Modulation Estimators and Confidence Sets

Rudolf Beran* and Lutz Dümbgen†

University of California, Berkeley and Universität Heidelberg

January 1996, Revised August 1997

Abstract. An unknown signal plus white noise is observed at n discrete time points. Within a large convex class of linear estimators of ξ, we choose the estimator $\widehat{\xi}$ that minimizes estimated quadratic risk. By construction, $\widehat{\xi}$ is nonlinear. This estimation is done after orthogonal transformation of the data to a reasonable coordinate system. The procedure adaptively tapers the coefficients of the transformed data. If the class of candidate estimators satisfies a uniform entropy condition, then $\widehat{\xi}$ is asymptotically minimax in Pinsker's sense over certain ellipsoids in the parameter space and shares one such asymptotic minimax property with the James-Stein estimator. We describe computational algorithms for $\widehat{\xi}$ and construct confidence sets for the unknown signal. These confidence sets are centered at $\widehat{\xi}$, have correct asymptotic coverage probability, and have relatively small risk as set-valued estimators of ξ.

AMS 1991 subject classifications. Primary 62H12; secondary 62M10.

Key words and phrases. Adaptivity, asymptotic minimax, bootstrap, bounded variation, coverage probability, isotonic regression, orthogonal transformation, signal recovery, Stein's unbiased estimator of risk, tapering.

*Research supported in part by National Science Foundation Grant DMS95-30492 and in part by Sonderforschungsbereich 373 at Humboldt-Universität zu Berlin.

†Research supported in part by European Union Human Capital and Mobility Program ERB CHRX-CT 940693.

1 Introduction

The problem of recovering a signal from observation of the signal plus noise may be formulated as follows. Let $X = X_n = [X(t)]_{t \in T}$ be a random function observed on the set $T = T_n = \{1, 2, \ldots, n\}$. The components $X(t)$ are independent with $\mathbb{E}\, X(t) = \xi(t) = \xi_n(t)$ and $\mathrm{Var}[X(t)] = \sigma^2$ for every $t \in T$. Working with functions on T rather than vectors in $\mathbf{R}^n$ is very convenient for the present purposes. As just indicated, we will usually drop the subscript n for notational simplicity. The signal ξ and the noise variance σ^2 are both unknown. For simplicity we assume throughout that X is Gaussian. Portions of the argument that hold for non-Gaussian X are expressed by the lemmas in Section 6.2.

For any $g \in \mathbf{R}^T$, the space of real-valued functions defined on T, let

$$\mathrm{ave}(g) \;:=\; n^{-1} \sum_{t \in T} g(t).$$

The loss of any estimator $\widehat{\xi}$ for ξ is defined to be

$$L(\widehat{\xi}, \xi) \;:=\; \mathrm{ave}[(\widehat{\xi} - \xi)^2] \tag{1.1}$$

and the corresponding risk of $\widehat{\xi}$ is

$$\rho(\widehat{\xi}, \xi, \sigma^2) \;:=\; \mathbb{E}\, L(\widehat{\xi}, \xi).$$

The first goal is to devise an estimator that is efficient in terms of this risk. If ξ and X are electrical voltages, then $\mathrm{ave}(\xi^2)$ and $L(\widehat{\xi}, \xi)$ are the time-averaged powers dissipated in passing the signal ξ and the error $\widehat{\xi} - \xi$ through a unit resistance.

Any estimator $\widehat{\xi}$ of ξ is governed by the asymptotic minimax bound

$$\liminf_{n \to \infty} \; \inf_{\widehat{\xi}} \; \sup_{\mathrm{ave}(\xi^2) \le c} \rho(\widehat{\xi}, \xi, \sigma^2) \;\ge\; \frac{\sigma^2 c}{\sigma^2 + c} \tag{1.2}$$

for every positive c and σ^2. Inequality (1.2) follows from a more general bound proved by Pinsker (1980) for signal recovery in Gaussian noise (see Nussbaum 1996 and Section 2). It may also be derived from ideas in Stein (1956) by considering best orthogonally equivariant estimators in the submodel where $\mathrm{ave}(\xi^2) = c$ (see Beran 1996b). Let $\widehat{\sigma}^2 = \widehat{\sigma}_n^2$ be an estimator of σ^2 that is consistent as in display (2.2) of Section 2. Then

$$\widehat{\xi}_S \;:=\; [1 - \widehat{\sigma}^2 / \mathrm{ave}(X^2)]^+ X$$

is essentially the James-Stein (1961) estimator, where $[\cdot]^+$ denotes the positive part function. It achieves the Pinsker bound (1.2) because

$$\lim_{n\to\infty} \sup_{\text{ave}(\xi^2)\le c} \rho(\widehat{\xi}_S, \xi, \sigma^2) \;=\; \frac{\sigma^2 c}{\sigma^2 + c} \tag{1.3}$$

for every positive c and σ^2. The limit (1.3) follows from Corollary 2.3 or from asymptotics in Casella and Hwang(1982). For the maximum likelihood estimator $\widehat{\xi}_{\text{ML}} = X$, the risk is always σ^2, which is strictly greater than the Pinsker bound.

Section 2 of this paper constructs estimators of ξ that are asymptotically minimax over a variety of ellipsoids in the parameter space while achieving, in particular, the asymptotic minimax bound (1.2) for every $c > 0$. These *modulation* estimators take the form $\widehat{f}X = [\widehat{f}(t)X(t)]_{t\in T}$. Here $\widehat{f} : T \to [0,1]$ depends on X and is chosen to minimize the estimated risk of the linear estimator fX over all functions f in a class $\mathcal{F} = \mathcal{F}_n \subset [0,1]^T$. Many well-known estimators are of this form with special classes $\mathcal{F}$. In the present paper we analyze such estimators under rather general assumptions on $\mathcal{F}$. How large this class may be is at the heart of the analysis. Taking $\mathcal{F}$ to be the set of all functions from T to $[0,1]$ leads to a poor modulation estimator. Successful is to let $\mathcal{F}$ be a closed convex set of functions with well-behaved uniform covering numbers. One example is the set of all functions in $[0,1]^T$ that are nonincreasing. The asymptotic theory of such modulation estimators, including links with the literature, is the subject of Section 2. Section 4 develops algorithms for computing $\widehat{f}X$ in the example of $\mathcal{F}$ just cited.

Section 3 constructs confidence sets that are centered at a modulation estimator $\widehat{f}X$ and have asymptotic coverage probability α for ξ. The risk of the modulation estimator at the center is shown to determine the risk of the confidence set, when that is viewed as a set-valued estimator for ξ. In this manner, efficiency of a modulation estimator determines the efficiency of the associated confidence set.

Before estimation of ξ, the data X may be transformed orthogonally without changing its Gaussian character. A modulation estimator computed in the new coordinate system can be transformed back into the original coordinate system to yield an estimator of ξ. Standard choices for such preliminary orthogonal transformation include Fourier transforms, wavelet transforms, or analysis-of-variance transforms. When applied in this

manner, modulation estimators perform data-driven tapering of empirical Fourier, wavelet or analysis-of-variance coefficients. Section 5 includes numerical examples of modulation estimators and confidence bounds after Fourier transformation.

2 Modulation estimators

After defining modulation estimators, this section obtains uniform asymptotic approximations to their risks. Let $\mathcal{F} = \mathcal{F}_n$ be a given subset of $[0,1]^T$. Each function $f \in \mathcal{F}$ is called a *modulator* and defines a candidate linear estimator $fX = [f(t)X(t)]_{t\in T}$ for ξ. The risk of this candidate estimator under quadratic loss (1.1) is

$$\rho(fX, \xi, \sigma^2) \;=\; \mathbb{E}\, L(fX, \xi) = \mathrm{ave}[\sigma^2 f^2 + \xi^2(1-f)^2]. \tag{2.1}$$

For brevity, we will write $R(f, \xi, \sigma^2)$ in place of $\rho(fX, \xi, \sigma^2)$.

We will first construct a suitably consistent estimator $\widehat{R}(f)$ of this risk. Suppose that $\widehat{\sigma}^2 = \widehat{\sigma}_n^2$ is an estimator of σ^2, constructed (for instance) by one of the methods described later. Let X^* be a bootstrap random vector in $\mathbf{R}^T$ such that $\mathcal{L}(X^* \mid X, \widehat{\sigma}^2) = \mathcal{N}_T(X, \widehat{\sigma}^2 I)$. The corresponding bootstrap risk estimator for $R(f, \xi, \sigma^2)$ is

$$\mathbb{E}\Big(L(fX^*, X)\,\Big|\, X, \widehat{\sigma}^2\Big) \;=\; R(f, X, \widehat{\sigma}^2).$$

We call $R(f, X, \widehat{\sigma}^2)$ the *naive* risk estimator because it is badly biased upwards, even asymptotically. The key point is

$$\mathbb{E}\, R(f, X, \sigma^2) \;=\; \mathrm{ave}[f^2\sigma^2 + (1-f)^2(\xi^2 + \sigma^2)] \;=\; R(f, \xi, \sigma^2) + \mathrm{ave}[(1-f)^2\sigma^2].$$

Two possible corrections to the naive risk estimator are

$$\begin{aligned}
\widehat{R}_C(f) \;&:=\; \mathrm{ave}[f^2\widehat{\sigma}^2 + (1-f)^2(X^2 - \widehat{\sigma}^2)] \;=\; R(f, X, \widehat{\sigma}^2) - \mathrm{ave}[(1-f)^2\widehat{\sigma}^2], \\
\widehat{R}_B(f) \;&:=\; \max\Big\{\mathrm{ave}(f^2\widehat{\sigma}^2), \widehat{R}_C(f)\Big\} \;=\; \mathrm{ave}(f^2\widehat{\sigma}^2) + \mathrm{ave}[(1-f)^2(X^2 - \widehat{\sigma}^2)]^+.
\end{aligned}$$

Risk estimator $\widehat{R}_C$ is essentially Mallows' (1973) C_L criterion or Stein's (1981) unbiased estimator of risk, with estimation of σ^2 incorporated. Risk estimator $\widehat{R}_B$ corrects the possible negativity in $\mathrm{ave}[(1-f)^2(X^2 - \widehat{\sigma}^2)]$ as an estimator for $\mathrm{ave}[(1-f)^2\xi^2]$. Let X^*

be a random vector in R^T such that $\mathcal{L}(X^* \mid X, \widehat{\sigma}^2)$ is $\mathcal{N}_T(\widehat{\xi}, \widehat{\sigma}^2 I)$, where $\widehat{\xi} = \widehat{\xi}(X, \widehat{\sigma}^2)$ is a vector such that $\mathrm{ave}(\widehat{\xi}^2) = \mathrm{ave}[(1-f)^2(X^2 - \widehat{\sigma}^2)]^+ / \mathrm{ave}[(1-f)^2 X^2]$. Then the bootstrap risk estimator $\mathbb{E}[L(fX^*, \widehat{\sigma}^2) | X, \widehat{\sigma}^2]$ is precisely $\widehat{R}_B$.

Let $\widehat{R}$ denote either $\widehat{R}_C$ or $\widehat{R}_B$. We propose to estimate ξ by the *modulation* estimator $\widehat{f}X$, where $\widehat{f}$ is any function in $\mathcal{F}$ that minimizes $\widehat{R}(f)$. Unless stated otherwise it is assumed throughout that

$$\mathcal{F} \text{ is a closed convex subset of } [0,1]^T \text{ containing all constants } c \in [0,1].$$

Because both $\widehat{R}_C(\cdot)$ and $\widehat{R}_B(\cdot)$ are convex functions on $[0,1]^T$, the minimizer $\widehat{f}$ over $\mathcal{F}$ exists in each case. These minimizers are unique with probability one because $\widehat{R}_C(f)$ is strictly convex in f whenever $X(t) \neq 0$ for every $t \in T$. Similarly, the risk function $R(f, \xi, \sigma^2)$ defined through (2.1) is strictly convex over $[0,1]^T$, with unique minimizer $\widetilde{f}$.

REMARK A. The modulation estimator $\widehat{f}X$ behaves poorly when the class $\mathcal{F}$ is too large. For instance, let $\mathcal{F}$ be the class of all functions in $[0,1]^T$. The minimizer of $R(\cdot, \xi, \sigma^2)$ over $[0,1]^T$ is the "oracle" modulator (cf. Donoho and Johnstone 1994)

$$\widetilde{g} \;:=\; \xi^2/(\sigma^2 + \xi^2),$$

the division being componentwise, while the minimizer of $\widehat{R}(\cdot)$ over $\mathcal{F}$ is now the *greedy* modulator $\widehat{g}^+$, where

$$\widehat{g} \;:=\; (X^2 - \widehat{\sigma}^2)/X^2.$$

To simplify the discussion, suppose that σ^2 is known and $\widehat{\sigma}^2 \equiv \sigma^2$. Then the estimator $\widehat{g}^+X$ is of the general form $\widehat{\xi} := \Big(S[X(t)]\Big)_{t \in T}$ for some measurable function S on the line. Since the maximum likelihood estimator X is componentwise admissible, the risk function $\rho(\widehat{\xi}, \cdot, \sigma^2)$ of $\widehat{\xi}$ is either identical to $\rho(X, \cdot, \sigma^2) \equiv \sigma^2$ or there is a real number θ such that $\int (\theta - S)^2 \, d\mathcal{N}(\theta, \sigma^2) > \sigma^2$. Then, if $\xi(\cdot) \equiv \theta$,

$$\rho(\widehat{\xi}, \xi, \sigma^2) \;>\; \sigma^2 \;=\; \rho(X, \xi, \sigma^2) \;>\; \sigma^2\theta^2/(\sigma^2 + \theta^2),$$

the latter being the asymptotic risk of the James-Stein estimator $\widehat{\xi}_S$. Thus, the maximum risk of $\widehat{g}^+X$ is worse than that of estimators achieving Pinsker's asymptotic minimax bound (1.2) and is even worse than that of the naive estimator X.

It should be mentioned that greedy modulation can be made successful in some sense if one overestimates the variance σ^2 systematically. Donoho and Johnstone (1994) propose threshold estimators of the form $\widehat{\xi} = (1 - \lambda_n \sigma/|X|)^+ X$ or $\widehat{\xi} = 1\{|X| \geq \lambda_n \sigma\} X$, and prove that they have surprising optimality properties if $\lambda_n = (2 \log n)^{1/2}(1 + \delta_n)$ with a suitable sequence $(\delta_n)_n$ tending to zero. These estimators are similar to $\widehat{g}^+ X$ if $\widehat{g}$ is computed with $\widehat{\sigma}_n^2 := \lambda_n^2 \sigma^2$. While showing good performance in case of "sparse signals", these estimators do not achieve the Pinsker bound (1.2) or the minimax bounds in Corollary 2.3 below. Also, the construction of confidence bounds for their loss seems to be intractable. Section 5 illustrates the possibly poor performance of hard thresholding for non-sparse signals.

