q}$ can now be bounded by combining Lemma 4.3 with Lemma 4.5. However, it turns out that this will result in a bound with an unnecessary $(\log n)$-term. The entropy bound without logarithmic factors can be found in BIRMAN and SOLOMJAK (1963) for the case of Sobolev spaces, and in BIRGÉ and MASSART (2000) for the generalization to Besov spaces.

We consider $\mathcal{B}_{s,p,q}$ as a subset of the Euclidean space $\mathbf{R}^{(2^J-2)}$, with Euclidean norm $\|\cdot\|$ (possibly $J = \infty$, in which case $\mathbf{R}^\infty$ should be understood as $l_2(N)$).

Lemma 4.6. Let $\rho = 2/(2s+1) < p$. For a constant A depending on p and s,

$$H(\delta, \mathcal{B}_{s,p,\infty}) \leq A\delta^{-\frac{1}{s}}, \quad \delta > 0. \tag{4.7}$$

Proof. This is shown in BIRGÉ and MASSART (2000). In fact, they show that the δ-entropy for the $L_{p'}$ (Lebesgue measure)-norm, of a Besov ball with radius 1 in $B_{\sigma,p,\infty}([0,1]^d)$, is bounded by $A\delta^{-\frac{1}{s}}$, provided $s = \frac{\sigma}{d} > \frac{1}{p} - \frac{1}{p'}$. $\qquad\square$

5. LEAST ABSOLUTE DEVIATIONS

The LAD estimator with soft thresholding type penalty is

$$\hat{\theta}_n = \min_{\theta \in \Theta, \ \theta = \sum_{j=1}^n \alpha_j \psi_j} \left[\frac{1}{n} \sum_{i=1}^n |Y_i - \theta(z_i)| + \lambda_n^2 \sum_{j=1}^n |\alpha_j| \right]$$

$$= \sum_{j=1}^n \hat{\alpha}_{j,n} \psi_j.$$

Note that this estimator has a certain scale invariance, in the sense that the optimal value of the smoothing parameter does not depend on the scale of the data. In particular, the smoothing parameter will not depend on (an estimate of) the variance of the data.

In order to apply Theorem 3.1, we need to verify conditions (3.3) and (3.4). Condition (3.3) depends on the choice θ_* and moreover on the choice of the basis functions $\psi_1, \ldots, \psi_n$. To avoid digressions, we simply take $\theta_* = \theta_0$, so that (3.3) is automatically true.

We will require the boundedness condition (3.5), i.e., that for some constant K,

$$\sup_{\theta \in \Theta} \|\theta\|_\infty \leq K. \tag{6.1}$$

Let us (for simplicity) assume that $W_1, \ldots, W_n$ are i.i.d. copies of a random variable W (with median zero). Suppose W has density f_W with respect to Lebesgue measure, and that for some $\eta > 0$,

$$f_W(w) \geq \eta, \ \text{for all } |w| \leq \eta. \tag{6.2}$$

Then indeed, one may verify (see also VAN DE GEER (1990)) that (3.4) holds with $t_0 = \min(\frac{1}{2}, \frac{\eta}{2K})$. Thus, when (6.1) and (6.2) are met, then Theorem 3.1 holds with $\theta_* = \theta_0$.

Remark 6.1. It is not a good idea to apply LAD with soft thresholding in the sequence space. To see why, recall that the empirical coefficients are given by

$$\tilde{\alpha}_{j,n} = \frac{1}{n} \sum_{i=1}^n Y_i \psi_j(z_i), \ j = 1, \ldots, n.$$

Renormalize to

$$\tilde{Y}_j = \sqrt{n} \tilde{\alpha}_{j,n}, \ j = 1, \ldots, n,$$

with expectation (assuming the errors are centered)

$$\mathbf{E}\tilde{Y}_j = \sqrt{n}\alpha_{j,0} := \vartheta_{j,0}, \ j = 1, \ldots, n.$$

The LAD estimator of ϑ_0 is now

$$\hat{\vartheta}_n = \arg\min_{\vartheta \in \mathbf{R}^n} \left\{ \sum_{j=1}^n |\tilde{Y}_j - \vartheta_j| + \sqrt{n}\lambda_n^2 \sum_{j=1}^n |\vartheta_j| \right\}.$$

Thus, as soon as $\sqrt{n}\lambda_n^2 > 1$, $\hat{\vartheta}_n \equiv 0$.

