On asymptotic expansion of pseudovalues in nonparametric median regression

Eduard Belitser

Summary. We consider the median regression model $X_k = \theta(x_k) + \xi_k$, where the unknown signal $\theta : [0,1] \to \mathbb{R}$, is assumed to belong to a Hölder smoothness class, the ξ_k's are independent, but not necessarily identically distributed, noises with zero median. The distribution of the noise is assumed to be unknown and satisfying some weak conditions. Possible noise distributions may have heavy tails, so that, for example, no moments of the noises exist. Therefore, traditional estimation methods (for example, kernel methods) can not be applied directly in this situation.

On the basis of a preliminary recursive estimator, we construct certain variables Y_k's, called pseudovalues which do not depend on the noise distribution, and derive an asymptotic expansion (uniform over certain class of noise distributions): $Y_k = \theta(x_k) + \epsilon_k + r_k$, where ϵ_k's are binomial random variables and the rest terms r_k's are "small". This expansion mimics the nonparametric regression model with binomial noises. In so doing, we reduce our original observation model with "bad" (heavy-tailed) noises effectively to the nonparametric regression model with binomial noises.

1 Introduction

Suppose we want to recover a smooth signal $\theta(\cdot)$ on the basis of the observations

$$X_k = \theta(x_k) + \xi_k, \qquad k = 0, 1, \ldots, n\,, \tag{1}$$

where $\{x_k\} = \{x_{k,n}\} \subset [0,1] = I$ is a design to be specified later, $\theta(x) \in \mathcal{H}_\alpha(H', L')$, $x \in I$, with nonparametric class $\mathcal{H}_\alpha(H', L')$ which we define below.

The noises ξ_k's are assumed to be independent, each ξ_k is absolutely continuous with density f_k, all the f_k's are unknown. By $f_\xi(u) = \prod_{k=0}^n f_k(u_k)$, $u = (u_0, \ldots, u_n)$, we denote

AMS 1991 subject classifications. Primary: 62G20, 62G35; secondary: 62L20.

Key words and phrases: asymptotic expansion, nonparametric median, pseudovalues, recursive algorithm.

the joint distribution of the noise vector $(\xi_0, \xi_1, \ldots, \xi_n)$. Assume that $f_\xi \in \mathcal{P}_\xi$, where

$$\begin{aligned}\mathcal{P}_\xi &= \mathcal{P}_\xi(\delta, p, M_\xi, L_\xi) \\ &= \Big\{ \prod_{k=0}^{n} f_k(u_k) : \ f_0(0) = \ldots = f_n(0), \ f_k \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M_\xi) \cap \mathcal{P}_3(L_\xi) \Big\} \ ,\end{aligned}$$

where the sets $\mathcal{P}_i$'s are defined for positive constants δ, p, M_ξ and L_ξ as follows:

P1. $\mathcal{P}_1 = \Big\{ f(\cdot) : \ f \text{ is a density}, \ F(0) = \int_{-\infty}^{0} f(z)dz = 1/2 \Big\}$, i.e. the distribution of each ξ_k has zero median;

P2. $\mathcal{P}_2 = \mathcal{P}_2(\delta, p, M_\xi) = \big\{ f(\cdot) : \ 0 < p \le \inf_{|x| \le \delta} f(x) \le \sup_{|x| \le \delta} f(x) \le M_\xi \big\}$;

P3. $\mathcal{P}_3 = \mathcal{P}_3(L_\xi) = \big\{ f(\cdot) : \ |f(u_1) - f(u_2)| \le L_\xi |u_1 - u_2| \, , \ u_1, u_2 \in \mathbb{R} \big\}$.

The case when the ξ_k's are independent identically distributed random variables such that $f_0 \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M_\xi) \cap \mathcal{P}_3(L_\xi)$ certainly fits in this framework.

The function value $\theta(x_k)$ has a meaning of conditional median. The model (1) can therefore be called nonparametric median regression. Note further that the noise distribution from $\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3$ may have heavy tails, (say, Cauchy distribution) so that, for instance, the expectation of noises does not exist. This implies also that in general linear methods (for instance, kernel methods) can not be applied directly for estimating the signal θ in the situations when we measure the quality of estimator $\hat{\theta}_n(x)$ by a risk function of the form $E|\hat{\theta}_n(x) - \theta(x)|^\kappa$, $\kappa > 0$.

The design $\{x_k\}$ is assumed to satisfy the following conditions:

D1. $0 = x_0 \le x_1 \le \ldots \le x_n = 1$;

D2. $|x_l - x_m| \le D|l - m|/n$ for all $0 \le l, m \le n$ and some fixed positive constant D.

In fact, the above conditions represent the requirement for the design $\{x_k\}$ to have the same properties as the equidistant design.

Suppose now that the unknown function $\theta(x)$ on the interval $[0, 1]$ belongs to the Hölder function class $\mathcal{H}_\alpha = \mathcal{H}_\alpha(H', L')$ of the smoothness α: for some positive H' and L',

$$\begin{aligned}\mathcal{H}_\alpha(H', L') = \Big\{ \theta : \ |\theta^{(m)}(0)| \le H' \, , \ m = 0, \ldots, r \, ; \\ |\theta^{(r)}(u) - \theta^{(r)}(v)| \le L'|u - v|^{\beta_0}, \ u, v \in [0, 1] \Big\} \, ,\end{aligned}$$

where $r \in \mathbb{N}$, $0 < \beta_0 \le 1$, $\alpha = r + \beta_0$. Here $\theta^{(r)}(u)$ denotes the r-th derivative of $\theta(u)$.