REMARK B. Kneip's (1994) ordered linear smoothers are equivalent to certain modulation estimators computed after suitable orthogonal transformation of X. The conditions that we impose on $\mathcal{F}$ in this paper are substantially weaker than the ordering of $\mathcal{F}$ required by Kneip. Consequently, our results also apply to the ridge regression, spline estimation, and kernel estimation examples discussed in Kneip's paper. The earlier paper of Li (1987) treated non-diagonal linear estimators indexed by a parameter h. Li's optimality result may be compared with Theorem 2.1 below. However, it does not seem easy to relate Li's conditions on the range of h to our conditions on $\mathcal{F}$. The latter conditions give access to empirical process results that yield asymptotic distributions for the loss of $\widehat{f}X$ and hence confidence sets for ξ centered at modulation estimators.

REMARK C. Nussbaum (1996) surveyed constructions of adaptive estimators that achieve Pinsker-type asymptotic minimax bounds. For instance, Golubev and Nussbaum (1992) treated adaptive, asymptotically minimax estimation when $\xi_i = g(x_i)$ and g lies in an ellipsoid of unknown radius within a Sobolev space of unknown order. Corollary 2.3 below is of related character. However, our results make no smoothness assumptions on ξ. For instance, sample paths up to time n of suitably scaled, discrete-time, independent white noise ultimately lie, as $n \to \infty$, within the ball $\mathrm{ave}(\xi^2) \leq c$.

Useful classes of modulators $\mathcal{F}$ can be characterized through their uniform covering numbers, which are defined as follows. For any probability measure Q on T, consider the

pseudo-distance $d_Q(f,g)^2 := \int (f-g)^2\, dQ$ on $[0,1]^T$. For every positive u, let

$$N(u,\mathcal{F},d_Q) \;:=\; \min\Bigl\{\#\mathcal{F}_o : \mathcal{F}_o \subset \mathcal{F},\; \inf_{f_o\in\mathcal{F}_o} d_Q(f_0,f) \le u \;\forall f\in\mathcal{F}\Bigr\}.$$

Define the uniform covering number

$$N(u,\mathcal{F}) \;:=\; \sup_Q N(u,\mathcal{F},d_Q),$$

where the supremum is taken over all probabilities on T. Let

$$J(\mathcal{F}) \;:=\; \int_0^1 \sqrt{\log N(u,\mathcal{F})}\, du.$$

Throughout C denotes a generic universal real constant which does not depend on n, ξ, σ^2 or $\mathcal{F}$, but whose value may be different in various places.

THEOREM 2.1 *Let $\mathcal{F}$ be any closed subset of $[0,1]^T$ containing 0, let $\tilde{f}$ be a minimizer of $R(f,\xi,\sigma^2)$ over $f\in\mathcal{F}$, and let $\widehat{f}$ minimize either $\widehat{R}_C(f)$ or $\widehat{R}_B(f)$ over $f\in\mathcal{F}$. Then*

$$\mathbb{E}\Bigl|G - R(\tilde{f},\xi,\sigma^2)\Bigr| \;\le\; C\Bigl(J(\mathcal{F})\frac{\sigma^2+\sigma\sqrt{\mathrm{ave}(\xi^2)}}{\sqrt{n}} + \mathbb{E}\,|\widehat{\sigma}^2-\sigma^2|\Bigr),$$

where G is any one of the following quantities:

$$L(\widehat{f}X,\xi),\quad \inf_{f\in\mathcal{F}} L(fX,\xi),\quad \widehat{R}_C(\widehat{f}),\quad \widehat{R}_B(\widehat{f}).$$

In particular,

$$\Bigl|\rho(\widehat{f}X,\xi,\sigma^2) - R(\tilde{f},\xi,\sigma^2)\Bigr| \;\le\; C\Bigl(J(\mathcal{F})\frac{\sigma^2+\sigma\sqrt{\mathrm{ave}(\xi^2)}}{\sqrt{n}} + \mathbb{E}\,|\widehat{\sigma}^2-\sigma^2|\Bigr).$$

This theorem is about convergence of losses and risks. The next result uses convexity of $\mathcal{F}$ to establish that $\widehat{f}$ and $\tilde{f}$, as well as $\widehat{f}X$ and $\tilde{f}X$, converge to one another. Note that the second bound holds uniformly in $\xi\in\mathbf{R}^T$.

THEOREM 2.2 *Let $\widehat{f}$ be the minimizer of $\widehat{R}_C$. Then*

$$\begin{aligned}
\mathbb{E}\,\mathrm{ave}\Bigl((\sigma^2+\xi^2)(\widehat{f}-\tilde{f})^2\Bigr) &\le CJ(\mathcal{F})\frac{\sigma^2+\sigma\sqrt{\mathrm{ave}(\xi^2)}}{\sqrt{n}} + \mathbb{E}\,|\widehat{\sigma}^2-\sigma^2|,\\
\mathbb{E}\,\mathrm{ave}\Bigl((\widehat{f}X-\tilde{f}X)^2\Bigr) &\le CJ(\mathcal{F})\frac{\sigma^2}{\sqrt{n}} + \mathbb{E}\,|\widehat{\sigma}^2-\sigma^2|.
\end{aligned}$$

Given consistency of $\hat{\sigma}$ and boundedness of $\sigma^2 + \mathrm{ave}(\xi^2)$, a key assumption on $\mathcal{F}$ that ensures success of the modulation estimator $\hat{f}X$ defined above is that $J(\mathcal{F}) = o(n^{1/2})$. Here are some examples of modulator classes $\mathcal{F}$ to which Theorem 2.1 applies.

EXAMPLE 1 (Stein shrinkage). Suppose that $\mathcal{F}$ consists of all constant functions in $[0,1]^T$. The minimizer over $\mathcal{F}$ of $R(f, \xi, \sigma^2)$ is

$$\tilde{f}_S \equiv 1 - \sigma^2/[\sigma^2 + \mathrm{ave}(\xi^2)].$$

The minimizer of both $\widehat{R}_C$ and $\widehat{R}_B$ is

$$\hat{f}_S \equiv [1 - \hat{\sigma}^2/\,\mathrm{ave}(X^2)]^+.$$

The resulting modulation estimator $\hat{f}_S X$ is the (modified) James-Stein (1981) estimator $\hat{\xi}_S$. Here one easily shows that $N(u, \mathcal{F}) \le 1 + (2u)^{-1}$, whence $J(\mathcal{F})$ is bounded by a universal constant.

EXAMPLE 2 (Multiple Stein shrinkage). Let $\mathcal{B} = \mathcal{B}_n$ be a partition of T and define

$$\mathcal{F} := \Big\{ \sum_{B \in \mathcal{B}} 1_B \, c(B) : c \in [0,1]^{\mathcal{B}} \Big\},$$

where 1_B is the indicator function of B. The values of $c(B)$ that define $\tilde{f}$ and $\hat{f}$, respectively, are

$$\begin{aligned} \tilde{c}(B) &= \mathrm{ave}(1_B \xi^2)/\,\mathrm{ave}[1_B(\sigma^2 + \xi^2)], \\ \hat{c}(B) &= \mathrm{ave}[1_B(X^2 - \hat{\sigma}^2)]^+/\,\mathrm{ave}(1_B X^2). \end{aligned}$$

The modulation estimator $\hat{f}X$ now has the asymptotic form of the multiple shrinkage estimator in Stein (1966). Elementary calculations show that $N(u, \mathcal{F}) \le [1 + (2u)^{-1}]^{\#\mathcal{B}}$. Thus $J(\mathcal{F})$ is bounded by a universal constant times $(\#\mathcal{B})^{1/2}$, so that $J(\mathcal{F}) = o(n^{1/2})$ follows from the intuitively appealing condition $\#\mathcal{B} = o(n)$.

EXAMPLE 3 (Monotone shrinkage). Let $\mathcal{F}_{\mathrm{mon}}$ be the set of all nonincreasing functions in $[0,1]^T$. The class of candidate estimators $\{fX : f \in \mathcal{F}_{\mathrm{mon}}\}$ includes the nested model-selection estimators $f_k X$, $0 \le k \le n$, defined by $f_k(t) := 1\{t \le k\}$. In fact, $\mathcal{F}_{\mathrm{mon}}$ is the

convex hull of $\mathcal{D}_{MS} := \{f_0, f_1, \ldots, f_n\}$. Elementary calculations show that

$$N(u, \mathcal{D}_{MS}) \;\le\; 1 + u^{-2} \;\le\; 2u^{-2}$$

for $0 < u \le 1$. Together with Theorem 5.1 of Dudley (1987) it follows that

$$\log N(u, \mathcal{F}_{\mathrm{mon}}) \;\le\; Cu^{-1} \quad \text{for all } u \in]0, 1].$$

EXAMPLE 4 (Monotone shrinkage with respect to a quasi-order). Let $\preceq$ be a quasi-order relation on T (cf. Robertson *et al.* 1988, Chapter 1.3), and let $\mathcal{F}_{\preceq}$ be the set of all functions in $[0,1]^T$ that are nonincreasing with respect to $\preceq$. That means, for all $f \in \mathcal{F}_{\preceq}$ and $s, t \in T$,

$$f(s) \;\ge\; f(t) \quad \text{if } s \preceq t.$$

Here one can easily deduce from the conclusion of Example 3 that

$$\log N(u, \mathcal{F}_{\preceq}) \;\le\; CN_{\preceq} u^{-1}$$

for $0 < u \le 1$, where $N_{\preceq} = N_{\preceq,n}$ is the minimal cardinality of a partition of $(T, \preceq)$ into totally ordered subsets. Thus $J(\mathcal{F}_{\preceq})$ is of order $O(N_{\preceq}^{1/2})$. To give an example, suppose that X consists of $n = 2^{k+1} - 1$ empirical Haar (or wavelet) coefficients, arranged as a binary tree. If this tree is equipped with its natural order $\preceq$, then the monotonicity constraint $\widehat{f} \in \mathcal{F}_{\preceq}$ means that $\widehat{f}X$ is a mixture of histogram estimators (cf. Engel 1994). Here $N_{\preceq} = 2^k > n/2$. Therefore, in order to apply our theory one has to replace the class $\mathcal{F}_{\preceq}$ with suitable subclasses.

EXAMPLE 5 (Shrinkage with bounded total variation). Let $\mathcal{F}_{(M)}$ be all functions f in $[0,1]^T$ with total variation not greater than $M = M_n$, i.e.

$$\sum_{t=2}^{n} |f(t) - f(t-1)| \;\le\; M.$$

For instance, the class of functions $f(t) := \max\{\min\{p(t), 1\}, 0\}$, where p is a polynomial of degree less than or equal to M, belongs to $\mathcal{F}_{(M)}$. Any $f \in \mathcal{F}_{(M)}$ can be written as $(M+1)(f_1 - f_2)$ with $f_1, f_2 \in \mathcal{F}_{\mathrm{mon}}$. Hence

$$\log N(u, \mathcal{F}_{(M)}) \;\le\; 2 \log N\Big([2(M+1)]^{-1} u, \mathcal{F}_{\mathrm{mon}}\Big) \;\le\; C(M+1)u^{-1}$$

for $0 < u \le 1$. In particular, $J(\mathcal{F}_{(M)}) = O[(M+1)^{1/2}]$.

The minimizers $\widetilde{f}$ and $\widehat{f}$ in Examples 3-5 lack closed forms. Section 4 describes computational algorithms for $\widetilde{f}$ and $\widehat{f}$ in Examples 3-4. Example 5 differs from the remaining examples both theoretically as well as computationally and will be treated in detail elsewhere.

A particular consequence of Theorem 2.1 is that the modulation estimators are asymptotically minimax optimal for a large class of submodels for (ξ, σ^2). Namely, for $a \in [1,\infty]^T$ and $c > 0$ define the linear minimax risk

$$\nu^2(a, c, \sigma^2) \;:=\; \inf_{g \in [0,1]^T} \sup_{\mathrm{ave}(a\xi^2) \le c} R(g, \xi, \sigma^2).$$

It is shown by Pinsker (1980) that the linear minimax risk approximates the unrestricted minimax risk in that

$$\inf_{\widehat{\xi}} \sup_{\mathrm{ave}(a\xi^2) \le c} \rho(\widehat{\xi}, \xi, \sigma^2)/\nu^2(a, c, \sigma^2) \;\to\; 1 \quad \text{as } n\nu^2(a, c, \sigma^2) \to \infty.$$

Moreover,

$$\nu^2(a, c, \sigma^2) \;=\; \sup_{\mathrm{ave}(a\xi^2) \le c} R(g_o, \xi, \sigma^2) \;=\; R(g_o, \xi_o, \sigma^2),$$

where $g_o := [1 - (a/\mu_o)^{1/2}]^+$, $\xi_o^2 := \sigma^2[(\mu_o/a)^{1/2} - 1]^+$, and $\mu_o > 0$ is the unique real number satisfying $\mathrm{ave}(a[(\mu_o/a)^{1/2} - 1]^+) = c/\sigma^2$. The special case $a \equiv 1$ yields (1.2).

If the minimax modulator $g_o = g_o(\cdot \,|\, a, c/\sigma^2)$ happens to be in $\mathcal{F}$, which is certainly true for $a \equiv 1$, then

$$\sup_{\mathrm{ave}(a\xi^2) \le c} \rho(\widehat{f}X, \xi, \sigma^2) \;\le\; \sup_{\mathrm{ave}(a\xi^2) \le c} \left|\rho(\widehat{f}X, \xi, \sigma^2) - R(\widetilde{f}, \xi, \sigma^2)\right| + \nu^2(a, c, \sigma^2).$$

Thus Theorem 2.1 immediately implies the following minimax result, where the distribution of $(X, \widehat{\sigma}^2)$ is assumed to depend on (ξ, σ^2) only.

COROLLARY 2.3 *Suppose that $J(\mathcal{F}) = o(n^{1/2})$, and that for every $c, \sigma^2 > 0$,*

$$\epsilon_n(c, \sigma^2) := \sup_{\mathrm{ave}(\xi^2) \le c} \mathbb{E}\,|\widehat{\sigma}^2 - \sigma^2| \;\to\; 0 \quad (n \to \infty). \tag{2.2}$$

Then the modulation estimator $\widehat{f}X$ achieves the asymptotic minimax bound (1.2).