Suppose now that ϑ_0 remains bounded, say that $|\vartheta_{j,0}| \leq 1$ for all j. This is perhaps not a natural condition, but it is the sequence space counterpart of (6.1), and we need it in order to be able to apply Theorem 3.1. But then result of Theorem 3.1 is actually trivial:

$$\frac{1}{n}\sum_{j=1}^n \vartheta_{j,0}^2 = \frac{1}{n}\sum_{j\in\mathcal{J}_n} \vartheta_{j,0}^2 + \frac{1}{n}\sum_{j\notin\mathcal{J}_n} \vartheta_{j,0}^2$$

$$\leq \frac{N_n}{n} + \frac{1}{n}\sum_{j\notin\mathcal{J}_n} |\vartheta_{j,0}| = \frac{N_n}{n} + \frac{1}{\sqrt{n}}\sum_{j\notin\mathcal{J}_n} |\alpha_{j,0}|.$$

6. SIMULATION STUDY

In our simulation study, we consider LS and LAD. In the LAD case, we will not restrict the functions θ to be bounded in sup-norm by some constant (condition (6.1)).

The signals have been generated using the software MATLAB. We consider the Heavisine function and the Doppler function with $n = 100$ observations, and decompose these two functions onto a wavelet basis using a Daubechies wavelet with 8 vanishing moments.

As error distribution, we considered the standard centered Gaussian distribution with variance 3, and also the Laplacian (i.e., double exponential) distribution, with mean zero and variance 3.

The LS estimator is computed using its explicit expression whereas the LAD estimator must be numerically computed. Since the LAD estimator is defined as the solution of an L_1-minimization, a standard minimization algorithm does not give good results. It *is* however a standard L_1-fitting problem. To see why, write $Y_i = 0$ for $i = n+1, \ldots, 2n$ (data augmentation). Moreover, when $i \in \{n+1, \ldots, 2n\}$ take $\psi_j(z_i) = \lambda_n^2$ for $i = j+n$, and $\psi_j(z_i) = 0$ for $i \neq j+n$, $j = 1, \ldots, n$. To obtain the LAD estimator with soft thresholding type penalty, we now have to minimize over $\alpha \in \mathbf{R}^n$

$$\sum_{i=1}^{2n} |Y_i - \sum_{j=1}^n \alpha_j\psi_j(z_i)|.$$

This is a standard L_1-regression problem (with $2n$ observations and n parameters).

Following ideas developed by BRUCE, SARDY and TSENG (1999) for the Huber loss function, we consider the minimization problem as an optimization problem with Lagrange multipliers and the associated dual, see for instance ROCKAFELLAR (1970). A primal-dual algorithm with a log-barrier penalty, described by CHEN, DONOHO and SAUNDERS (1999) provides an efficient numerical method to compute the LAD estimator.

We have looked at 9 cases, corresponding to different values of the smoothing-parameter λ_n^2 including the theoretical optimal value for the Gaussian case $\lambda_n^2/\sigma = \sqrt{2\log n/n} = 0.303$ that corresponds to the 8th line. The four tables (Figures 4-7) summarize the performance of the LS and LAD estimators in terms of mean square error (MSE). The numbers represent an average over 20 simulations. In order to make comparison of LS and LAD relevant, we have put on a same line the results with λ_n^2/σ for the LS and λ_n^2 for the LAD. We also added a line

where comparisons are made for the optimal cases, i.e., smallest $\|\hat{\theta}_n - \theta_0\|_{Q_n}$ (corresponding to different smoothing parameters).