The parameter n stands for the frequency of observations, i.e. a number of observations per unit interval. We study the estimation problem in asymptotic setup when this parameter tends to infinity. Note that in fact we deal with the sequence of models $X_{k,n} = \theta(x_{k,n}) + \xi_{k,n}$, $k = 1, \ldots, n$. To ease the notations, we omit the subscript n, for instance $X_k = X_{k,n}$,

$\xi_k = \xi_{k,n}$ etc. Related estimation problems have been studied by Tsybakov [8], Korostelev [5], Truong [7], Belitser and Korostelev [1], Belitser and van de Geer [2].

Belitser and Korostelev, [1], considered the problem of pointwise signal estimation and estimation of a smooth functional of the signal; the noises are iid with zero median and possibly heavy tails. In that paper, on the basis of a consistent recursive estimator special random variables, called pseudovalues, are introduced, for which an asymptotic expansion was derived. By using this expansion, a minimax estimator for the signal and an efficient estimator for the smooth functional of the signal were constructed. One of the main assumptions from the above paper, used in all the constructions, is that the value of noise density at point zero (median of noises) is fixed and known, which is rather restrictive.

In Belitser and van de Geer, [2], the result on convergence properties of recursive estimator is improved in several respects (among others almost sure convergence rate is derived). Using this recursive estimator, we construct a version of pseudovalues Y_ks, adaptive with respect to noise distribution, and derive a robust (uniform over certain class of noise distributions) stochastic asymptotic expansion for these pseudovalues: $Y_k = \theta(x_k) + \epsilon_k + r_k$, where ϵ_k's are binomial random variables and the rest terms r_k's are small, in a certain sense. This expansion mimics the nonparametric regression model with binomial noises. In so doing, we reduce our original observation model with "bad" (heavy-tailed) noises effectively to the nonparametric regression model with "nice" (binomial) noises. This expansion reduces the original observation model with heavy noises effectively to the nonparametric regression model with binomial noises. We also elucidate how one can utilize this expansion, when constructing minimax estimator: we consider the example of kernel estimator based on the pseudovalues.

2 A preliminary recursive estimator

Introduce the Lipschitz function class $\Theta_\beta = \Theta_\beta(H, L)$ with the smoothness β. For some positive H, L and $0 < \beta \le 1$ define

$$\Theta_\beta(H, L) = \left\{\theta : \ |\theta(0)| \le H\,,\ |\theta(u) - \theta(v)| \le L|u - v|^\beta,\ u, v \in [0, 1]\right\}.$$

It is easy to see that if $\theta \in \mathcal{H}_\alpha$, then, of course $\theta \in \Theta_\beta$ with $\beta = \min\{\alpha, 1\}$ for the appropriate choice of constants H and L in the definition of Θ_β, i.e.

$$\mathcal{H}_\alpha(H', L') \subset \Theta_\beta(H, L)\,.$$

Therefore, if some property holds uniformly over $\theta \in \Theta_\beta(H, L)$, then certainly the same property holds uniformly over $\theta \in \mathcal{H}_\alpha(H', L')$ as well.

Now we introduce a preliminary recursive estimator, whose convergence properties were studied by Belitser and Korostelev, [1], and by Belitser and van de Geer, [2].

First some more notations. Let $I\{S\}$ denote the indicator function of the set S and by c and C generic constants which may be different in different expressions. Define the functions $\text{sign}(u) = I\{u \geq 0\} - I\{u < 0\}$ and

$$S(u,v) = \begin{cases} \text{sign}(u-v), & |v| \leq M \\ -v, & |v| > M \end{cases}$$

for some fixed $M > H + L$, and the sequence $\gamma_n = n^{-2\beta/(2\beta+1)} \log n$, where the constants β, H and L appear in the definition of the class Θ_β. For brevity, denote also $\theta_k = \theta(x_k)$.

The following recursive formula gives an estimator for the function value θ_k:

$$\hat\theta_{k+1} = \hat\theta_k + \gamma_n S(X_k, \hat\theta_k), \qquad k = 0, 1, \ldots, n-1\,, \tag{2}$$

with the initial value $\hat\theta_0 = 0$. We will use the uniform (over the class $\mathcal{P}_1 \cap \mathcal{P}_2$ of the noise distributions) version of the result from Belitser and van de Geer, [2], which describes the mean square convergence rate of the above recursive estimator.

Theorem 1 *Let*

$$K_n = \{k \in \mathbb{N} :\, C_0 n^{2\beta/(2\beta+1)} \leq k \leq n\}\,,$$

where $C_0 = 2/\min\{2p, 2p\delta/(M+H+L), 1/2\}$. *Then, for some positive constant* C_1, *the relation*

$$\limsup_{n\to\infty} \max_{k\in K_n} \frac{n^{2\beta/(2\beta+1)}}{(\log n)^2} E_\theta(\hat\theta_k - \theta_k)^2 \leq C_1 \tag{3}$$

holds uniformly over $\theta \in \Theta_\beta$ *and* $f_\xi \in \left\{ \prod_{k=0}^n f_k(u_k) :\, f_k \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M_\xi) \right\}$.

The proof of the above theorem is exactly the same as the proof of the corresponding result from Belitser and van de Geer, [2], and is therefore omitted, see also Remark 7 in that paper about the uniformity over over $f_\xi \in \left\{ \prod_{k=0}^n f_k(u_k) :\, f_k \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M) \right\}$. In this paper the rate of convergence for the above recursive algorithm is derived also in almost sure sense, but we do not utilize this here.

If we interpret index k as time moment then the above estimating procedure is a recursive stochastic algorithm. Such estimators are most appropriate in the situations when observations appear successively so that, when constructing an estimator at a fixed time moment k, we can use only those observations which have been obtained up to this moment, i.e. only X_i's with $i \leq k$.