More generally, let $a = a_n \in [1,\infty]^T$ *such that*

$$[1 - (a/\mu)^{1/2}]^+ \ \in \ \mathcal{F} \quad \textit{for all constants } \mu \geq 1. \tag{2.3}$$

Then for every $c, \sigma^2 > 0$,

$$\sup_{\mathrm{ave}(a\xi^2)\leq c} \rho(\widehat{f}X, \xi, \sigma^2) \ \leq \ \nu^2(a, c, \sigma^2) \ + \ O[n^{-1/2}J(\mathcal{F}) + \epsilon_n(c, \sigma^2)]. \qquad \square$$

Specifically, let $a(t) = 1$ for $t \in A \subset T$ and $a(t) = \infty$ otherwise. Then $\mathrm{ave}(a\xi^2) \leq c$ is equivalent to $\mathrm{ave}(\xi^2) \leq c$ and $\xi^2 = 0$ on $T \setminus A$. Here one can easily see that condition (2.3) is equivalent to $1_A \in \mathcal{F}$. The linear minimax risk equals

$$\nu^2(a, c, \sigma^2) \ = \ \frac{\sigma^2 \,\mathrm{ave}(1_A)c}{\sigma^2 \,\mathrm{ave}(1_A) + c},$$

which can be significantly smaller than the bound in (1.2).

In case of $\mathcal{F} = \mathcal{F}_{\mathrm{mon}}$ condition (2.3) is equivalent to a being nondecreasing on T.

We end this section with some examples for $\widehat{\sigma}$. Internal estimators of σ^2 depend only on X and require additional smoothness or dimensionality restrictions on the possible values of ξ to achieve the consistency property (2.2). One internal estimator of σ^2, analyzed by Rice (1984) and by Gasser et al. (1986) is

$$\widehat{\sigma}^2_{(1)} = [2(n-1)]^{-1} \sum_{t=2}^{n} [X(t) - X(t-1)]^2. \tag{2.4}$$

Here $\mathbb{E}\,|\widehat{\sigma}^2_n - \sigma^2_n| \to 0$ as $n \to \infty$ and

$$n^{-1} \sum_{t=2}^{n} [\xi_n(t) - \xi_n(t-1)]^2 \ \to \ 0.$$

External estimators of variance are available in linear models, where one observes an N-dimensional normal random vector Y with mean $\mathbb{E}\,Y = D\xi$ and covariance matrix $\mathrm{Cov}(Y) = \sigma^2 I_N$ for some design matrix $D \in \mathbf{R}^{N\times n}$, $N = N_n > n$. After suitable linear transformation of Y and ξ one may assume that ξ is the expectation of the vector $X := (Y_1, Y_2, \ldots, Y_n)$. Then the standard estimator for σ^2 is given by

$$\widehat{\sigma}^2_{(2)} \ := \ (N-n)^{-1} \sum_{i=n+1}^{N} Y_i^2,$$

which is independent from X with $(N-n)\sigma^{-2}\widehat{\sigma}^2_{(2)} \sim \chi_{N-n}$. This estimator also satisfies (2.2), provided that $N - n \to \infty$.

3 Confidence sets

Having replaced the maximum likelihood estimator X with $\widehat{f}X$, a natural question is to what extent $\widehat{f}X$ is closer to the unknown signal ξ than X. More precisely we want to compare the distance $L(X, \widehat{f}X)^{1/2}$ with an upper confidence bound $\widehat{r} = \widehat{r}(X, \widehat{\sigma}^2)$ for $L(\xi, \widehat{f}X)^{1/2}$. In geometrical terms, the confidence ball of primary interest is defined by

$$\widehat{C} = \widehat{C}_n \ := \ \{\gamma \in \mathbf{R}^T : L(\widehat{f}X, \gamma) \leq \widehat{r}^2\}.$$

The radius $\widehat{r}$ is chosen so that the coverage probability $\mathbb{P}(\xi \in \widehat{C})$ converges to $\alpha \in]0,1[$ as n increases. The full definition of $\widehat{C}$ follows the theorem below. Underlying the construction is the confidence set idea sketched at the end of Stein (1981). The quality of $\widehat{C}$ as a set-valued estimator of ξ will be measured through the quadratic loss

$$L(\widehat{C}, \xi) \ := \ \sup_{\gamma \in \widehat{C}} L(\gamma, \xi) \ = \ [L(\widehat{f}X, \xi)^{1/2} + \widehat{r}]^2. \tag{3.1}$$

This is a natural extension of the quadratic loss defined in (1.1) and has an appealing projection-pursuit interpretation; see Beran (1996a).

One main assumption for this section is that

$$\begin{aligned} &X_n \text{ and } \widehat{\sigma}_n^2 \text{ are independent with } \mathcal{L}(\sigma_n^{-2}\widehat{\sigma}_n^2) \text{ depending only on } n \\ &\text{such that} \quad \lim_{n\to\infty} m\Big(\mathcal{L}[n^{1/2}(\sigma^{-2}\widehat{\sigma}_n^2 - 1)], \mathcal{N}(0, \beta^2)\Big) \ = \ 0. \end{aligned} \tag{3.2}$$

Here $\beta^2 \geq 0$ is a given constant and $m(\cdot, \cdot)$ metrizes weak convergence of distributions on the line. For instance, the estimator $\widehat{\sigma}^2_{(2)}$ of Section 2 satisfies Condition (3.2) with $\beta := 2 \lim_{n\to\infty} n/(N_n - n)$, provided that this limit exists. Condition (3.2) is made for the sake of simplicity. It could be replaced with weaker, but more technical conditions in order to include special internal estimators of variance such as $\widehat{\sigma}^2_{(1)}$. A second key assumption is that

$$\int_0^1 \sqrt{\sup_n N(u, \mathcal{F}_n)}\, du \ < \ \infty. \tag{3.3}$$

Roughly speaking, this condition allows us to pretend that $\widehat{f}$ is equal to $\widetilde{f}$. It is satisfied in all Examples 1-5, provided that $\#\mathcal{B}_n = O(1)$ in Example 2, $N_{\preceq,n} = O(1)$ in Example 4, and $M_n = O(1)$ in Example 5.

At first let us consider confidence balls centered at the naive estimator X. Since $n\sigma^{-2}\,\mathrm{ave}[(X-\xi)^2]$ has a chi-squared distribution with n degrees of freedom, we consider

$$\widehat{C}_{\mathrm{N}} := \Big\{\gamma \in \mathbf{R}^T : \mathrm{ave}[(X-\gamma)^2] \le \widehat{\sigma}^2(1+n^{-1/2}c)\Big\}$$

for some fixed c. The inequality $\mathrm{ave}[(X-\xi)^2] \le \widehat{\sigma}^2(1+n^{-1/2}c)$ is equivalent to

$$n^{1/2}\Big(\sigma^{-2}\,\mathrm{ave}[(X-\xi)^2]-1\Big) - n^{1/2}(\sigma^{-2}\widehat{\sigma}^2-1) \;\le\; \sigma^{-2}\widehat{\sigma}^2 c \;=\; c+o_p(1).$$

Thus the Central Limit Theorem for the chi-squared distribution together with Condition 3.2 implies that $c=(2+\beta^2)^{1/2}\Phi^{-1}(\alpha)$ yields a confidence set $\widehat{C}_{\mathrm{N}}$ with

$$\lim_{n\to\infty}\sup_{\xi\in\mathbf{R}^T,\ \sigma^2>0}\Big|\mathbb{P}\{\xi\in\widehat{C}_{\mathrm{N}}\}-\alpha\Big| \;=\; 0,$$

where $\Phi^{-1}(\alpha)$ denotes the α-th quantile of $\mathcal{N}(0,1)$. Moreover,

$$\lim_{n\to\infty}\sup_{\xi\in\mathbf{R}^T}\mathbb{P}\Big\{|L(\widehat{C}_{\mathrm{N}},\xi)-4\sigma^2|>\epsilon\Big\} \;=\; 0 \quad \forall\,\epsilon>0.$$

In what follows we shall see that confidence sets centered at a good modulation estimator $\widehat{f}X$ dominate the naive confidence set $\widehat{C}_{\mathrm{N}}$ in terms of the loss $L(\widehat{C},\xi)$.

To construct these confidence sets, we first determine the asymptotic distribution of

$$\widehat{d}=\widehat{d}_n \;:=\; n^{1/2}[L(\widehat{f}X,\xi)-\widehat{R}_C(\widehat{f})].$$

This difference compares the loss of $\widehat{f}X$ with an estimate for the expected loss of $\widehat{f}X$.

THEOREM 3.1 *Under Conditions (3.2, 3.3),*

$$\lim_{n\to\infty}\sup_{\mathrm{ave}(\xi^2)\le c} m[\mathcal{L}(\widehat{d}),\mathcal{N}(0,\tau^2)] \;=\; 0,$$

for arbitrary $c,\sigma^2>0$, where

$$\tau^2=\tau_n^2(\xi,\sigma^2) \;:=\; 2\sigma^4\,\mathrm{ave}[(2\widetilde{f}-1)^2]+\beta^2\sigma^4[\mathrm{ave}(2\widetilde{f}-1)]^2+4\sigma^2\,\mathrm{ave}[\xi^2(1-\widetilde{f})^2].$$

A consistent estimator $\widehat{\tau}^2=\widehat{\tau}_n^2$ of τ^2 is obtained by substituting $\widehat{\sigma}^2$ for σ^2, $\widehat{f}$ for $\widetilde{f}$ and $X^2-\widehat{\sigma}^2$ for ξ^2 in the expression for τ^2. The implied estimator of the approximating normal

distribution $\mathcal{N}(0,\tau^2)$ is $\mathcal{N}(0,\widehat{\tau}^2)$. This leads to the following definition of a confidence ball for ξ that is centered at the modulation estimator $\widehat{f}X$:

$$\widehat{C} \;:=\; \Big\{\gamma \in \mathbf{R}^T : L(\widehat{f}X,\gamma) \le \widehat{R}_C(\widehat{f}) + n^{-1/2}\widehat{\tau}\,\Phi^{-1}(\alpha)\Big\}.$$

The intended coverage probability of $\widehat{C}$ is α. The next theorem establishes asymptotic properties of this confidence set construction. Beran (1994) treats in detail the example where $\widehat{f}X$ is the James-Stein estimator. That situation is much easier to analyze than the general case.

THEOREM 3.2 *Under the conditions of Theorem 3.1, for arbitrary $c, \sigma^2 > 0$,*

$$\lim_{n\to\infty,K\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\Big\{|L(\widehat{C},\xi) - 4R(\widetilde{f},\xi,\sigma^2)| \ge Kn^{-1/2}\Big\} \;=\; 0$$

$$\textit{and} \quad \lim_{n\to\infty,K\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\Big\{|\widehat{r}^2 - R(\widetilde{f},\xi,\sigma^2)| \ge Kn^{-1/2}\Big\} \;=\; 0.$$

Moreover, $\widehat{\tau}^2$ is consistent in that

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\Big\{|\widehat{\tau}^2 - \tau^2| > \epsilon\Big\} \;=\; 0 \quad \forall\, \epsilon > 0.$$

If

$$\liminf_{n\to\infty} \inf_{\mathrm{ave}(\xi^2)\le c} \tau_n^2(\xi,\sigma^2) \;>\; 0, \tag{3.4}$$

then

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \Big|\mathbb{P}\{\xi \in \widehat{C}\} - \alpha\Big| \;=\; 0.$$

A sufficient condition for (3.4) is the following: For every n, $\mathcal{F} = \mathcal{F}_n$ is such that

$$1\{f \ge c\}f \;\in\; \mathcal{F} \quad \textit{for all } f \in \mathcal{F} \textit{ and } c \in [0,1]. \tag{3.5}$$

Condition (3.4) ensures that $\mathcal{L}(\widehat{d})$ does not approach a degenerate distribution. Note that Condition (3.5) is satisfied in Examples 1-4. When $R(\widetilde{f},\xi,\sigma^2) = O(n^{-1/2})$ our confidence ball has loss $L(\widehat{C},\xi) = O_p(n^{-1/2})$. In fact, according to Theorem 2.1 of Li (1989) this is the smallest possible order of magnitude for a Euclidean confidence ball, unless one imposes further constraints on the signal ξ. The result (3.2) on asymptotic coverage of $\widehat{C}$ may be compared with the lower bound in Theorem 3.2 of Li (1989).

A key step in the proof of Theorem 3.1 is that in the definition of $\widehat{d}$ one may replace $\widetilde{f}$ with $\widehat{f}$. Instead of the normal approximation underlying $\widehat{C}$ a bootstrap approximation of $H = H_n := \mathcal{L}(\widehat{d})$ that imitates the estimation of $\widetilde{f}$ seems to be more reliable in moderate dimension. Precisely, let $\widehat{H} = \widehat{H}_n$ be the conditional distribution (function) of $\widehat{d}^*$ given $(X, \widehat{\sigma}^2)$, where $\widehat{d}^*$ is computed as $\widehat{d}$ with the pair $(X^*, \widehat{\sigma}^{2*})$ in place of $(X, \widehat{\sigma}^2)$. More precisely, let $\widehat{\xi} = \widehat{\xi}(\cdot \,|\, X, \widehat{\sigma}^2)$ be an estimator for ξ. Let S_n^2 be a random variable with a specified distribution depending only on n such that

$$\lim_{n\to\infty} m\Big(\mathcal{L}[n^{1/2}(S_n^2 - 1)], \mathcal{N}(0, \beta^2)\Big) \;=\; 0,$$

where S_n and $(X, \widehat{\sigma}^2)$ are independent. Then

$$\mathcal{L}(X^*, \widehat{\sigma}^{2*} \,|\, X, \widehat{\sigma}^2) \;=\; \mathcal{N}(\widehat{\xi}, \widehat{\sigma}^2 I) \otimes \mathcal{L}(\widehat{\sigma}^2 S_n^2 \,|\, X, \widehat{\sigma}^2),$$

the product of the probability measures $\mathcal{N}(\widehat{\xi}, \widehat{\sigma}^2 I)$ and $\mathcal{L}(\widehat{\sigma}^2 S_n^2 \,|\, X, \widehat{\sigma}^2)$. The resulting bootstrap confidence bound $\widehat{r}_b(\alpha)$ for $L(\xi, \widehat{f}X)$ is given by

$$\widehat{r}_b^2(\alpha) \;=\; \widehat{R}(\widehat{f}) + n^{-1/2}\widehat{H}^{-1}(\alpha).$$

The last theorem of this section states conditions, under which $\widehat{H}$ is a consistent estimator for H. An interesting fact is that neither $\widehat{\xi} = X$ nor $\widehat{\xi} = \widehat{f}X$ satisfy these conditions.