In these simulations, we can see that LAD works better in the Laplacian case, and LS works better in the Gaussian case (as is to be expected). In the LS case, the value $\lambda_n^2/\sigma = \sqrt{2\log n/n} = 0.303$ is optimal when the errors are Gaussian, but it is too large when the errors are Laplacian.

We show the results for some significant simulations. The LS estimator is represented in dotted lines and the LAD estimator is represented with solid lines. The figures 1-2 show the results obtained for the two different functions Doppler and Heavisine with Gaussian noise. The last figure, figure 3, shows the results when taken Heavisine function corrupted by Laplacian noise. We can observe that LAD catches better the irregularity of the two functions Heavisine and Doppler, when LS is too smooth. However LAD with a wavelet basis may have limitations because its ability to estimate spatially inhomogeneous signals might conflict with the goal of robustness to filter noise.
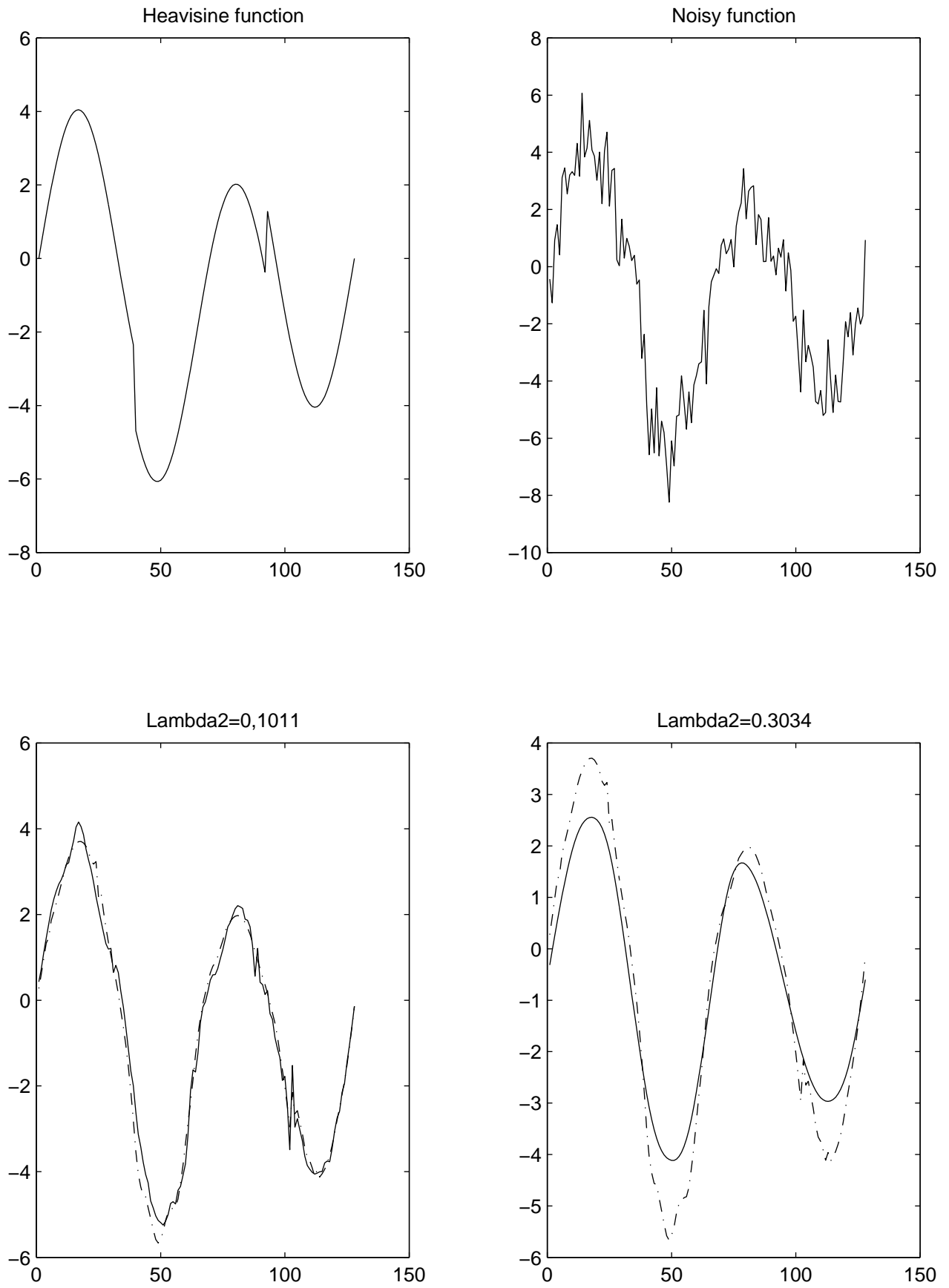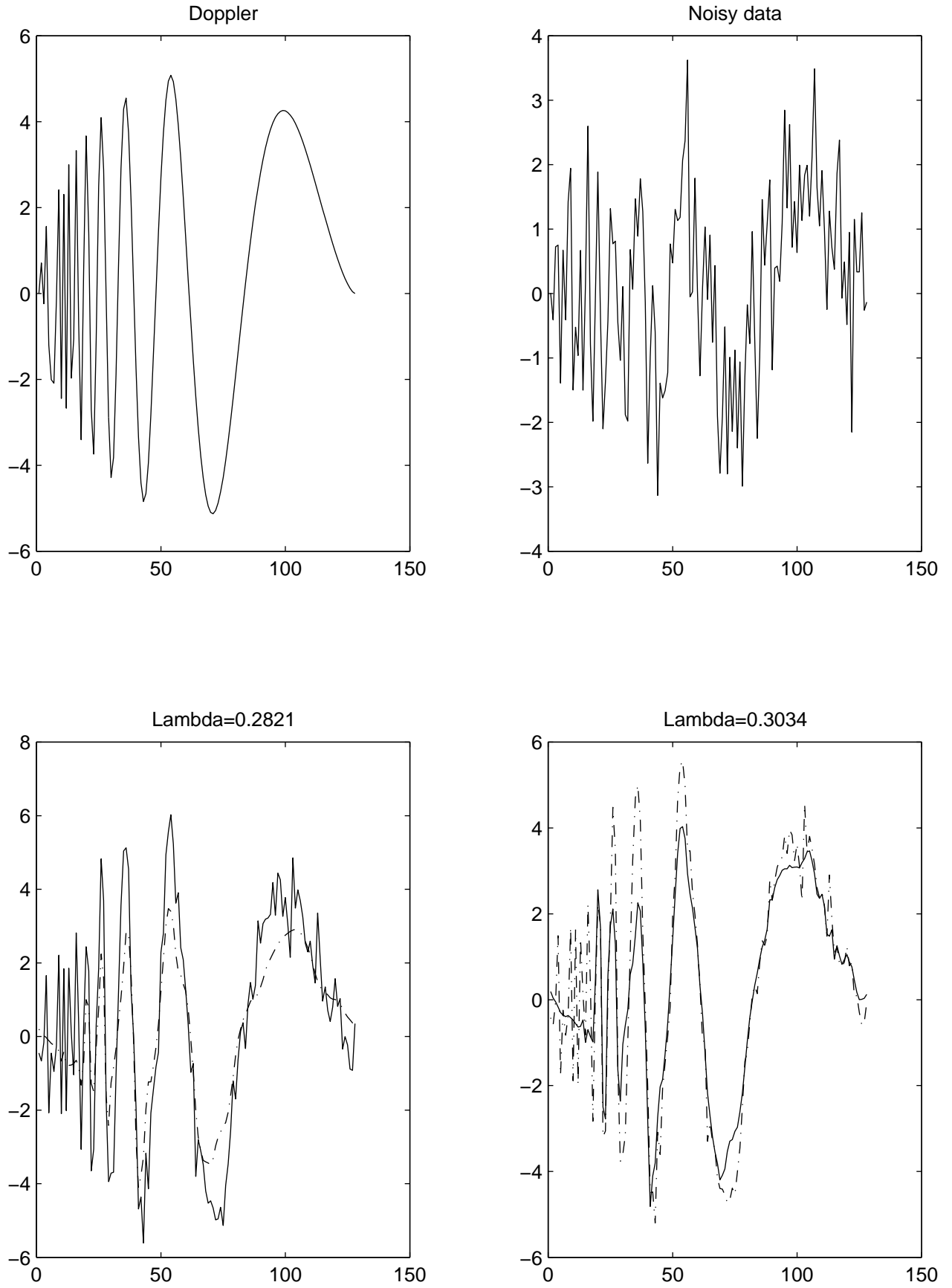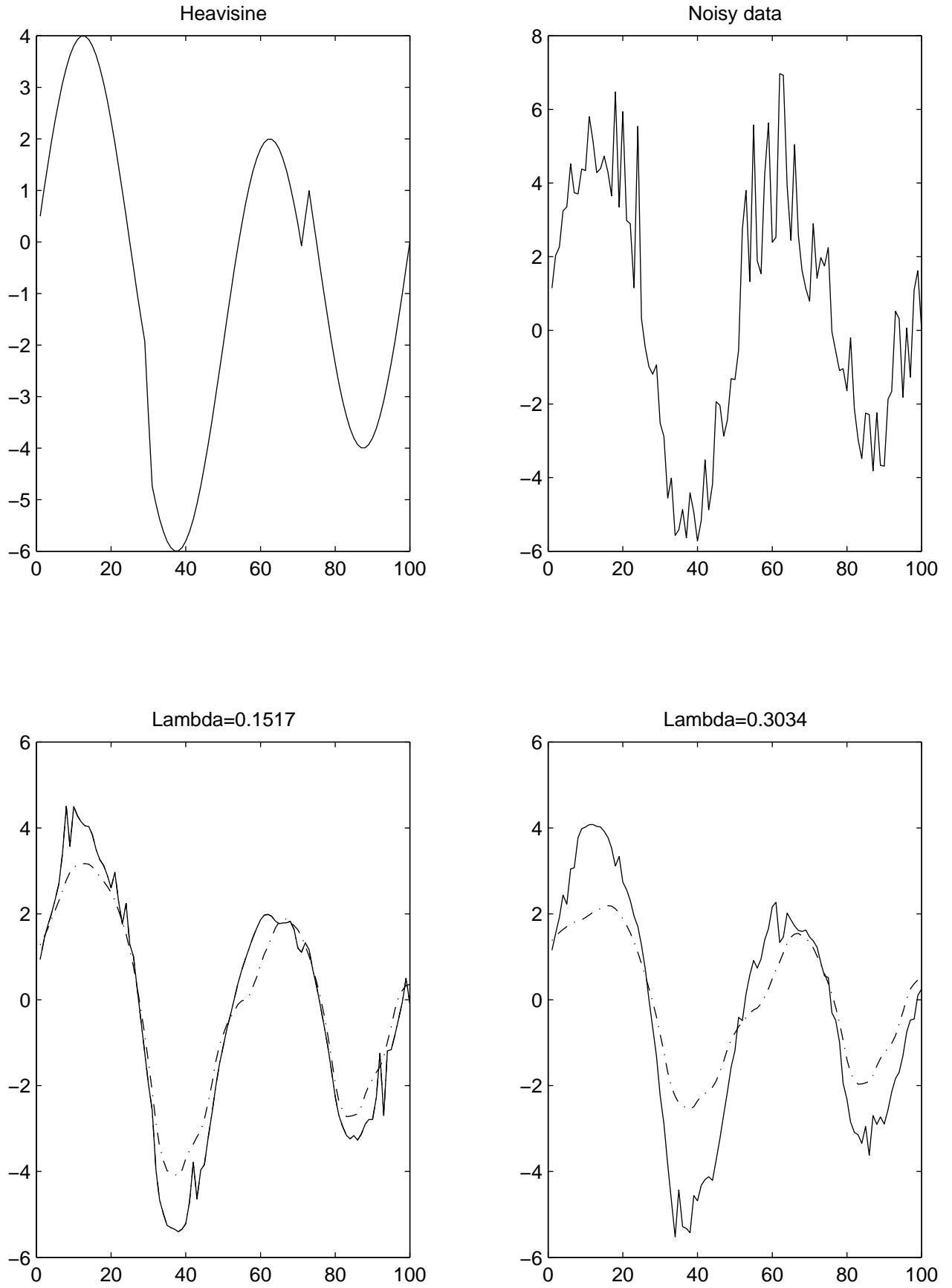
FIGURE 1