Since we start our algorithm with zero value $\hat\theta_0 = 0$ which need not be equal to the true value θ_0, the estimator $\hat\theta_k$ can not be consistent for all values θ_k, $0 \leq k \leq n$. There is a so called "burn-in" part $k = 0, 1, \ldots, \lfloor C_0 n^{2\beta/(2\beta+1)} \rfloor$. We can handle this problem by running the algorithm in the opposite direction.

In terms of signal estimation we have the following implication of the above theorem and the Lipschitz condition on functions from the class Θ_β. Let $\hat\theta_n(u)$ be a piecewise constant

continuation of $\hat{\theta}_k = \hat{\theta}(x_k)$, $k = 0, 1, \ldots, n$, i.e. $\hat{\theta}_n(u) = \hat{\theta}_k$ for $x_k \le u < x_{k+1}$, $i = 0, 1, \ldots, n-1$, and $\hat{\theta}_n(1) = \hat{\theta}_n$. Let also

$$T_n = \{x : \ C_0 n^{-1/(2\beta+1)} \le x \le 1\}\,.$$

Then, for some positive constant C_1 the relation

$$\limsup_{n\to\infty} \max_{u\in T_n} \frac{n^{2\beta/(2\beta+1)}}{(\log n)^2} E_\theta(\hat{\theta}_n(u) - \theta(u))^2 \le C_1$$

hold uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \left\{ \prod_{k=0}^n f_k(u_k) \ : f_k \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M) \right\}$. Notice that, for any $x \in (0, 1]$, $x \in T_n$ for sufficiently large n.

3 Pseudovalues

In this section we introduce special statistics called pseudovalues, a nonparametric analogue of Tukey's pseudovalues. The idea of introducing such random variables is due to Korostelev, [5], in a different estimation problem, who considered these random variables as a nonparametric counterpart of Tukey's pseudovalues. We derive astochastic asymtotic expansion for these pseudovalues which enables us to construct robust minimax estimators. First we define robust minimax estimators.

The following proposition is given for illustrative purposes. It says essentially what estimation quality can in principle be achieved for the class $\mathcal{H}_\alpha$ in terms of convergence rate. We consider the pointwise minimax risk of the form (convenient for our purposes)

$$r_n(\mathcal{H}_\alpha) = \inf_{\hat{\theta}_n} \sup_{\theta\in\mathcal{H}_\alpha} E_\theta |\hat{\theta}_n(x) - \theta(x)| \tag{4}$$

where the infimum is taken over all possible estimators and the supremum over all curves from the nonparametric class $\mathcal{H}_\alpha$. This quantity reflects in a way the difficulty of the estimation problem over the class $\mathcal{H}_\alpha$. Then the following lower bound for the minimax risk holds.

Proposition 1 *Let the distributions of ξ's be standard Gaussian. For any fixed $x \in [0, 1]$ there exists a positive constant c such that*

$$\liminf_{n\to\infty} n^{\alpha/(2\alpha+1)} r_n(\mathcal{H}_\alpha) \ge c\,,$$

where the minimax risk $r_n(\mathcal{H}_\alpha)$ is defined by (4).

The proof is omitted since this is a folklore result (by now) originating from Ibragimov and Hasminskii, [4], and Stone, [6].

If the distributions of ξ's are standard Gaussian, then the corresponding $f_\xi \in \mathcal{P}_\xi$ for the appropriate choice of the constants in the definition of class $\mathcal{P}_\xi$. This implies trivially the lower bound for the supremum of the minimax risk $r_n(\mathcal{H}_\alpha)$ over all possible $f_\xi \mathcal{P}_\xi$:

$$\liminf_{n\to\infty} \sup_{f_\xi \in \mathcal{P}_\xi} n^{\alpha/(2\alpha+1)} r_n(\mathcal{H}_\alpha) \geq c\,.$$

An estimator $\tilde{\theta}_n$ is called *robust minimax* in point x if there exists a constant $C > 0$ such that

$$\limsup_{n\to\infty} \sup_{f_\xi \in \mathcal{P}_\xi} \sup_{\theta \in \mathcal{H}_\alpha} n^{\alpha/(2\alpha+1)} E_\theta |\tilde{\theta}_n(x) - \theta(x)| \leq C\,.$$

It is easy to see that if $\theta \in \mathcal{H}_\alpha$, then, of course $\theta \in \Theta_\beta$ with $\beta = \min\{\alpha, 1\}$ (we use this notation throughout this section). Therefore, according to the above results, our recursive estimator achieves the rate $n^{\beta/(2\beta+1)}$ up to a log factor. Comparing this with the lower bound, we find that our recursive estimator $\hat{\theta}$ never attains the optimal rate. The estimator is almost optimal for the class $\mathcal{H}_\alpha$ (in fact this estimator was originally designed for this class) with $0 < \alpha \leq 1$, i.e. $r = 0$ and $\alpha = \beta_0$. To be precise, it attains the optimal rate up to a log factor. Informally, one can regard the log factor as a price for recursiveness.

The difference with the optimal rate becomes significant when $\alpha > 1$, i.e. when at least one derivative exists. We are going to fix this shortcoming by introducing some statistics, which we call *pseudovalues*.

First introduce some notations:

$$A_k = A_{k,n} = \frac{n^{1/(2\beta+1)}}{2k} \sum_{i=0}^{k-1} I\Big\{|X_i - \hat{\theta}_i| \leq n^{-1/(2\beta+1)}\Big\}$$

for $k = 1, \ldots, n$. Define further

$$\hat{f}_0 = \hat{f}_{0,k} = \hat{f}_{0,k,n} = \begin{cases} A_k, & A_k \geq p \\ p, & A_k < p\,, \end{cases}$$

$$\varepsilon_k = \frac{\text{sign}(\xi_k)}{2f_0}\,, \tag{5}$$

$f_0 = f_0(0)$, $\mathcal{A}_k = \sigma(\xi_0, \ldots, \xi_k)$, the σ-algebra generated by $\xi_0, \ldots, \xi_k$, $k = 1, \ldots, n$. Recall that the constant p appears in the definition of $\mathcal{P}_\xi$ and f_0 is the density function of the noise variable ξ_0. Recall also that for $f_\xi \in \mathcal{P}_\xi$ $f_0(0) = \ldots = f_n(0)$.