THEOREM 3.3 *Under the assumptions of Theorem 3.1,*

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\Big\{|m(\widehat{H}_n, H_n)| > \epsilon\Big\} \;=\; 0 \quad \forall\, \epsilon > 0,$$

provided that

$$\widehat{f} \;=\; \arg\min_{f\in\mathcal{F}} R(f, \widehat{\xi}, \widehat{\sigma}^2) \quad \textit{almost surely}, \tag{3.6}$$

$$\limsup_{n\to\infty, K\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\{\mathrm{ave}(\widehat{\xi}^2) > K\} = 0, \tag{3.7}$$

$$\lim_{n\to\infty} \sup_{\mathrm{ave}(\xi^2)\le c} \mathbb{P}\Big\{\Big|\mathrm{ave}[\widehat{\xi}^2(1-\widehat{f})^2] - \mathrm{ave}[\xi^2(1-\widetilde{f})^2]\Big| > \epsilon\Big\} \;=\; 0 \quad \forall\, \epsilon > 0. \tag{3.8}$$

In particular, suppose that each $\mathcal{F}_n$ has the following property: For all $X, \xi \in \mathbf{R}^T$ with $X^2 > 0$ and any $c \in [0,1]$,

$$c \;=\; \mathrm{ave}[1\{\widehat{f} = c\}(X^2 - \widehat{\sigma}^2)]^+ / \,\mathrm{ave}(1\{\widehat{f} = c\}X^2) \quad \textit{if } \{\widehat{f} = c\} \neq \emptyset,$$

$$c \;=\; \mathrm{ave}(1\{\widetilde{f} = c\}\xi^2) / \,\mathrm{ave}[1\{\widetilde{f} = c\}(\sigma^2 + \xi^2)] \quad \textit{if } \{\widetilde{f} = c\} \neq \emptyset.$$

Then the function $\widehat{\xi} := \left(\widehat{\sigma}^2 \widehat{f}/(1-\widehat{f})\right)^{1/2}$ *satisfies Conditions (3.6, 3.7, 3.8).*

One can show that the last part of Theorem 3.3 applies to Examples 1-4. This follows from the representation of $\widehat{f}$ given in Section 4 and Robertson *et al.* (1988, Theorem 1.3.5). Here one can also show that $\widehat{\xi} := \widehat{f}^{1/2} X$ satisfies these requirements, too. This yields a natural extension of the bootstrap method proposed by Beran (1994).

4 Computation of $\widehat{f}$ and $\widetilde{f}$

We restrict our attention to $\widehat{R} = \widehat{R}_C$. With the oracle modulator $\widetilde{g} := \xi^2/(\sigma^2+\xi^2)$ and its naive estimator $\widehat{g} := (X^2 - \widehat{\sigma}^2)/X^2$ one can write

$$\begin{aligned} R(f,\xi,\sigma^2) &= \text{ave}[(f-\widetilde{g})^2(\sigma^2+\xi^2)] + \text{ave}(\sigma^2\widetilde{g}), \\ \widehat{R}_C(f) &= \text{ave}[(f-\widehat{g})^2 X^2] + \text{ave}(\widehat{\sigma}^2\widehat{g}). \end{aligned}$$

Hence both functions $\widetilde{f}$ and $\widehat{f}$ are metric projections of some function g onto $\mathcal{F}$.

Now we consider the family $\mathcal{F}_{\preceq}$ of Example 4. Note that this case includes Examples 1-3 as special cases. The family $\mathcal{F}_{\preceq}$ can be written as $\mathcal{H}_{\preceq} \cap [0,1]^T$, where

$$\mathcal{H}_{\preceq} := \{\text{nonincreasing functions } h \in \mathbf{R}^T \text{ with respect to } \preceq\}.$$

Given arbitrary $h \in \mathcal{H}_{\preceq}$ and $a \in \mathbf{R}$, the functions $\max\{h,a\}$ and $\min\{h,a\}$ also belong to $\mathcal{H}_{\preceq}$. Consequently, since $0 \le \widetilde{g} \le 1$ and $\widehat{g} \le 1$,

$$\widetilde{f} = \operatorname*{arg\,min}_{h \in \mathcal{H}_{\preceq}} \text{ave}[(h-\widetilde{g})^2(\sigma^2+\xi^2)] \quad \text{and} \quad \widehat{f} = \operatorname*{arg\,min}_{h \in \mathcal{H}^+_{\preceq}} \text{ave}[(h-\widehat{g})^2 X^2],$$

where $\mathcal{H}^+_{\preceq} := \{h \in \mathcal{H} : h \ge 0\}$. Hence $\widetilde{f}$ can be computed by any algorithm for projections onto $\mathcal{H}_{\preceq}$, while in case of $\widehat{f}$ one has to deal with a nonnegativity constraint. (Replacing $\widehat{g}$ with $\widehat{g}^+$ would yield an inconsistent estimator for $\widetilde{f}$ in general.) The latter problem is easy to solve. Let $\widehat{h}$ be the unique unrestricted projection

$$\widehat{h} := \operatorname*{arg\,min}_{h \in \mathcal{H}_{\preceq}} \text{ave}[(h-\widehat{g})^2 X^2]$$

of $\widehat{g}$. Then

$$\widehat{f} = \widehat{h}^+.$$

For if $\widehat{h} = \widehat{h}^+ - \widehat{h}^-$, then $\widehat{f} - \widehat{h}^- \in \mathcal{H}_{\preceq}$ and $\widehat{h}^+ \in \mathcal{H}^+_{\preceq}$. Hence

$$\begin{aligned}
\operatorname{ave}[(\widehat{h} - \widehat{g})^2 X^2] &\leq \operatorname{ave}[(\widehat{f} - \widehat{h}^- - \widehat{g})^2 X^2] \\
&= \operatorname{ave}[(\widehat{f} - \widehat{g})^2 X^2] + \operatorname{ave}[(\widehat{h}^- + \widehat{g})^2 X^2] - \operatorname{ave}[\widehat{g}^2 X^2] \\
&\quad - 2 \operatorname{ave}[\widehat{f}\widehat{h}^- X^2] \\
&\leq \operatorname{ave}[(\widehat{f} - \widehat{g})^2 X^2] + \operatorname{ave}[(\widehat{h}^- + \widehat{g})^2 X^2] - \operatorname{ave}[\widehat{g}^2 X^2] \\
&\leq \operatorname{ave}[(\widehat{h}^+ - \widehat{g})^2 X^2] + \operatorname{ave}[(\widehat{h}^- + \widehat{g})^2 X^2] - \operatorname{ave}[\widehat{g}^2 X^2] \\
&= \operatorname{ave}[(\widehat{h} - \widehat{g})^2 X^2],
\end{aligned}$$

by definition of $\widehat{f}$ and $\widehat{h}$. Thus $\widehat{f} - \widehat{h}^-$ equals $\widehat{h}$, whence $\widehat{f} = \widehat{h}^+$.

Explicit algorithms for computing $\widehat{h}$ are described in Robertson *et al.* (1988, Section 1). Because of the special form of $\widetilde{g}$ and $\widehat{g}$ one can even replace weighted least squares by ordinary least squares. Namely, let

$$H_\xi \;:=\; \operatorname*{arg\,min}_{h \in \mathcal{H}_{\preceq}} \operatorname{ave}\Big((h - \xi^2)^2\Big) \quad \text{and} \quad H_X \;:=\; \operatorname*{arg\,min}_{h \in \mathcal{H}_{\preceq}} \operatorname{ave}\Big((h - X^2)^2\Big).$$

Then

$$\widetilde{f} \;=\; H_\xi/(\sigma^2 + H_\xi) \quad \text{and} \quad \widehat{f} \;=\; (H_X - \widehat{\sigma}^2)^+/H_X.$$

This follows from the min-max formula for antitonic regression (Robertson *et al.* 1988, Theorem 1.4.4). For let L and U be generic lower and upper sets, respectively. That means,

$$L \;=\; \Big\{y \in T : y \preceq x \text{ for some } x \in L\Big\} \quad \text{and} \quad U \;=\; \Big\{y \in T : x \preceq y \text{ for some } x \in U\Big\}.$$

Then

$$\begin{aligned}
\widehat{h}(t) &= \max_{L: t \in L} \min_{U: t \in U} \operatorname{ave}(1_{L \cap U} X^2 \widehat{g}) / \operatorname{ave}(1_{L \cap U} X^2) \\
&= \max_{L: t \in L} \min_{U: t \in U} \Big(1 - \widehat{\sigma}^2 \operatorname{ave}(1_{L \cap U}) / \operatorname{ave}(1_{L \cap U} X^2)\Big) \\
&= 1 - \widehat{\sigma}^2 \Big/ \max_{L: t \in L} \min_{U: t \in U} \Big(\operatorname{ave}(1_{L \cap U} X^2) / \operatorname{ave}(1_{L \cap U})\Big) \\
&= 1 - \widehat{\sigma}^2 / H_X(t).
\end{aligned}$$

The formula for $\widetilde{f}$ is proved analogously.

5 Numerical examples

In this section we apply the proposed methods to empirical Fourier coefficients. We simulated data

$$Y(z) \;=\; \mu(z/n) + \epsilon(z), \quad z \in T,$$

where $n = 1000$, $\epsilon \sim \mathcal{N}_T(0, I)$, and μ is one of the following two functions on $[0,1]$:

$$\text{Case 1:} \quad \mu(u) \;:=\; 2\,[6.75u^2(1-u)]^3,$$

$$\text{Case 2:} \quad \mu(u) \;:=\; \begin{cases} 1.5 & \text{if } 1.5 < u < 0.3, \\ 0.5 & \text{if } 0.3 < u \le 0.6, \\ 2 & \text{if } 0.6 < u \le 0.8, \\ 0 & \text{else,} \end{cases}$$

With the orthonormal functions $\phi_1 \equiv n^{1/2}$, $\phi_{2k-1}(z) := (2n)^{1/2}\cos(2\pi kz/n)$ and $\phi_{2k}(z) := (2n)^{1/2}\sin(2\pi kz/n)$ $(1 \le k < n/2)$, and $\phi_{n/2}(z) := n^{1/2}(-1)^z$ on T these data were transformed into

$$X \;:=\; [\mathrm{ave}(\phi_t Y)]_{t\in T} \;=\; [\mathrm{ave}(\phi_t \mu(\cdot/n))]_{t\in T} + [\mathrm{ave}(\phi_t \epsilon)]_{t\in T} \;=:\; \xi + E,$$

so that $E \sim \mathcal{N}_T(0, I)$. Then we computed the modulation estimator $\widehat{f}X$ using $\widehat{\sigma}^2 \equiv 1$ (for simplicity) and the modulator class

$$\mathcal{F} \;:=\; \Big\{ f \in \mathcal{F}_{\mathrm{mon}} : f_{2k-1} = f_{2k} \text{ for } 1 \le k < n/2 \Big\}.$$

This yielded the estimator

$$\widehat{\mu}(z/n) \;:=\; \sum_{t\in T} \mathrm{ave}(\phi_t \widehat{f} X)\phi_t(z)$$

for $\mu(z/n)$. The additional requirement on $f \in \mathcal{F}_{\mathrm{mon}}$ takes into account the ambiguity of labeling sine and cosine functions of the same frequency. It also makes the resulting estimator $\widehat{\mu}$ equivariant under cyclical shifts of the data Y.

Figures 1a and 2a depict the data Y and the estimator $\widehat{\mu}(\cdot/n)$ in the two cases, respectively. Figures 1b and 2b show the estimator $\widehat{\mu}(\cdot/n)$ and the true function $\mu(\cdot/n)$. Figures 1a' and 2a' contain Y and $\widehat{\mu}(\cdot/n)$, too, but this time the modulator $\widehat{f}$ was computed with $\widehat{\sigma} \equiv 0.9$ and $\widehat{\sigma} \equiv 1.1$, respectively. A higher estimated variance leads to a

smoother estimate. These plots show clearly that estimating the variance *is* a crucial step. They also indicate the possibility to pick $\widehat{\sigma}$ visually.

Figures 1c and 2c show what is going on in the Fourier domain. The first plot shows the ideal greedy modulator $\widetilde{g}$ and the ideal monotone modulator $\widetilde{f} \in \mathcal{F}$. The second plot gives the empirical counterparts $\widehat{g}^+$ and $\widehat{f}$. Apparently the functions $\widetilde{g}$ and $\widehat{g}^+$ have little in common. On the other hand, the estimator $\widehat{f}$ and $\widetilde{f}$ are close to one another, as predicted by Theorem 2.2.

Note that a (hard) threshold estimator keeps all coefficients $X(t)$ of X such that $\widehat{g}(t)$ is above a certain level while replacing the remaining coefficients with zero. In examples the authors looked at, this often led to peculiar estimators using only very few low frequencies or including some high frequencies. For cases 1 and 2, Figure 3 shows "oracle" (in a strong sense defined by the next display) threshold estimators $\widetilde{\mu}_{\mathrm{th}}(\cdot/n) := \sum_{t\in T} \mathrm{ave}(\phi_t \widetilde{\xi}_{\mathrm{th}})\phi_t$, where $\widetilde{\xi}_{\mathrm{th}}(t) := 1\{|X(t)| \geq c_{\mathrm{th}}\}X(t)$ and

$$c_{\mathrm{th}} = c_{\mathrm{th}}(X,\xi) \;:=\; \operatorname*{arg\,min}_{c\geq 0} L\Big(1\{|X| \geq c\}X, \xi\Big).$$

In Case 1 the function μ is very smooth, thus leading to a sparse signal ξ. In fact, the threshold fit is excellent. In Case 2 the threshold fit seems useless.

Finally we computed the bootstrap upper confidence bounds $\widehat{r}_b^2(0.9)$ for the actual loss $L(\widehat{f}X,\xi)$ as described in Section 3. The quantile $\widehat{H}^{-1}(0.9)$ was estimated in 3000 Monte-Carlo simulations. Table 1 contains the distance $L(X,\widehat{f}X)$, the estimated risk $\widehat{R}(\widehat{f})$, the bootstrap bound $\widehat{r}_b^2(0.9)$, the actual loss $L(\widehat{f}X,\xi)$ as well as the loss $L(\widetilde{\xi}_{\mathrm{th}},\xi)$.

	$L(\widehat{f}X,X)$	$\widehat{R}(\widehat{f})$	$\widehat{r}_b^2(0.9)$	$L(\widehat{f}X,\xi)$	$L(\widetilde{\xi}_{\mathrm{th}},\xi)$
Case 1	0.9317	0.0108	0.0801	0.0109	0.0083
Case 2	0.9170	0.0755	0.1441	0.0546	0.1013

Table 1

6 Proofs

6.1 Auxiliary results

Our results utilize well-known techniques from empirical process theory. Theorem 6.1 below follows from standard symmetrization arguments and Pisier's (1983) version of the

Chaining lemma (see also Pollard 1990, Sections 2 and 3). Theorem 6.2 is a simplified and modified version of Alexander's (1987) general results (see also Pollard 1990, Theorem 10.6).