FIGURE 2

FIGURE 3

Heavisine function with Gaussian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.7535	0.605
0.0607 (2)	0.5229	0.3994
0.1011 (3)	0.4782	0.3737
0.1517 (4)	0.4934	0.4507
0.2124 (5)	0.4749	0.4612
0.2427 (6)	0.3451	0.4828
0.2731 (7)	0.2821	0.5003
0.3034 (8)	0.2238	0.5601
0.6070 (9)	0.5852	0.6242
optimum	0.2238 at (8)	0.3737 at (3)

FIGURE 4.

Heavisine function with Laplacian noise.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	1.7051	1.5157
0.0607 (2)	1.010	0.954
0.1011 (3)	.8201	0.6238
0.1517 (4)	0.7853	0.5896
0.2124 (5)	0.6021	0.4324
0.2427 (6)	0.5925	0.4654
0.2731 (7)	0.5896	0.5870
0.3034 (8)	0.6012	0.6925
0.607 (9)	0.6238	0.7021
optimum	0.5896 at (7)	0.4324 at (5)

FIGURE 5.

Doppler signal with Gaussian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.5862	0.8103
0.0607 (2)	0.5210	0.7801
0.1011 (3)	0.3521	0.6610
0.1517 (4)	0.2625	0.5218
0.2124 (5)	0.2212	0.3451
0.2427 (6)	0.1521	0.2821
0.2731 (7)	0.1330	0.3299
0.3034 (8)	0.090	0.4445
0.6070 (9)	0.3928	0.5510
optimum	0.090 at (8)	0. 2821 at (6)

FIGURE 6.

Doppler signal with Laplacian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.736	0.901
0.0607 (2)	0.6260	0.7700
0.1011 (3)	0.5218	0.6101
0.1517 (4)	0.5680	0.5018
0.2124 (5)	0.6321	0.3451
0.2427 (6)	0.7081	0.2821
0.2731 (7)	0.8588	0.8229
0.3034 (8)	0.9097	0.8429
0.607 (9)	0.9254	0.9545
optimum	0.5218 at (3)	0.2521 at (6)

FIGURE 7.

References

BARAUD Y. (2000), Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117: 467–493.

BARRON A. and C.-H. SHEU (1991), Approximation of density functions by sequences of exponential families. *The Annals of Statistics* 19: 1347–1369.

BARRON A., L. BIRGÉ, and P. MASSART (1999), Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113: 301–413.

BERGH J. and J. LÖFSTRÖM (1976), *Interpolation spaces - An Introduction.* Springer, New York.

BESOV O., V. P. IL'IN, and S. M. NIKOL'SKIĬ (1978), *Integral Representations of Functions and Embedding Theorems. Vol. I.* V. H. Winston & Sons, Washington, D.C. Translated from the Russian, Scripta Series in Mathematics, Edited by Mitchell H. Taibleson.