Remark 1 As one can see, we need to know the constant p when constructing the estimator $\hat{f}_0$. Although this seems to be restrictive, we can assume this without loss of generality, because we can use a sequence p_n converging to zero sufficiently slowly instead of constant p. We will have only to modify slightly the proof.

Now define the pseudovalues

$$Y_k = \hat{\theta}_k + \frac{S(X_k, \hat{\theta}_k)}{2\hat{f}_{0,k}}\,, \quad k = 1, \ldots, n\,,$$

and the set

$$K_{\epsilon,n} = \{k \in \mathbb{N} : \ \epsilon \le k/n \le 1\}. \tag{6}$$

These statistics represent a nonparametric analogue of Tukey's pseudovalues; cf. Tukey (1958), [9], Efron, [3]. Indeed,

$$Y_k = N\hat{\theta}_{k+1} - (N-1)\hat{\theta}_k$$

with $N = \gamma_n^{-1}$. Here N is nonparametric analogue of n, the "effective" number of observations used in estimating the signal θ at point x.

Without loss of generality we can assume that n is large enough so that $K_{\epsilon,n} \subset K_n$. In the following theorem a stochastic expansion of the pseudovalues is given.

Theorem 2 *Let $\beta = \min\{\alpha, 1\}$ and the set $K_{\epsilon,n}$ be defined by (6). Then*

$$Y_k = \theta_k + \varepsilon_k + \mu_k + \nu_k\,, \quad k = 1, \ldots, n\,,$$

where ε_k's are defined by (5), $\{\mu_k\}$, $k = 1, \ldots, n$, forms a martingale difference with respect to the filtration $\{\mathcal{A}_k\}$, $\mathcal{A}_k = \sigma(\xi_0, \ldots, \xi_k)$, i.e. $E[\mu_k|\mathcal{A}_{k-1}] = 0$. Moreover, for any fixed $\epsilon > 0$, there exist some positive constants B_μ, B_ν such that

$$\limsup_{n\to\infty} \max_{k\in K_{\epsilon,n}} \frac{n^{\beta/(2\beta+1)}}{\log n} E_\theta \mu_k^2 \le B_\mu\,, \tag{7}$$

$$\limsup_{n\to\infty} \max_{k\in K_{\epsilon,n}} \frac{n^{2\beta/(2\beta+1)}}{(\log n)^2} E_\theta |\nu_k| \le B_\nu\,, \tag{8}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$.

Remark 2 Recall that for some H and L

$$\mathcal{H}_\alpha(H', L') \subset \Theta_\beta(H, L)\,,$$

with $\beta = \min\{\alpha, 1\}$. Therefore, the assertion in the above Theorem holds uniformly over $\mathcal{H}_\alpha$ as well.

Remark 3 All the results will still be valid if we require the Lipschitz condition in the definition of $\mathcal{P}_3$ only in a fixed neighborhood of zero instead of the whole real line $\mathbb{R}$.

The proof of the theorem is an immediate consequence of the following two lemmas. The proofs of these lemmas are somewhat lengthy, so we defer them to the next section.

Introduce the auxiliary variables

$$Y'_k = \hat{\theta}_k + \frac{S(X_k, \hat{\theta}_k)}{2f_0}\,, \quad k = 0, \ldots, n\,.$$

Lemma 1 *The following stochastic expansion holds:*

$$Y'_k = \theta_k + \varepsilon_k + \mu'_k + \nu'_k\,, \quad k = 1, \ldots, n\,,$$

where the sequences $\{\mu'_k\}$ *and* $\{\nu'_k\}$ *have the same properties as in Theorem 2.*

Remark 4 In fact, a slightly more general version of Lemma 1 holds: the set $K_{\epsilon,n}$ defined by (6) can be replaced by a bigger set K_n defined in Theorem 1.

Lemma 2 *The following stochastic expansion holds:*

$$Y_k = Y'_k + \mu''_k + \nu''_k\,, \quad k = 1, \ldots, n\,,$$

where the sequences $\{\mu''_k\}$ *and* $\{\nu''_k\}$ *have the same properties as in Theorem 2.*

Now we illustrate how one can utilize the stochastic expansion above. Suppose we had the observations

$$Z_k = \theta(x_k) + \varepsilon_k, \qquad k = \lfloor \epsilon n \rfloor, \ldots, n\,,$$

with ε_k defined by (5). This is a classical nonparametric regression model, where the noises ε_k's are independent binomial random variables taking just two values, each with probability $1/2$:

$$P\Big(\varepsilon_k = \frac{1}{2f_0}\Big) = P\Big(\varepsilon_k = -\frac{1}{2f_0}\Big) = \frac{1}{2}\,.$$

In this case some routine estimators, say kernel estimator with an appropriate kernel function, are known to attain the optimal convergence rate specified in the lower bound proposition. The benefit in constructing the pseudovalues Y_k's is that we can treat them as if they were Z_k's irrespective of the noise distribution $f_\xi \in \mathcal{P}_\xi$, i.e. the rest terms $\mu_k + \nu_k$ are negligible, beginning with a certain moment.