Let $S = \sum_{i=1}^n \phi_i$ with independent stochastic processes $\phi_1, \phi_2, \ldots, \phi_n$ on an index set $\mathcal{T}$. Examples for S are empirical processes and partial sum processes; see also Pollard (1990). For technical reasons we suppose that all ϕ_i have continuous paths with respect to some metric d on $\mathcal{T}$ such that $(\mathcal{T}, d)$ is separable. Now define a random pseudodistance $\widehat{\rho}$ on $\mathcal{T}$ via

$$\widehat{\rho}(s,t)^2 \ := \ \sum_{i=1}^n [\phi_i(s) - \phi_i(t)]^2.$$

Further let

$$\rho(s,t) \ := \ \mathbb{E}(\widehat{\rho}(s,t)^2)^{1/2}.$$

For any pseudo-metric ν on $\mathcal{T}$ define the covering numbers

$$N(u, \mathcal{T}, \nu) \ := \ \min\Big\{ \#\mathcal{T}_o : \mathcal{T}_o \subset \mathcal{T}, \ \inf_{t_o \in \mathcal{T}_o} \nu(t_o, t) \le u \ \forall\, t \in \mathcal{T} \Big\}.$$

THEOREM 6.1 *Suppose that $S(t_o) \equiv 0$ for some $t_o \in \mathcal{T}$. Then*

$$\mathbb{E}\, \|S - \mathbb{E}\, S\|_{\mathcal{T}} \ \le \ C\, \mathbb{E} \int_0^{\widehat{D}} \sqrt{\log N(u, \mathcal{T}, \widehat{\rho})}\, du,$$

where $\widehat{D} := \sup_{t \in \mathcal{T}} \widehat{\rho}(t, t_o)$ and $\|x\|_{\mathcal{T}} := \sup_{t \in \mathcal{T}} |x(t)|$. □

THEOREM 6.2 *Suppose that $\mathcal{T} = \mathcal{T}_n$, $\phi_i = \phi_{n,i}$ depend on n such that the following conditions are satisfied as $n \to \infty$:*

$$(6.1) \quad \mathbb{E} \sum_{i=1}^n \|\phi_i\|_{\mathcal{T}}^2 = O(1) \quad \textit{and} \quad \mathbb{E} \sum_{i=1}^n 1\{\|\phi_i\|_{\mathcal{T}}^2 > u\}\, \|\phi_i\|_{\mathcal{T}}^2 = o(1) \ \textit{for all } u > 0;$$

$$(6.2) \qquad \int_0^{\epsilon(n)} \sqrt{\log N(u, \mathcal{T}, \widehat{\rho})}\, du \ \to_p \ 0 \quad \textit{whenever } \epsilon(n) \downarrow 0.$$

Then

$$\mathbb{E}\, \|\widehat{\rho}^2 - \rho^2\|_{\mathcal{T} \times \mathcal{T}} = o(1) \quad \textit{and} \quad N(u, \mathcal{T}, \rho) = O(1) \ \textit{for all } u > 0,$$

$$\sup_{s,t \in \mathcal{T}: \rho(s,t) \le \alpha} \Big| (S - \mathbb{E}\, S)(s) - (S - \mathbb{E}\, S)(t) \Big| \ \to_p \ 0 \quad \textit{as } n \to \infty, \alpha \downarrow 0,$$

$$\|S - \mathbb{E}\, S\|_{\mathcal{T}} \ = \ O_p(1).$$

Moreover, let $a_n : \mathcal{T} \to \mathbf{R}$ *be arbitrary functions such that* $\sum_{t\in\mathcal{T}} |a_n(t)| = O(1)$. *Then*

$$m\Big(\mathcal{L}\Big(\sum_{t\in\mathcal{T}} a_n(t)(S - \mathbb{E}\,S)(t)\Big), \mathcal{N}\Big(0, \mathrm{Var}\Big[\sum_{t\in\mathcal{T}} a_n(t)S(t)\Big]\Big)\Big) \;\to\; 0. \qquad \square$$

6.2 Proofs for Section 2

With the vector $E := X - \xi$ of residuals we define random functions

$$W_1 \;:=\; E^2 - \sigma^2 \quad \text{and} \quad W_2 \;:=\; \xi E$$

on T. Then one can write

$$\begin{aligned} L(fX, \xi) - R(f, \xi, \sigma^2) &= \mathrm{ave}\{[fE - (1-f)\xi]^2 - f^2\sigma^2 - (1-f)^2\xi^2\} \\ &= \mathrm{ave}[f^2 W_1 + 2(f^2 - f)W_2]. \end{aligned}$$

Moreover, $X^2 = E^2 + 2\xi E + \xi^2 = W_1 + 2W_2 + \sigma^2 + \xi^2$, and with

$$V \;:=\; \widehat{\sigma}^2 - \sigma^2$$

one obtains

$$\begin{aligned} &\widehat{R}_C(f) - R(f, \xi, \sigma^2) \\ &\quad= \mathrm{ave}[f^2\widehat{\sigma}^2 + (1-f)^2(X^2 - \widehat{\sigma}^2) - f^2\sigma^2 - (1-f)^2\xi^2] \\ &\quad= \mathrm{ave}[(f^2 - 2f + 1)(W_1 + 2W_2) + (2f-1)V], \\ &\widehat{R}_B(f) - R(f, \xi, \sigma^2) \\ &\quad= \mathrm{ave}\Big((1-f)^2\xi^2 + (1-f)^2(W_1 + 2W_2 - V)\Big)^+ - \mathrm{ave}[(1-f)^2\xi^2 + f^2 V] \\ &\quad= \lambda(f)\,\mathrm{ave}\Big((f^2 - 2f + 1)(W_1 + 2W_2) + [f^2 - \lambda(f)(1-f)^2]V\Big), \end{aligned}$$

where $0 \le \lambda(f) \le 1$. Hence the following inequalities hold:

LEMMA 6.3

$$\begin{aligned} \sup_{f\in\mathcal{F}} \Big|\widehat{R}(f) - R(f, \xi, \sigma^2)\Big| &\le 4 \sup_{g\in\mathcal{G}} |\,\mathrm{ave}(gW_1)| + 8 \sup_{g\in\mathcal{G}} |\,\mathrm{ave}(gW_2)| + |V|, \\ \sup_{f\in\mathcal{F}} \Big|L(fX, \xi) - R(f, \xi, \sigma^2)\Big| &\le \sup_{g\in\mathcal{G}} |\,\mathrm{ave}(gW_1)| + 4 \sup_{g\in\mathcal{G}} |\,\mathrm{ave}(gW_2)|, \end{aligned}$$

where $\mathcal{G} := \{fg : f, g \in \mathcal{F}\}$.

If $\mathcal{F}$ is the convex hull of a family $\mathcal{D}$ of (indicators of) subsets of T, which is closed under intersection, then $\mathcal{G} = \mathcal{F}$. In fact, in that case one may replace $J(\mathcal{F})$ in Theorems 2.1, 2.2 and Corollary 2.3 with $J(\mathcal{D})$.

Theorem 2.1 follows easily from Lemma 6.3 and the following result, which is stated for potentially non-Gaussian error E.

LEMMA 6.4 *Let $X = \xi + E$, where E has independent components with mean zero and variance σ^2. Then*

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathcal{G}} |\operatorname{ave}(gW_1)| &\leq C J(\mathcal{F}) \sqrt{\frac{\mathbb{E} \operatorname{ave}(E^4)}{n}}, \\ \mathbb{E} \sup_{g \in \mathcal{G}} |\operatorname{ave}(gW_2)| &\leq C J(\mathcal{F}) \sqrt{\frac{\sigma^2 \operatorname{ave}(\xi^2)}{n}}. \end{aligned}$$

Proof of Lemma 6.4. Elementary calculations show that the uniform covering numbers of $\mathcal{G}$ satisfy

$$N(u, \mathcal{G}) \leq N(u/2, \mathcal{F})^2 \quad \text{for all } u > 0.$$

Thus $J(\mathcal{G}) \leq 4J(\mathcal{F})$, and $\mathcal{F}$ may be replaced with $\mathcal{G}$. Now the asserted inequalities for W_1, W_2 are direct consequences of Theorem 6.1. For let $\mathcal{T} := \mathcal{G}$ and

$$\phi_i(g) := \begin{cases} n^{-1} E(i)^2 g(i) & \text{for } j = 1, \\ n^{-1} \xi(i) E(i) g(i) & \text{for } j = 2. \end{cases}$$

Then $\sup_{g \in \mathcal{G}} |\operatorname{ave}(gW_j)| = \|S - \mathbb{E}\, S\|_{\mathcal{G}}$.

For $j = 1$, and arbitrary $g, h \in \mathcal{G}$,

$$\begin{aligned} \widehat{\rho}(g,h)^2 &= n^{-1} \operatorname{ave}(E^4 (g-h)^2) \\ &= n^{-1} \operatorname{ave}(E^4) d_{\widehat{Q}}(g,h)^2 \\ &\leq n^{-1} \operatorname{ave}(E^4) \end{aligned}$$

for some (random) probability measure $\widehat{Q}$ on T. Thus $\widehat{D} \leq n^{-1/2} \operatorname{ave}(E^4)^{1/2}$ and

$$N(u, \mathcal{G}, \widehat{\rho}) \leq N[n^{1/2} \operatorname{ave}(E^4)^{-1/2} u, \mathcal{G}].$$

Hence

$$\begin{aligned} \int_0^{\widehat{D}} \sqrt{\log N(u, \mathcal{G}, \widehat{\rho})}\, du &\leq \int_0^{n^{-1/2} \operatorname{ave}(E^4)^{1/2}} \sqrt{\log N[n^{1/2} \operatorname{ave}(E^4)^{-1/2} u, \mathcal{G}]}\, du \\ &= n^{-1/2} \operatorname{ave}(E^4)^{1/2} \int_0^1 \sqrt{\log N(u, \mathcal{G})}\, du, \end{aligned}$$

and $\mathbb{E}[\mathrm{ave}(E^4)^{1/2}] \le [\mathbb{E}\,\mathrm{ave}(E^4)]^{1/2}$.

For $j = 2$ the same arguments yield

$$\int_0^{\widehat{D}} \sqrt{\log N(u, \mathcal{G}, \widehat{\rho})}\, du \;\le\; n^{-1/2}\,\mathrm{ave}(\xi^2 E^2)^{1/2} \int_0^1 \sqrt{\log N(u, \mathcal{G})}\, du,$$

and $\mathbb{E}[\mathrm{ave}(\xi^2 E^2)^{1/2}] \le \sigma\,\mathrm{ave}(\xi^2)^{1/2}$. □

Proof of Theorem 2.2. Let $w_1 := X^2$, $w_2 := \sigma^2 + \xi^2$, $g_1 := \widehat{g}$ and $g_2 := \widetilde{g}$. Then $\widehat{f} = f_1$ and $\widetilde{f} = f_2$, where generally

$$f_i \;:=\; \arg\min_{f \in \mathcal{F}} \mathrm{ave}\Big(w_i (f - g_i)^2\Big).$$

Let (i, j) be either $(1, 2)$ or $(2, 1)$. By convexity of $\mathcal{F}$,

$$\frac{\partial}{\partial t}\Big|_{t=0} \mathrm{ave}\Big(w_i [f_i + t(f_j - f_i) - g_i]^2\Big) \;=\; 2\,\mathrm{ave}\Big(w_i (f_i - g_i)(f_j - f_i)\Big) \;\ge\; 0. \tag{6.3}$$

Hence,

$$\begin{aligned}
&\frac{\partial}{\partial t}\Big|_{t=\mathbf{1}} \mathrm{ave}\Big(w_i [f_i + t(f_j - f_i) - g_i]^2\Big) \\
&\quad= \; 2\,\mathrm{ave}\Big(w_i (f_i - g_i)(f_j - f_i)\Big) + 2\,\mathrm{ave}\Big(w_i (f_j - f_i)^2\Big) \\
&\quad\ge \; 2\,\mathrm{ave}\Big(w_i (f_j - f_i)^2\Big).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\frac{\partial}{\partial t}\Big|_{t=\mathbf{1}} \mathrm{ave}\Big(w_i [f_i + t(f_j - f_i) - g_i]^2\Big) \;&=\; -\frac{\partial}{\partial t}\Big|_{t=\mathbf{0}} \mathrm{ave}\Big(w_i [f_j + t(f_i - f_j) - g_i]^2\Big) \\
&=\; -2\,\mathrm{ave}\Big(w_i (f_j - g_i)(f_i - f_j)\Big) \\
&\le\; 2\,\mathrm{ave}\Big([w_j (f_j - g_j) - w_i (f_j - g_i)](f_i - f_j)\Big),
\end{aligned}$$

where the latter inequality follows from (6.3) with i and j interchanged. Thus we end up with the inequality

$$\mathrm{ave}(w_i (f_j - f_i)^2) \;\le\; \mathrm{ave}\Big([w_j (f_j - g_j) - w_i (f_j - g_i)](f_i - f_j)\Big).$$

Elementary algebra yields

$$w_j (f_j - g_j) - w_i (f_j - g_i) \;=\; \begin{cases} (\widehat{f} - 1)(W_1 + 2W_2) + V & \text{if } (i, j) = (1, 2), \\ (1 - \widetilde{f})(W_1 + 2W_2) - V & \text{if } (i, j) = (2, 1). \end{cases}$$

Consequently, Lemma 6.4 yields

$$\begin{aligned}
\mathbb{E}\,\mathrm{ave}\Big(X^2(\widehat{f}-\widetilde{f})^2\Big) &\le 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\mathrm{ave}\Big((1-\widetilde{f})(W_1+2W_2)f\Big)\Big| + \mathbb{E}\,|V| \\
&\le CJ(\mathcal{F})n^{-1/2}\Big(\sigma^2+\sigma\,\mathrm{ave}[(1-\widetilde{f})^2\xi^2]^{1/2}\Big) + \mathbb{E}\,|V| \\
&\le CJ(\mathcal{F})n^{-1/2}\sigma^2 + \mathbb{E}\,|V|, \\
\mathbb{E}\,\mathrm{ave}\Big((\sigma^2+\xi^2)(\widehat{f}-\widetilde{f})^2\Big) &\le 4\,\mathbb{E}\sup_{g\in\mathcal{G}}\Big|\mathrm{ave}((W_1+2W_2)g)\Big| + \mathbb{E}\,|V| \\
&\le CJ(\mathcal{F})n^{-1/2}\Big(\sigma^2+\sigma\,\mathrm{ave}(\xi^2)^{1/2}\Big) + \mathbb{E}\,|V|.
\end{aligned}$$

In the first case we applied Lemma 6.4 with $(1-\widetilde{f})\xi$ in place of ξ and utilized the fact that $\mathrm{ave}\Big((1-\widetilde{f})^2\xi^2\Big) \le R(\widetilde{f},\xi,\sigma^2) \le R(1,\xi,\sigma^2) = \sigma^2$. □

6.3 Proofs for Section 3

Throughout this subsection asymptotic statements are meant as $n\to\infty$ uniformly in $\mathrm{ave}(\xi_n^2)\le c$, where $c,\sigma^2>0$ are arbitrary and fixed.