BIRGÉ L. and P. MASSART (1997), From model selection to adaptive estimation. In: *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics,* D. Pollard, E. Torgersen, G. Yang (Eds.), Springer, New York, 55–87.

BIRGÉ L. and P. MASSART (2000), An adaptive compression algorithm in Besov spaces. *Journal of Constructive Approximation*, 16: 1–36.

BIRGÉ L. and P. MASSART (2001), Gaussian model selection. *Journal of the European Mathematical Society*, 3: 203–268.

BIRMAN M. and M. Z. SOLOMJAK (1967), Piecewise-polynomial approximations of functions of the classes $W_p{}^{\alpha}$. *Matematisheskiĭ Sbornik (Noveaya Seriya)*, 73: 331–355.

BRUCE A., S. SARDY, P. TSENG (1999), Robust wavelet denoising. *Preprint.*

CHEN S., D. DONOHO, and M. A. SAUNDERS (1999), Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing*, 20: 33–61 (electronic).

DEVORE R. and G. G. LORENTZ (1993), *Constructive Approximation.* Springer-Verlag, Berlin.

DONOHO D., and I. JOHNSTONE (1994a), Minimax risk for l_q losses over l_p-balls. *Probability Theory and Related Fields,* 99: 277–303.

DONOHO D. and I. JOHNSTONE (1994b), Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81: 425–455.

DONOHO D. (1995), De-noising via soft-thresholding. *IEEE Transaction in Information Theory* 41: 613–627.

DONOHO D., and I. JOHNSTONE (1996), Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli* 2: 39–62.

DONOHO D., I JOHNSTONE, G. KERKACHARIAN and D. PICARD (1996a), Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics,* D. Pollard, E. Torgersen, G. Yang (Eds.), Springer, New York, 183–218.

DONOHO D., I. JOHNSTONE, G. KERKYACHARIAN, and D. PICARD (1996b), Density estimation by wavelet thresholding. *The Annals of Statistics*, 24: 508–539.

EDMUNDS E. and H. TRIEBEL (1992), Entropy numbers and approximation numbers in function spaces. II. *Proceedings of the London Mathematical Society (3)*, 64: 153–169.

HOEFFDING W. (1963), Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58: 13–30.

LEDOUX M. and M. TALAGRAND (1991), *Probability in Banach Spaces, Isoperimetry and Processes.* Springer-Verlag, Berlin.

MALLAT S. (1998), *A Wavelet Tour of Signal Processing.* Academic Press Inc., San Diego, CA.

MAMMEN E. and S. A. VAN DE GEER (1997), Locally adaptive regression splines. *The Annals of Statistics* 25: 387–413.

MASSART P. (2000), Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, 9: 245–303.

MEYER Y. (1987), Les ondelettes. In *Contributions to nonlinear partial differential equations, Vol. II (Paris, 1985)*, pages 158–171. Longman Sci. Tech., Harlow.

PORTNOY S. (1997), Local asymptotics for quantile smoothing splines. *The Annals of Statistics* 25: 414–434.

Rockafellar R. (1997) *Convex Analysis*. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Silverman B. (1982), On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics* 10: 795–810.

Stone C. (1990), Large-sample inference for log-spline models. *The Annals of Statistics* 18: 717–741.

Tibshirani R. (1996), Regression analysis and selection via the LASSO. *Journal of the Royal Statistical Society B*, 58: 267–288.

van de Geer S. (1990), Estimating a regression function, *The Annals of Statistics* 18: 907–924.

van de Geer S. (1999), M-estimation using penalties or sieves. *Technical Report* MI 31-99, University of Leiden, 1999, To appear in *Journal of Statistical Planning and Inference.* (2002).

van de Geer S. (2000), *Empirical Processes in M-Estimation*. Cambridge University Press.

van der Vaart A. and J. Wellner (1996), *Weak Convergence and Empirical Processes, with Applications to Statistics*. Springer, New York.

Yang Y. (1999), Model selection for nonparametric regression. *Statistica Sinica* 9: 475–499.