We elucidate this idea by heuristic arguments for a kernel estimator based on the pseudovalues. Suppose kernel $K(u)$ and bandwidth $h_n = cn^{-1/(2\alpha+1)}$ are chosen in such a way that for any fixed $x \in (\epsilon, 1]$ and some positive C,

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{H}_\alpha} n^{\alpha/(2\alpha+1)} E_\theta |\bar\theta_n(x) - \theta(x)| \le C\,.$$

where

$$\bar\theta_n(x) = \frac{1}{(n - \lfloor \epsilon n \rfloor)h_n} \sum_{k=\lfloor \epsilon n \rfloor + 1}^{n} K\Big(\frac{x - x_k}{h_n}\Big) Z_k\,.$$

However, we do not observe Z_k's, but we have pseudovalues Y_k's instead. So, we can construct estimator $\tilde\theta_n(x)$:

$$\tilde\theta_n(x) = \frac{1}{(n - \lfloor \epsilon n \rfloor)h_n} \sum_{k=\lfloor \epsilon n \rfloor + 1}^{n} K\Big(\frac{x - x_k}{h_n}\Big) Y_k = \bar\theta_n(x) + R_n + r_n\,.$$

Now estimate the expectations of the rest terms $|R_n|$ and $|r_n|$: under appropriate conditions on the kernel K ($\int K^2(u)du < \infty$),

$$\begin{aligned} E|R_n| &\le \big(ER_n^2\big)^{1/2} \\ &= \frac{1}{\big((n-\lfloor \epsilon n\rfloor)h_n\big)^{1/2}}\Big(\frac{1}{(n-\lfloor \epsilon n\rfloor)h_n}\sum_{k=\lfloor \epsilon n\rfloor+1}^{n} K^2\Big(\frac{x-x_k}{h_n}\Big)E\mu_k^2\Big)^{1/2} \\ &\le \frac{Cn^{-\beta/(2\beta+1)}}{\big((n-\lfloor \epsilon n\rfloor)h_n\big)^{1/2}}\Big(\int K^2(u)du\Big)^{1/2}\big(\log n\big)^{1/2} \\ &\le cn^{-\alpha/(2\alpha+1)}n^{-\beta/(2\beta+1)}\big(\log n\big)^{1/2} \end{aligned}$$

and similarly

$$E|r_n| \le \frac{1}{(n-\lfloor \epsilon n\rfloor)h_n}\sum_{k=\lfloor \epsilon n\rfloor+1}^{n}\Big|K\Big(\frac{x-x_k}{h_n}\Big)\Big|E|\nu_k| \le Cn^{-2\beta/(2\beta+1)}\big(\log n\big)^2$$

uniformly over $\theta \in \mathcal{H}_\alpha$ and $f_\xi \in \mathcal{P}_\xi$. Assume that $\beta > 1/2$, which is a typical condition in estimation problems for the Hölder class. Then $2\beta/(2\beta+1) > \alpha/(2\alpha+1)$. Together with all the above relations, this yields that for any fixed $x \in (\epsilon, 1]$ and some positive C,

$$\limsup_{n\to\infty}\sup_{f_\xi\in\mathcal{P}_\xi}\sup_{\theta\in\mathcal{H}_\alpha} n^{\alpha/(2\alpha+1)}E_\theta|\tilde\theta_n(x)-\theta(x)| \le C\,,$$

which means that the estimator $\tilde\theta_n$ is robust minimax at any point $x \in (\epsilon, 1]$.

Remark 5 Notice that the rest terms μ_k and ν_k in the stochastic expansion for the pseudovalues become small beginning with a later moment $\lfloor \epsilon n\rfloor$, compared with the recursive estimator $\hat\theta$. This is because we need to construct a recursive estimator of $f_0 = f_0(0)$. If, however, for some reason we need a consistent estimator in this "burn in" period $k = 0, 1, \dots, \lfloor n\epsilon\rfloor$, we can run the algorithm (2) in the opposite direction:

$$\hat\theta_k = \hat\theta_{k+1} + \gamma_n S(X_{k+1}, \hat\theta_{k+1}), \qquad k = n-1, \dots, 0\,,$$

with the initial value $\hat\theta_n = 0$. Then we define the statistics A_k in the corresponding manner

$$A_{n-k} = \frac{n^{1/(2\beta+1)}}{2k}\sum_{i=0}^{k-1} I\Big\{|X_{n-i} - \hat\theta_{n-i}| \le n^{-1/(2\beta+1)}\Big\}\,.$$

The results of the above theorem will hold with set $K'_{\epsilon,n} = \{k \in \mathbb{N} :\ 0 \le k/n \le 1-\epsilon\}$ instead of set $K_{\epsilon,n}$ and σ-algebra $\mathcal{A}'_k = \sigma\{\xi_n, \xi_{n-1}, \dots, \xi_{n-k}\}$ instead of $\mathcal{A}_k$.

Remark 6 The construction of the pseudovalues $\{Y_k\}$ is clearly adaptive with respect to the number of derivatives r. So, one can use an estimation method for the pseudovalues $\{Y_k\}$, adaptive to r. Of course, one has to make sure that the rest terms do not effect the rate of convergence, when applying such a method.

Remark 7 One can try to generalize the results for a multivariate function $\theta(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, from a multivariate Hölder function class. We will not consider this issue here, we just remark that in this case a certain condition on smoothness α and dimension d is needed to make sure that the approximation error (due to discretization $\{\theta(\mathbf{x}_k)\}$) and the rest terms do not effect the rate of convergence.