Proof of Theorem 3.1. At first it is shown that

$$\widehat{d} \;=\; \widehat{d}(\widetilde{f}) + o_p(1), \tag{6.4}$$

where $\widehat{d}(f) := n^{1/2}[L(fX,\xi) - \widehat{R}_C(f)]$, i.e. $\widehat{d} = \widehat{d}(\widehat{f})$. The formulas of Section 6.2 for $L(fX,\xi)$ and $\widehat{R}_C(f)$ yield

$$\widehat{d}(f) \;=\; n^{1/2}\,\mathrm{ave}\Big((2f-1)W_1 + 2(f-1)W_2 + (1-2f)V\Big).$$

In particular,

$$\widehat{d}(f) - \widehat{d}(\widetilde{f}) \;=\; 2n^{1/2}\,\mathrm{ave}\Big((f-\widetilde{f})(W_1+W_2-V)\Big).$$

It follows from the proof of Theorem 2.2 that

$$\mathrm{ave}\Big((\sigma^2+\xi^2)(\widehat{f}-\widetilde{f})^2\Big) \;=\; o_p(1), \tag{6.5}$$

Hence $|n^{1/2}V\,\mathrm{ave}(\widehat{f}-\widetilde{f})| = o_p(1)$, and it suffices to show that

$$n^{1/2}\,\mathrm{ave}[(\widehat{f}-\widetilde{f})W_j] \;=\; o_p(1) \quad \text{for } j=1,2. \tag{6.6}$$

For $j = 1$ one can apply Theorem 6.2 with $\mathcal{T} = \mathcal{F}$ and $\phi_i(f) := n^{-1/2} f(i) E(i)^2$ similarly as in the proof of Theorem 2.1. The assumptions of Theorem 6.2 are satisfied, because

$$\begin{aligned}
\mathbb{E} \sum_{i=1}^{n} \|\phi_i\|_{\mathcal{F}}^2 &= \mathbb{E}\,\mathrm{ave}(E^4) \;=\; 3\sigma^4, \\
\mathbb{E} \sum_{i=1}^{n} 1\{\|\phi_i\|_{\mathcal{F}}^2 > u\}\, \|\phi_i\|_{\mathcal{F}}^2 &= \mathbb{E}\,\mathrm{ave}(1\{E^4 > nu\} E^4) \;=\; o(1) \quad \forall\, u > 0, \\
\int_0^{\epsilon(n)} \sqrt{\log N(u, \mathcal{F}, \widehat{\rho})}\, du &\leq \mathrm{ave}(E^4)^{1/2} \int_0^{\epsilon(n)/\,\mathrm{ave}(E^4)^{1/2}} \sqrt{N(u, \mathcal{F})}\, du \\
&\to_p \; 0 \quad \text{whenever } \epsilon(n) \downarrow 0.
\end{aligned}$$

Hence assertion (6.6) for $j = 1$ follows from the fact that

$$\rho(\widehat{f}, \widetilde{f})^2 \;=\; \mathbb{E}\,\mathrm{ave}[(f - \widetilde{f})^2 E^4]\Big|_{f=\widehat{f}} \;=\; 3\sigma^4\,\mathrm{ave}[(\widehat{f} - \widetilde{f})^2] \;=\; o_p(1),$$

according to (6.5).

As for $j = 2$, $f \mapsto n^{1/2}\,\mathrm{ave}(fW_2)$ is a centered Gaussian process with continuous paths with respect to $\rho(f, g) = \sigma\,\mathrm{ave}[\xi^2 (f - g)^2]^{1/2} = \mathrm{Var}\{n^{1/2}\,\mathrm{ave}[(f - g)W_2]\}^{1/2}$. Hence Pisier's (1983) maximal inequality yields

$$\mathbb{E} \sup_{f, g \in \mathcal{F}: \rho(f,g) \leq \epsilon(n)} \Big| n^{1/2}\,\mathrm{ave}((f - g)W_2) \Big| \;\to_p\; 0 \quad \text{whenever } \epsilon(n) \downarrow 0,$$

and (6.6) follows from $\rho(\widehat{f}, \widetilde{f}) = o_p(1)$, again a consequence of (6.5).

Because of expansion (6.4) it suffices to show that the distribution of

$$\widehat{d}(\widetilde{f}) \;=\; n^{1/2}\,\mathrm{ave}[(2\widetilde{f} - 1)W_1 + 2(\widetilde{f} - 1)W_2 + (1 - 2\widetilde{f})V]$$

is asymptotically normal with mean zero and variance τ^2. By assumption (3.2), (W_1, W_2) and V are independent with $m\Big(\mathcal{L}(n^{1/2}V), \mathcal{N}(0, \beta^2\sigma^4)\Big) \to 0$. Further, Lindeberg's Central Limit Theorem entails

$$m\Big(\mathcal{L}\Big(n^{1/2}\,\mathrm{ave}[(2\widetilde{f} - 1)W_1]\Big), \mathcal{N}\Big(0, 2\sigma^4\,\mathrm{ave}[(2\widetilde{f} - 1)^2]\Big)\Big) \;\to\; 0,$$

whereas $\mathcal{L}\Big(n^{1/2}\,\mathrm{ave}[(\widetilde{f} - 1)W_2]\Big) = \mathcal{N}\Big(0, \sigma^2\,\mathrm{ave}[\xi^2 (1 - \widetilde{f})^2]\Big)$. Thus it remains to be shown that $n^{1/2}\,\mathrm{ave}[(2\widetilde{f} - 1)W_1)$ and $n^{1/2}\,\mathrm{ave}[(\widetilde{f} - 1)W_2]$ are asymptotically independent. For

that purpose we split T into the two subsets $T^{(1)} := \{\xi^2 \le n^{1/2}\}$ and $T^{(2)} := T \setminus T^{(1)}$. For any $a \in \mathbf{R}^T$ let $a^{(k)}(t) := 1\{t \in T^{(k)}\}a(t)$. Note that $(W_1^{(1)}, W_2^{(1)})$ and $(W_1^{(2)}, W_2^{(2)})$ are independent. In addition,

$$n\,\mathbb{E}\,\mathrm{ave}[(2\widetilde{f}-1)W_1^{(2)}]^2 \;\le\; 2\sigma^4 \#T^{(2)}/n \;\le\; 2\sigma^4\,\mathrm{ave}(\xi^2)/n^{1/2} \;=\; o(1),$$

while $n^{1/2}\,\mathrm{ave}[(2\widetilde{f}-1)W_1^{(1)}]$ and $n^{1/2}\,\mathrm{ave}[(\widetilde{f}-1)W_2^{(1)}]$ are asymptotically independent, according to the bivariate CLT. □

Proof of Theorem 3.2. It follows from Theorem 3.1 and the proof of Theorem 2.1 that

$$\begin{aligned} \hat{r}^2 \;=\; \widehat{R}_C(\widehat{f}) + O_p(n^{-1/2}) \;&=\; R(\widetilde{f},\xi,\sigma^2) + O_p(n^{-1/2}) \\ \text{and} \quad L(\widehat{f}X,\xi) \;&=\; R(\widetilde{f},\xi,\sigma^2) + O_p(n^{-1/2}). \end{aligned}$$

Consequently,

$$L(\widehat{C},\xi) \;=\; [L(\widehat{f}X,\xi)^{1/2} + \hat{r}]^2 \;=\; 4R(\widetilde{f},\xi,\sigma^2) + O_p(n^{-1/2}).$$

As for the consistency of $\hat{\tau}^2$, we know already that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$ and $\mathrm{ave}[(\widehat{f}-\widetilde{f})^2] = o_p(1)$. Thus the assertion follows from Theorem 2.1 via

$$\begin{aligned} \mathrm{ave}[(X^2-\hat{\sigma}^2)(1-\widehat{f})^2] \;&=\; \widehat{R}_C(\widehat{f}) - \hat{\sigma}^2\,\mathrm{ave}(\widehat{f}^2) \\ &=\; R(\widetilde{f},\xi,\sigma^2) - \sigma^2\,\mathrm{ave}(\widetilde{f}^2) + o_p(1) \\ &=\; \mathrm{ave}[\xi^2(1-\widetilde{f})^2] + o_p(1). \end{aligned}$$

Given consistency of $\hat{\tau}^2$, the fact that (3.4) implies asymptotic level α of $\widehat{C}$ follows from standard arguments. It remains to be shown that (3.5) implies (3.4). Suppose that the latter condition is violated. This is equivalent to

$$\liminf_{n\to\infty} \inf_{\mathrm{ave}(\xi^2)\le c} \Big(\mathrm{ave}[(\widetilde{f}-1/2)^2] + \mathrm{ave}[\xi^2(1-\widetilde{f})^2]\Big) \;=\; 0. \tag{6.7}$$

However,

$$R(\widetilde{f},\xi,\sigma^2) \;\ge\; \sigma^2\,\mathrm{ave}(\widetilde{f}^2) \;\ge\; \sigma^2/4 - \mathrm{ave}[(\widetilde{f}-1/2)^2]^{1/2}$$

while for $f_o := 1\{\widetilde{f} \geq 3/4\}\widetilde{f} \in \mathcal{F}$,

$$\begin{aligned} R(\widetilde{f},\xi,\sigma^2) &\leq R(f_o,\xi,\sigma^2) \\ &= \operatorname{ave}[\xi^2(1-f_o)^2] + \sigma^2 \operatorname{ave}(f_o^2) \\ &\leq 16 \operatorname{ave}[\xi^2(1-\widetilde{f})^2] + 9\sigma^2 \operatorname{ave}[(\widetilde{f}-1/2)^2]. \end{aligned}$$

These two inequalities are incompatible with (6.7). □

Proof of Theorem 3.3. It follows from the fact that $\widehat{\sigma}^2 = \sigma^2 + o_p(1)$ and Theorem 3.1, applied to $\mathcal{L}(X^*, \widehat{\sigma}^{2*} \mid X, \widehat{\sigma}^2)$ in place of $\mathcal{L}(X, \widehat{\sigma}^2)$, that $m[\widehat{H}, \mathcal{N}(0,\widehat{\tau}^2)] = o_p(1)$, where $\widehat{\tau}^2$ is defined as τ^2 with $(\widehat{\xi}, \widehat{\sigma}^2, \widehat{f})$ in place of $(\xi, \sigma^2, \widetilde{f})$. One easily checks that $\widehat{\tau}^2 = \tau^2 + o_p(1) = O_p(1)$, whence

$$m[\widehat{H}, H] \;\leq\; m[\widehat{H}, \mathcal{N}(0,\widehat{\tau}^2)] + m[\mathcal{N}(0,\widehat{\tau}^2), \mathcal{N}(0,\tau^2)] + m[H, \mathcal{N}(0,\tau^2)] \;=\; o_p(1).$$

Now suppose that $\mathcal{F}$ has the special properties described in Theorem 3.3. If $\widehat{\sigma}^2 > 0$, then $\widehat{f} < 1$, so that $\widehat{\xi}$ is well-defined. Condition 3.6 follows from

$$\operatorname*{arg\,min}_{g \in [0,1]^T} R(g, \widehat{\xi}, \widehat{\sigma}^2) \;=\; \widehat{\xi}^2/(\widehat{\sigma}^2 + \widehat{\xi}^2) \;=\; \widehat{f}.$$

The special construction of $\widehat{\xi}$ entails that

$$\widehat{\xi}^2 \;=\; \sum_{c \in \widehat{f}(T)} 1\{\widehat{f} = c\} \operatorname{ave}[1\{\widehat{f} = c\}(X^2 - \widehat{\sigma}^2)]^+ / \operatorname{ave}(1\{\widehat{f} = c\}).$$

Consequently the moment condition (3.7) follows from

$$\operatorname{ave}(\widehat{\xi}^2) \;=\; \sum_{c \in \widehat{f}(T)} \operatorname{ave}[1\{\widehat{f} = c\}(X^2 - \widehat{\sigma}^2)]^+ \;\leq\; \operatorname{ave}(X^2) \;=\; O_p(1),$$

while (3.8) follows from

$$\begin{aligned} \operatorname{ave}[\widehat{\xi}^2(1-\widehat{f})^2] &= \operatorname{ave}[\widehat{\sigma}^2 \widehat{f}(1-\widehat{f})] \\ &= \operatorname{ave}[\sigma^2 \widetilde{f}(1-\widetilde{f})] + o_p(1) \\ &= \operatorname{ave}[\sigma^2 \widetilde{f}/(1-\widetilde{f})\,(1-\widetilde{f})^2] + o_p(1) \\ &= \sum_{c \in \widetilde{f}(T)} \operatorname{ave}\Big[1\{\widetilde{f} = c\} \operatorname{ave}(1\{\widetilde{f} = c\}\xi^2)/\operatorname{ave}(1\{\widetilde{f} = c\})\,(1-\widetilde{f})^2\Big] + o_p(1) \\ &= \sum_{c \in \widetilde{f}(T)} \operatorname{ave}[1\{\widetilde{f} = c\}\xi^2(1-\widetilde{f})^2] + o_p(1) \\ &= \operatorname{ave}[\xi^2(1-\widetilde{f})^2] + o_p(1). \qquad \square \end{aligned}$$

References

Alexander, K.S. (1987). Central limit theorems for stochastic processes under random entropy conditions. *Probab. Theory Rel. Fields* **75**, 351–378.

Beran, R. (1994). Stein confidence sets and the bootstrap. *Statistica Sinica* **5**, 109–127.

Beran, R. (1996a). Confidence sets centered at C_p-estimators. *Ann. Inst. Statist. Math.* **48**, 1–15.

Beran, R. (1996b). Stein estimation in high dimensions: a retrospective. *Festschrift in Honor of Madan L. Puri* (E. Brunner and M. Denker, eds.) 91–110. VSP, Zeist.

Casella, G.and J.T. Hwang (1982). Limit expressions for the risk of James-Stein estimators. *Canad. J. Statist.* **10**, 305–309.

Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Dudley, R.M. (1987). Universal Donsker classes and metric entropy. *Ann. Prob.* **15**, 1306–1326.

Engel, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **49**, 242–254.

Gasser, T., L. Sroka and C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Golubev, G.K. and Nussbaum, M. (1992). Adaptive spline estimates in a nonparametric regression model. *Theory Probab. Appl.* **37**, 521–529.

James, W. and C. Stein (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** (J. Neyman, ed.), 361–380. University of California Press, Berkeley.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835–866.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958–976.

Li, K.-C. (1989). Honest confidence sets for nonparametric regression. *Ann. Statist.* **17**, 1001–1008.

Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Nussbaum, M. (1996). The Pinsker bound: a review. Unpublished preprint.

Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16**, 120–133.

Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Banach Spaces, Harmonic Analysis, and Probability Theory* (R.C. Blei and S.J. Sidney, eds.). Lect. Notes Math. **995**, 123–154. Springer, Berlin.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conf. Ser. Prob. Statist. **2**. IMS, Hayward.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–1230.

Robertson, T., F.T. Wright and R.L. Dykstra (1988). *Order Restricted Statistical Inference.* Wiley, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* (J. Neyman, ed.), 197–206. University of California Press, Berkeley.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. *Research Papers in Statistics. Festschrift for Jerzy Neyman* (F.N. David, ed.), 351–366. Wiley, London.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151.

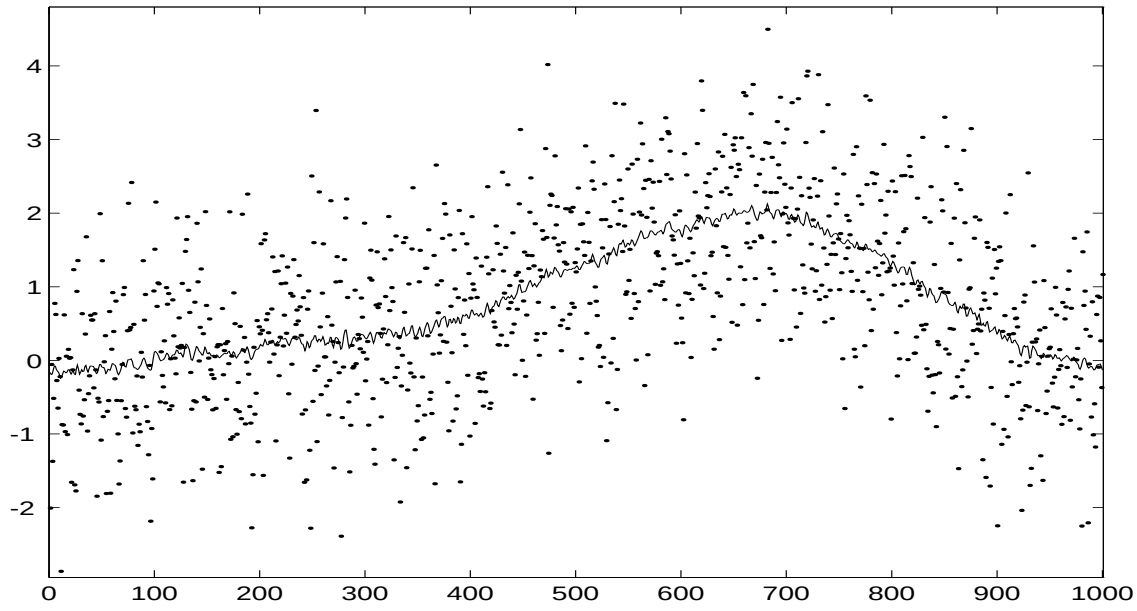

Figure 1a. Y and $\widehat{\mu}(\cdot/n)$.

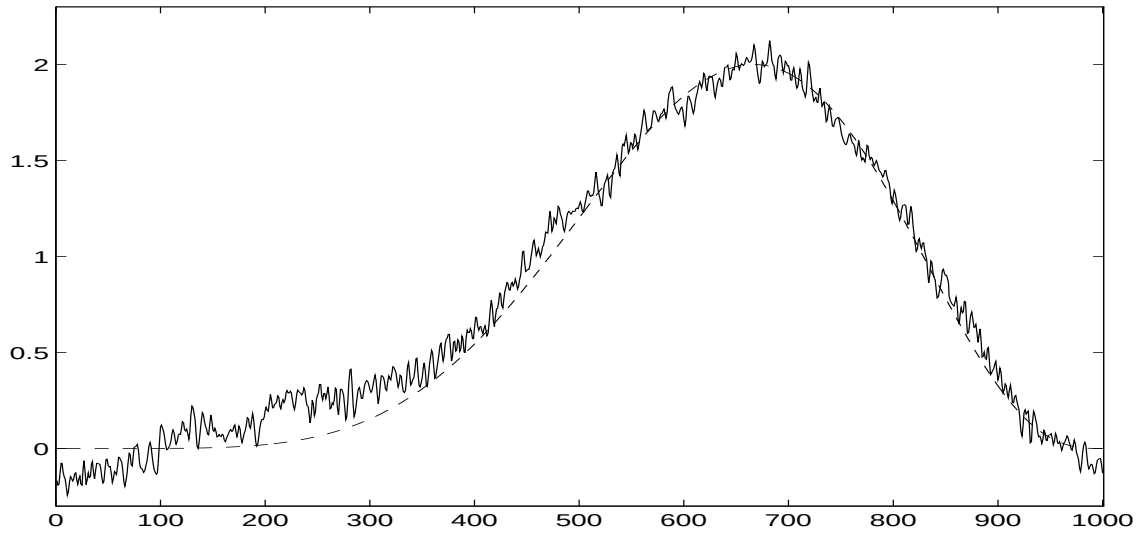

Figure 1b. Estimator $\widehat{\mu}(\cdot/n)$ (line) and true function $\mu(\cdot/n)$ (broken line).

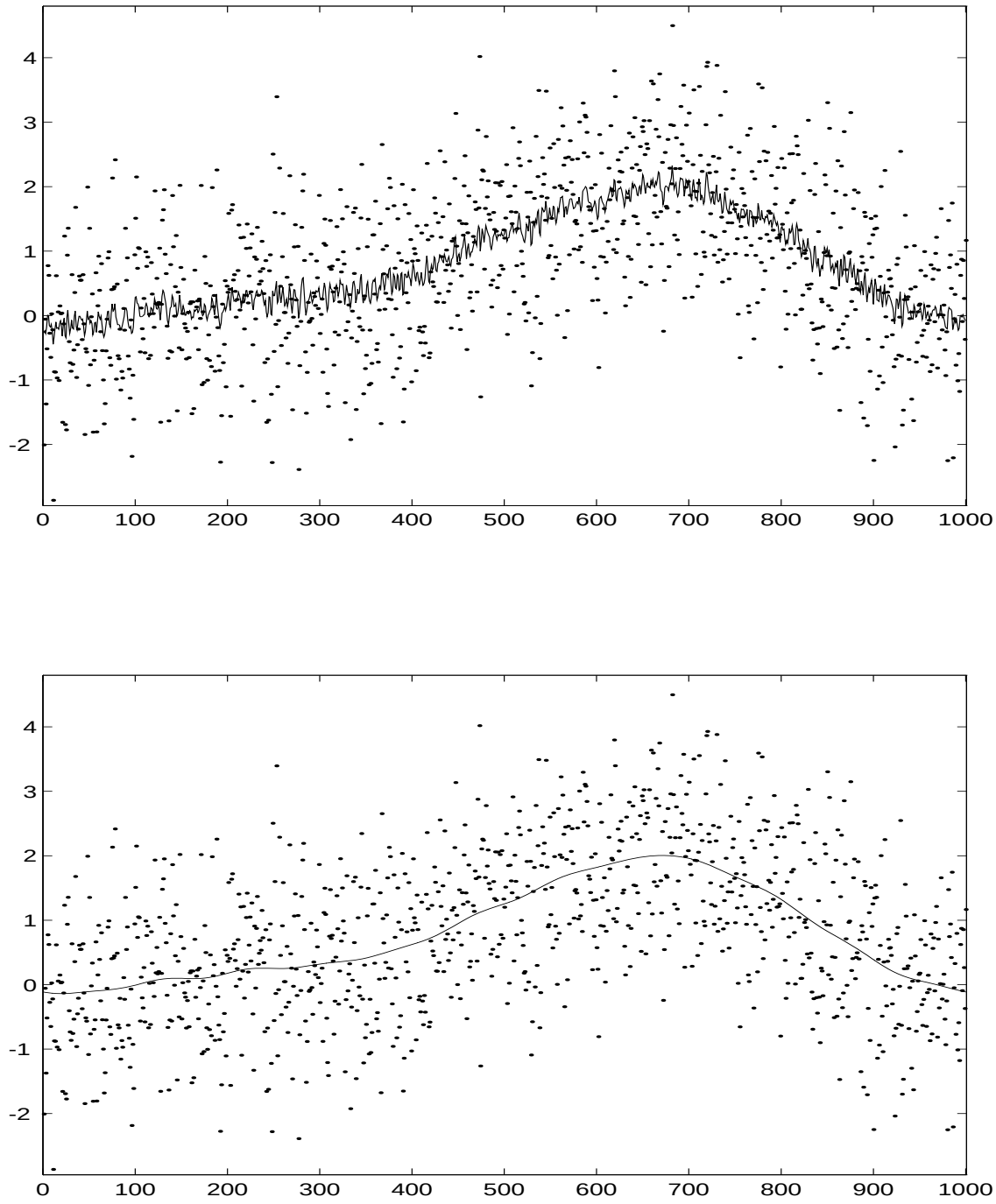

Figure 1a'. Y and $\widehat{\mu}(\cdot/n)$ with $\widehat{\sigma} = 0.9$ (1st plot) and $\widehat{\sigma} = 1.1$ (2nd plot).

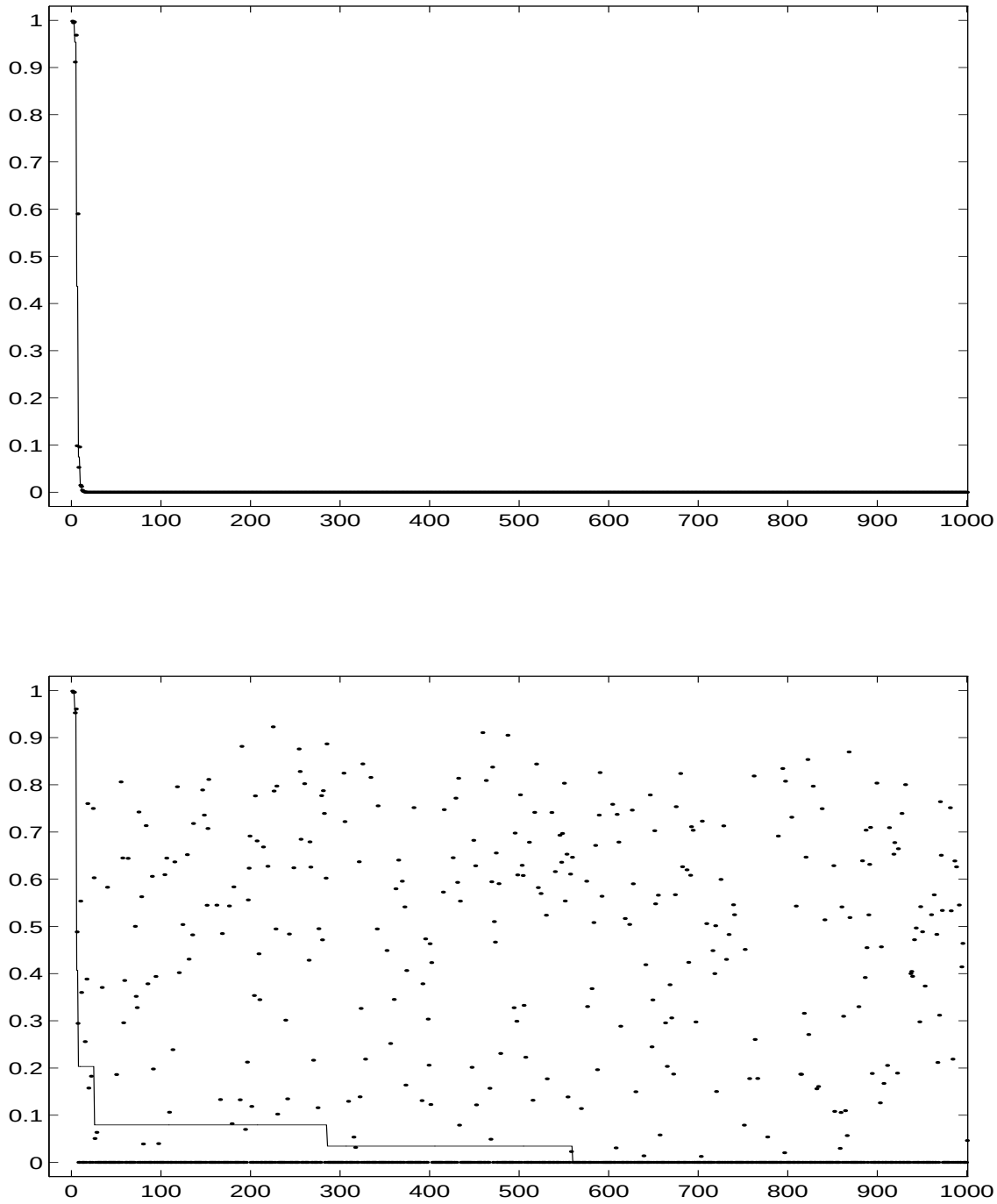

Figure 1c. Greedy and monotone modulators: $\widetilde{g}$ and $\widetilde{f}$ (1st plot); $\widehat{g}^{+}$ and $\widehat{f}$ (2nd plot).

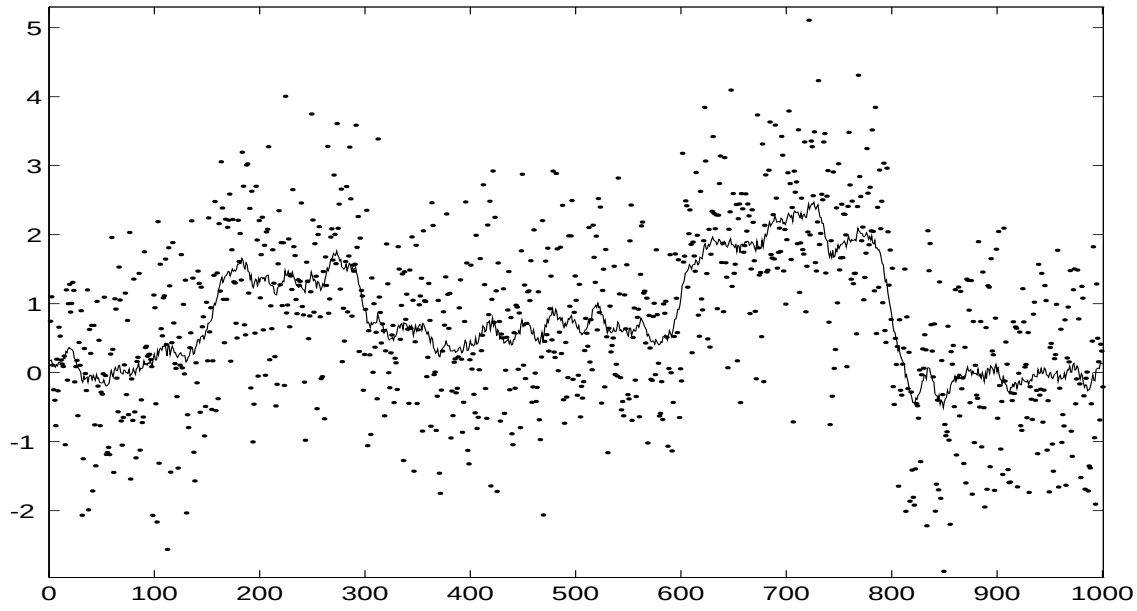

Figure 2a. Y and $\widehat{\mu}(\cdot/n)$.