4 Proof of Lemmas

Proof of Lemma 1. First of all, notice that since $\theta \in \Theta_\beta$, we have by Theorem 1 that for the estimator $\hat\theta$ defined by (2) there exists some positive C such that

$$\limsup_{n\to\infty} \max_{k\in K_n} \frac{n^{2\beta/(2\beta+1)}}{(\log n)^2} E_\theta(\hat\theta_k - \theta_k)^2 \le C \tag{9}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$ because

$$\mathcal{P}_\xi \subset \Big\{ \prod_{k=0}^n f_k(u_k) : \ f_k \in \mathcal{P}_1 \cap \mathcal{P}_2(\delta, p, M_\xi) \Big\} .$$

Introduce the following random variables, for $k = 1, \dots, n$,

$$T_k = \frac{S(X_k, \hat\theta_k)}{2f_0} - \varepsilon_k \,,$$

$$\mu'_k = (T_k - E[T_k|\mathcal{A}_{k-1}]) I\{|\hat\theta_k| \le M\} \,,$$

$$\nu'_k = (T_k + \delta_k) I\{|\hat\theta_k| > M\} + (E[T_k|\mathcal{A}_{k-1}] + \delta_k) I\{|\hat\theta_k| \le M\} \,,$$

where $\delta_k = \hat\theta_k - \theta_k$. Trivially, $\{\mu'_k\}$, $k = 1, \dots, n$, is a martingale difference with respect to the filtration $\{\mathcal{A}_k\}$. We have obviously that, for $k = 1, \dots, n$,

$$Y'_k = \theta_k + \varepsilon_k + \mu'_k + \nu'_k \,.$$

Let us show that the random variables μ'_k's and ν'_k's satisfy (7) and (8) respectively.

Denote for brevity $I_k = I\{|\hat\theta_k| \le M\}$. Recall also that $f_k(0) = f_0(0) = f_0$, $k = 1, \dots, n$. Since

$$E[\text{sign}(\xi_k + \theta_k - \hat\theta_k) - \text{sign}(\xi_k)|\mathcal{A}_{k-1}] = 2 \int_{\hat\theta_k - \theta_k}^0 f_k(u) du \tag{10}$$

and, due to the fact that $f_k \in \mathcal{P}_3(L_\xi)$, $k = 0, \ldots, n$,

$$\begin{aligned} & E\Big[\{\text{sign}(\xi_k + \theta_k - \hat{\theta}_k) - \text{sign}(\xi_k)\}^2 \,|\mathcal{A}_{k-1}\Big] \\ & \leq \; 4EI\Big\{|\xi_k| \leq |\hat{\theta}_k - \theta_k| \,\Big|\mathcal{A}_{k-1}\Big\} \\ & = \; 4\int_{|u|\leq|\hat{\theta}_k-\theta_k|} f_k(u)du \\ & \leq \; 4\int_{|u|\leq|\hat{\theta}_k-\theta_k|} |f_k(u) - f_k(0)|du + 8f_0|\hat{\theta}_k - \theta_k| \\ & \leq \; 4L_\xi \int_{|u|\leq|\hat{\theta}_k-\theta_k|} |u|du + 8f_0|\hat{\theta}_k - \theta_k| \\ & \leq \; C(\hat{\theta}_k - \theta_k)^2 + c|\hat{\theta}_k - \theta_k|\,, \end{aligned} \tag{11}$$

we evaluate

$$\begin{aligned} E[(\mu_k')^2|\mathcal{A}_{k-1}] \; &= \; I_k E[T_k^2|\mathcal{A}_{k-1}] - I_k\{E[T_k|\mathcal{A}_{k-1}]\}^2 \\ &= \; (2f_0)^{-2} I_k E\Big[\{\text{sign}(\xi_k + \theta_k - \hat{\theta}_k) - \text{sign}(\xi_k)\}^2|\mathcal{A}_{k-1}\Big] \\ &\quad -(2f_0)^{-2} I_k\Big\{E[\text{sign}(\xi_k + \theta_k - \hat{\theta}_k) - \text{sign}(\xi_k)|\mathcal{A}_{k-1}]\Big\}^2 \\ &\leq \; 2(f_0)^{-2} EI\Big\{|\xi_k| \leq |\hat{\theta}_k - \theta_k|\Big|\mathcal{A}_{k-1}\Big\} \\ &\leq \; C(\hat{\theta}_k - \theta_k)^2 + c|\hat{\theta}_k - \theta_k| \end{aligned}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Because $K_{\epsilon,n} \subset K_n$ for sufficiently large n, the assertion (7) follows from (9) and the last relation.

Let us show that the sequence $\{\nu_k'\}$ satisfies (8). According to (2), one can see that

$$|\hat{\theta}_k| \leq M + b\,,$$

with $b = \sup_{u\geq 1} u^{-2\beta/(2\beta+1)} \log u$. Thus, $|\hat{\theta}_k|$ is bounded uniformly over $\theta \in \Theta_\beta$. As $|\theta_k| \leq H + L$ for each $\theta \in \Theta_\beta$, $|\delta_k| = |\hat{\theta}_k - \theta_k|$ is bounded uniformly over $\theta \in \Theta_\beta$, and so is $|T_k + \delta_k|$. Denote $h = M - (H + L) > 0$. Now it is easy to bound the first term in the expression for ν_k':

$$\begin{aligned} E|(T_k + \delta_k)I\{|\hat{\theta}_k| > H + L + h\}| \; &\leq \; CEI\{|\hat{\theta}_k| > H + L + h\} \\ &\leq \; CEI\{|\hat{\theta}_k - \theta_k| > h\} \\ &\leq \; cE(\hat{\theta}_k - \theta_k)^2\,, \end{aligned} \tag{12}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$ because

$$\{|\hat{\theta}_k| > H + L + h\} \subseteq \{|\hat{\theta}_k - \theta_k| > h\}\,.$$

To evaluate the second term in the expression for ν'_k, we first compute, using (10) (recall the notation $I_k = I\{|\hat\theta_k| \le M\}$):

$$\begin{aligned} & I_k(E[T_k|\mathcal{A}_{k-1}] + \delta_k) \\ &= I_k\Big((2f_0)^{-1}E[\text{sign}(\xi_k + \theta_k - \hat\theta_k) - \text{sign}(\xi_k)|\mathcal{A}_{k-1}] + \delta_k\Big) \\ &= I_k\Big(\hat\theta_k - \theta_k - f_0^{-1}\int_0^{\hat\theta_k-\theta_k} f_k(u)du\Big) \\ &= I_k f_0^{-1}\int_0^{\hat\theta_k-\theta_k}(f_0 - f_k(u))du\,. \end{aligned}$$

By using this and the fact $f_k \in \mathcal{P}_3(L_\xi)$, we obtain that

$$\begin{aligned} |I_k(E[T_k|\mathcal{A}_{k-1}] + \delta_k)| &\le f_0^{-1}\int_0^{\hat\theta_k-\theta_k}|f_0 - f_k(u)|du \\ &\le f_0^{-1}L_\xi\int_0^{\hat\theta_k-\theta_k}|u|du \\ &\le C(\hat\theta_k - \theta_k)^2 \end{aligned}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Combining the last inequality with (12) and (9), we conclude that the sequence $\{\nu'_k\}$ satisfies (8). The lemma is proved. □

Proof of Lemma 2. Write

$$\mu''_k = \frac{I_k(f_0 - \hat f_{0,k})\text{sign}(\xi_k)}{2\hat f_{0,k}f_0}\,,$$

$$\nu''_k = \frac{I_k S(X_k,\hat\theta_k)}{2\hat f_{0,k}} - \frac{I_k S(X_k,\hat\theta_k)}{2f_0} - \mu''_k + (Y_k - Y'_k)I\{|\hat\theta_k| > M\}\,,$$

$k = 1,\dots,n$. We have trivially

$$Y_k = Y'_k + \mu''_k + \nu''_k\,.$$

Let us show that the sequences $\{\mu''_k\}$ and $\{\nu''_k\}$ satisfy the properties (7) and (8) respectively. Obviously, $\{\mu''_k\}$ is a martingale difference with respect to the filtration $\{\mathcal{A}_k\}$, $k = 1,\dots,n$.

Suppose now that there exists a positive constant B_f such that

$$\limsup_{n\to\infty}\max_{k\in K_{\epsilon,n}}\frac{n^{2\beta/(2\beta+1)}}{(\log n)^2}E(f_0 - \hat f_{0,k})^2 \le B_f \tag{13}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Then the property (7) would follow immediately because $\hat f_{0,k} \ge p$ and $f_0 > p$. The property (8) would follow too. Indeed, recall that $|\hat\theta_k|$ is bounded uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$, $\hat f_{0,k} \ge p$ and $f_0 > p$. Therefore, $|Y_k - Y'_k|$

is also bounded uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. The same reasoning as in (12) yields that

$$E|Y_k - Y_k'|I\{|\hat{\theta}_k| > H + L + h\} \leq cE(\hat{\theta}_k - \theta_k)^2$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. The expectation of the absolute value of the first term in the expression for ν_k'' is bounded as follows: in view of (11), $\hat{f}_{0,k} \geq p$ and $f_0 > p$,

$$\begin{aligned}
&E\Big|\frac{I_k S(X_k, \hat{\theta}_k)}{2\hat{f}_{0,k}} - \frac{I_k S(X_k, \hat{\theta}_k)}{2f_0} - \mu_k''\Big| \\
&= E\Big|\frac{I_k(f_0 - \hat{f}_{0,k})(\text{sign}(\xi_k + \theta_k - \hat{\theta}_k) - \text{sign}(\xi_k))}{2\hat{f}_{0,k} f_0}\Big| \\
&\leq cE\Big\{|f_0 - \hat{f}_{0,k}|E\Big[|\text{sign}(\xi_k + \theta_k - \hat{\theta}_k) - \text{sign}(\xi_k)|\Big|\mathcal{A}_{k-1}\Big]\Big\} \\
&\leq cE\Big\{|f_0 - \hat{f}_{0,k}|2E\Big[I\{|\xi_k| \leq |\hat{\theta}_k - \theta_k|\Big|\mathcal{A}_{k-1}\Big]\Big\} \\
&\leq CE\Big\{|f_0 - \hat{f}_{0,k}||\hat{\theta}_k - \theta_k|\Big\} \\
&\leq C\Big\{E(f_0 - \hat{f}_{0,k})^2 E(\hat{\theta}_k - \theta_k)^2\Big\}^{1/2}
\end{aligned}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. The property (8) for the sequence $\{\nu_k''\}$ follows now from the definition for ν_k'', the last two bounds, (9) and (13).

Thus, it remains to show (13). Denote, for $i = 1, \ldots, n$,

$$D_i = I\Big\{|X_i - \hat{\theta}_i| \leq n^{-1/(2\beta+1)}\Big\} \quad \text{and} \quad G_i = E\big[D_i|\mathcal{A}_{i-1}\big]\,.$$

As $f_i \in \mathcal{P}_2$, we have for $k = 1, \ldots, n$

$$\begin{aligned}
E(f_0 - \hat{f}_{0,k})^2 &\leq E\Big[\frac{n^{1/(2\beta+1)}}{2k}\sum_{i=0}^{k-1} D_i - f_0\Big]^2 \\
&= E\Big[\frac{n^{1/(2\beta+1)}}{2k}\sum_{i=0}^{k-1}(D_i - G_i)\Big]^2 \\
&\quad + E\Big[\frac{n^{1/(2\beta+1)}}{2k}\sum_{i=0}^{k-1} G_i - f_0\Big]^2. \qquad (14)
\end{aligned}$$

Because $f_i \in \mathcal{P}_2 \cap \mathcal{P}_3$, it follows that $f_i(u) \leq C$ uniformly over any bounded interval and $i = 0, \ldots, n$. Recall also that $\delta_i = \hat{\theta}_i - \theta_i$ is bounded uniformly in i and over $\theta \in \Theta_\beta$.