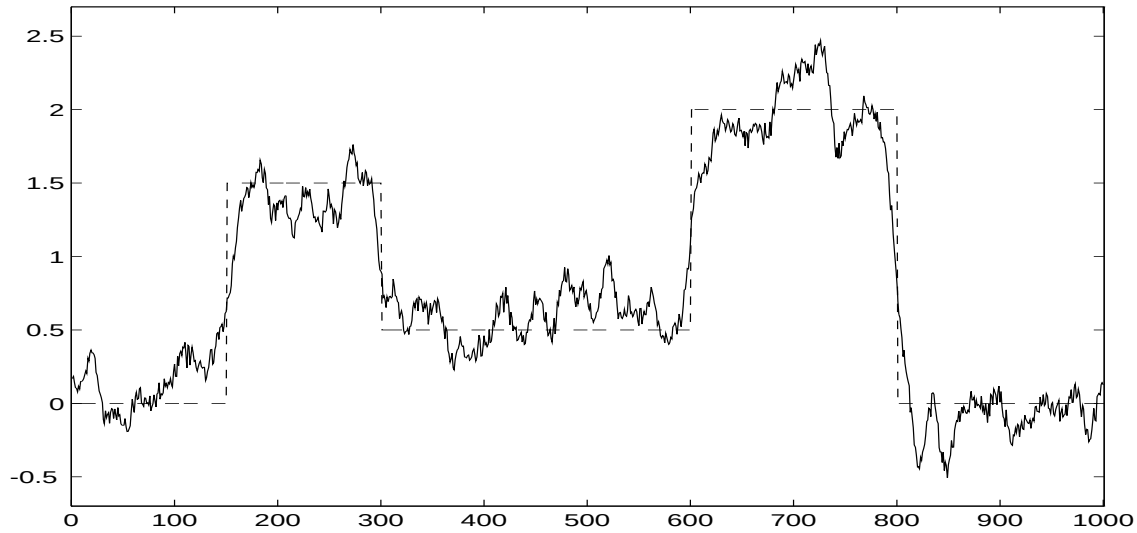

Figure 2b. Estimator $\widehat{\mu}(\cdot/n)$ (line) and true function $\mu(\cdot/n)$ (broken line).

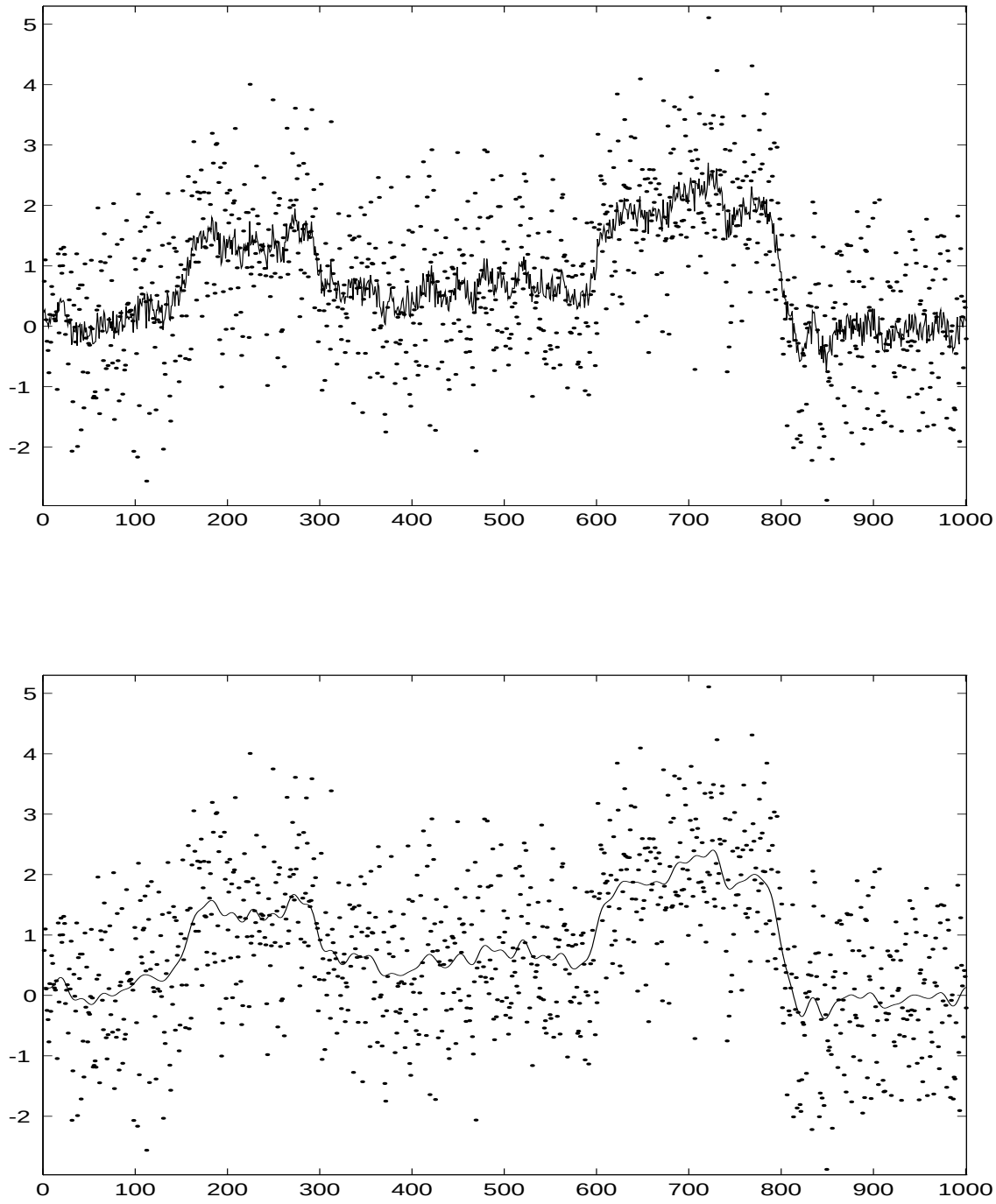

Figure 2a'. Y and $\widehat{\mu}(\cdot/n)$ with $\widehat{\sigma} = 0.9$ (1st plot) and $\widehat{\sigma} = 1.1$ (2nd plot).

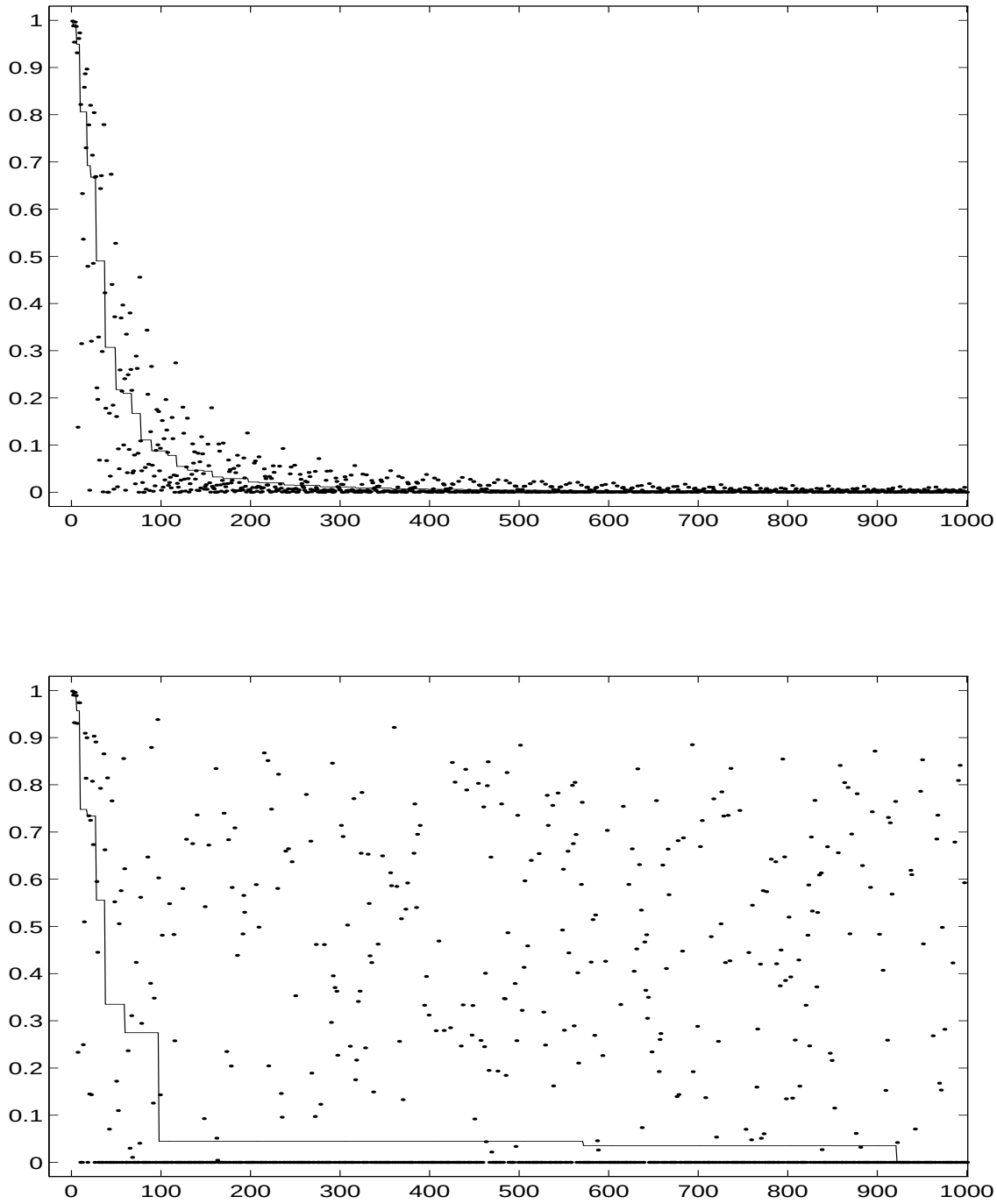

Figure 2c. Greedy and monotone modulators: $\widetilde{g}$ and $\widetilde{f}$ (1st plot); $\widehat{g}^{+}$ and $\widehat{f}$ (2nd plot).

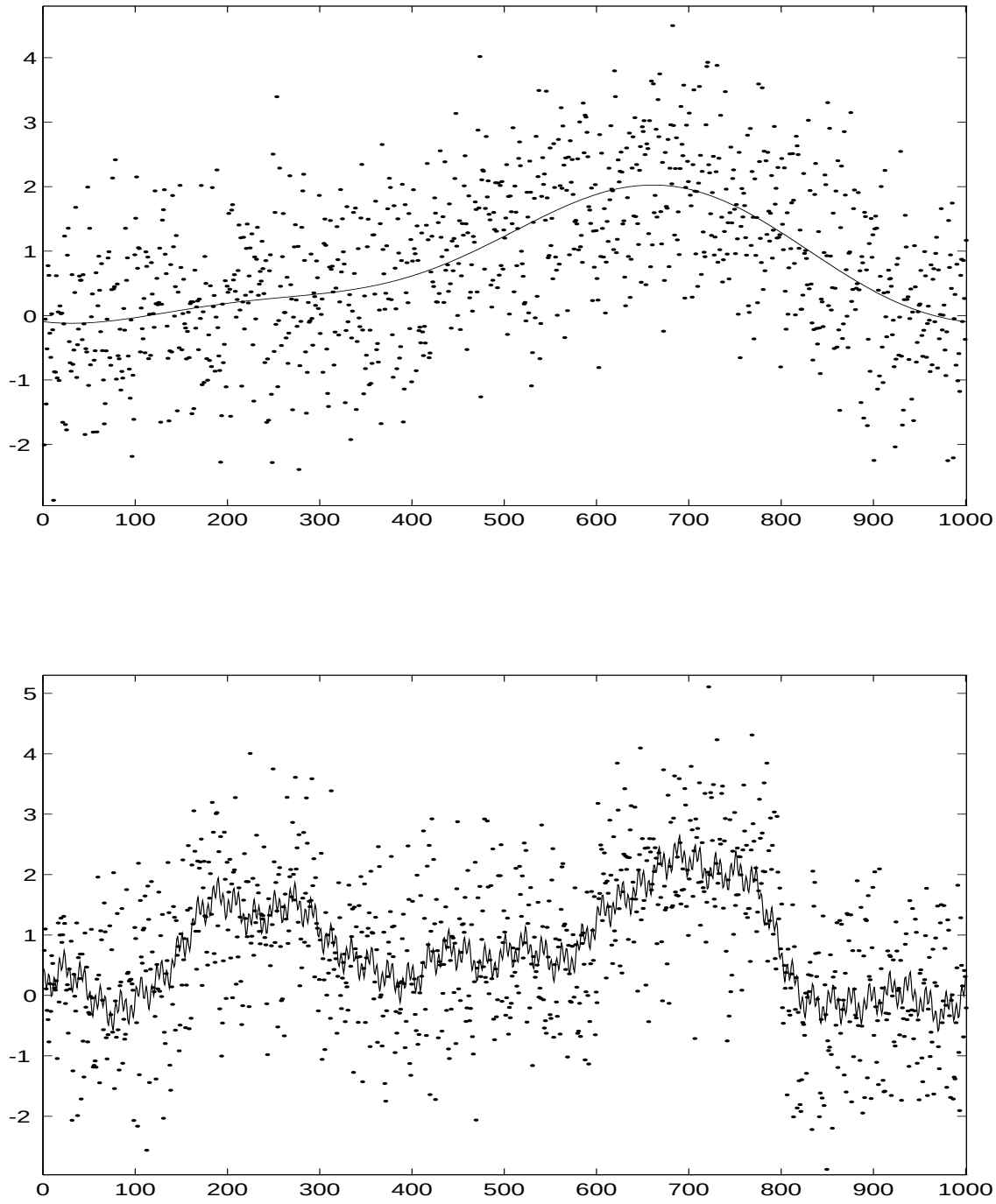

Figure 3. Oracle threshold estimator $\widetilde{\mu}_{\mathrm{th}}(\cdot/n)$ for Case 1 (1st plot) and Case 2 (2nd plot).