Therefore, we obtain, for $i = 1, \dots, n$,

$$\begin{aligned} G_i &= E\Big[I\{|X_i - \hat\theta_i| \le n^{-1/(2\beta+1)}\}\big|\mathcal{A}_{i-1}\Big] \\ &= \int_{\delta_i - n^{-1/(2\beta+1)}}^{\delta_i + n^{-1/(2\beta+1)}} f_i(u)du \\ &\le \int_{\delta_i - n^{-1/(2\beta+1)}}^{\delta_i + n^{-1/(2\beta+1)}} |f_i(u) - f_i(0)|du + 2n^{-1/(2\beta+1)} f_0 \\ &\le Cn^{-1/(2\beta+1)} \end{aligned}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Taking this and the fact that $ED_i = EG_i$ into account, we bound the first term in the right hand side of (14): uniformly over $k \in K_{\epsilon,n}$ (i.e. $\epsilon n \le k$),

$$\begin{aligned} E\Big[\frac{n^{1/(2\beta+1)}}{k}\sum_{i=0}^{k-1}(D_i - G_i)\Big]^2 &= \frac{n^{2/(2\beta+1)}}{k^2}\sum_{i=0}^{k-1} E(D_i - G_i)^2 \\ &\le \frac{n^{2/(2\beta+1)}}{k^2}\sum_{i=0}^{k-1}(ED_i + EG_i^2) \\ &\le Cn^{2/(2\beta+1)}k^{-1}n^{-1/(2\beta+1)} \\ &\le cn^{-2\beta/(2\beta+1)} \end{aligned} \tag{15}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$.

Further,

$$\begin{aligned} \frac{n^{1/(2\beta+1)}G_i}{2} - f_0 &= \frac{n^{1/(2\beta+1)}}{2}\int_{\delta_i - n^{-1/(2\beta+1)}}^{\delta_i + n^{-1/(2\beta+1)}} (f_i(u) - f_0)du \\ &\le \frac{L_\xi n^{1/(2\beta+1)}}{2}\int_{\delta_i - n^{-1/(2\beta+1)}}^{\delta_i + n^{-1/(2\beta+1)}} |u|du \\ &\le C|\hat\theta_i - \theta_i| + cn^{-1/(2\beta+1)} \end{aligned}$$

uniformly over $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Let $j_0 = j_{0,n} = \min\{j : j \in K_n\}$, so $j_0 \le Cn^{2\beta/(2\beta+1)}$. Now, using the previous inequality, we evaluate the second term in the right

hand side of (14) as follows:

$$
\begin{aligned}
& E\Big[\frac{n^{1/(2\beta+1)}}{2k}\sum_{i=0}^{k-1} G_i - f_0\Big]^2 \\
& \le \; Cn^{-2/(2\beta+1)} + cE\Big[\frac{1}{k}\sum_{i=0}^{k-1}|\hat\theta_i - \theta_i|\Big]^2 \\
& \le \; Cn^{-2/(2\beta+1)} + \frac{c}{k^2}E\Big[\sum_{i=0}^{j_0}|\hat\theta_i - \theta_i| + \sum_{i=j_0+1}^{k-1}|\hat\theta_i - \theta_i|\Big]^2 \\
& \le \; cn^{-2/(2\beta+1)} + \frac{Cj_0^2}{k^2} + \frac{c}{k^2}\sum_{i,j=j_0+1}^{k-1} E\Big[|\hat\theta_i - \theta_i||\hat\theta_j - \theta_j|\Big] \\
& \le \; Cn^{-2/(2\beta+1)} + c\max_{i\in K_n} E(\hat\theta_i - \theta_i)^2
\end{aligned}
$$

uniformly over $k \in K_{\epsilon,n}$, $\theta \in \Theta_\beta$ and $f_\xi \in \mathcal{P}_\xi$. Because $\beta \le 1$ and $K_{\epsilon,n} \subset K_n$ for sufficiently large n, the assertion (13) follows from (9), (14), (15) and the last relation. The lemma is proved. □

References

[1] E.N. Belitser and A.P. Korostelev. Pseudovalues and minimax filtering algorithms for the nonparametric median. *Adv. in Sov. Math.* 12:115-124, 1992.

[2] E. Belitser and S. van de Geer (2000). On robust recursive nonparametric curve estimation. *High dimensional probability II*, 391–404, Progr. Probab., 47, Birkhäuser.

[3] B. Efron *The jackknife, the bootstrap and other resampling plans*. SIAM, Philadelphia, 1982.

[4] I.A. Ibragimov and R.Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer Verlag, New York, 1981

[5] A.P. Korostelev. Asymptotic minimax filtering of nonparametric signals, Technical report, Institute of System Studies, Moscow, 1987.

[6] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040-1053, 1982.

[7] Y.K. Truong. Asymptotic properties of kernel estimators based on local medians. *Annals of Statistics*, 18:606-617, 1989.

[8] A.B. Tsybakov. Nonparametric estimation under incomplete information on the noise distribution. *Problems Inform. Transmission*, 18:44-66, 1982.

[9] J. Tukey. Bias and confidence in not quite large samples. *Annals of Math. Statistics*, 29:614, 1958.

Eduard Belitser
Mathematical Institute
Utrecht University
P.O. Box 80010
3508 TA Utrecht
the Netherlands
e-mail: belitser@math.uu.nl