Pointwise adaptive estimation for robust and quantile regression

Markus Reiβ† Yves Rozenholc‡ Charles A. Cuenod§

This Version: April 3, 2009

Abstract

A nonparametric procedure for robust regression estimation and for quantile regression is proposed which is completely data-driven and adapts locally to the regularity of the regression function. This is achieved by considering in each point M-estimators over different local neighbourhoods and by a local model selection procedure based on sequential testing. Non-asymptotic risk bounds are obtained, which yield rate-optimality for large sample asymptotics under weak conditions. Simulations for different univariate median regression models show good finite sample properties, also in comparison to traditional methods. The approach is extended to image denoising and applied to CT scans in cancer research.

2000 Mathematics Subject Classification. Primary 62G08; secondary 62G20, 62G35, 62F05, 62P10.

Keywords and Phrases. M-estimation, median regression, robust estimation, local model selection, unsupervised learning, local neighbourhood, median filter, Lepski procedure, minimax rate, image denoising, edge detection.

*This research has been partially supported by a grant Bonus Qualité Recherche 2007 from Université Paris Descartes.
†Institut für Mathematik, Humboldt-Universität, Unter den Linden 6, D-10099 Berlin, Germany, mreiss@mathematik.hu-berlin.de (corresponding author)
‡Université Paris René Descartes, yves.rozenholc@univ-paris5.fr
§George Pompidou Hospital Paris, charles-andre.cuenod@egp.aphp.fr
1 Introduction

We consider a generalized regression model

\[ Y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \]

with \((\varepsilon_i)\) i.i.d., \(x_1, \ldots, x_n\) in the design space \(\mathcal{X}\) and \(g : \mathcal{X} \to \mathbb{R}\). The problems we have in view are those of robust nonparametric estimation of \(g\) in the presence of heavy-tailed noise \((\varepsilon_i)\) and of nonparametric quantile estimation, which is becoming more and more popular in applications. One main application will be robust image denoising. In the spirit of classical M-estimation (Huber 1964) we therefore consider \(g(x_i)\) as the location parameter in the observation \(Y_i\), that is

\[ g(x_i) = \operatorname{argmin}_{m \in \mathbb{R}} \mathbb{E}[\rho(Y_i - m)] \quad (1.1) \]

for some convex function \(\rho : \mathbb{R} \to \mathbb{R}^+\) with \(\rho(0) = 0\). We shall assume that \(g(x_i)\) is uniquely defined by (1.1), which is true in all cases of interest. If the \(Y_i\) have Lebesgue densities, then often an equivalent description is given by the first order condition \(\mathbb{E}[\rho'(\varepsilon_i)] = 0\) where \(\rho'\) denotes the (weak) derivative. Standard examples are \(\rho(x) = x^2/2\) for the classical mean regression model (\(\mathbb{E}[\varepsilon_i] = 0\)), \(\rho(x) = |x|\) for the median regression model (\(\mathbb{P}(\varepsilon_i \leq 0) = \mathbb{P}(\varepsilon_i \geq 0) = 1/2\)) and the intermediate case \(\rho(x) = x^2/2\) for \(|x| \leq k\) and \(\rho(x) = k|x| - k^2/2\) for \(|x| \geq k\) with some \(k > 0\) for the Huber estimator (\(\mathbb{E}[\min(\max(\varepsilon_i, -k), k)] = 0\)). The quantile regression model is obtained for \(\rho(x) = |x| + (2\alpha - 1)x\) (\(\mathbb{P}(\varepsilon_i \leq 0) = \alpha\) with quantile \(\alpha \in (0, 1)\)), see e.g. Koenker (2005). Since we shall care about robustness, we merely assume a mild moment condition \(\varepsilon_i \in L^r\) for some \(r \geq 1\) and measure the error in \(L^r\)-norm.

The function \(g\) is not supposed to satisfy a global smoothness criterion, but we aim at estimating it locally in each point \(x \in \mathcal{X}\) as efficiently as possible. The risk will then depend on local regularity properties, which we do not assume to be known. For spatially inhomogeneous functions, in the presence of jumps or for image denoising pointwise adaptive methods are much more appropriate than global smoothing methods. In classical mean regression local adaptivity can be achieved using wavelet thresholding or kernels with locally varying bandwidths, see Lepski et al. (1997) for a discussion. In this ideal situation a data-driven choice among linear empirical quantities is performed. M-estimators are typically nonlinear and the standard approaches do not necessarily transfer directly. Brown et al. (2008), for example, use an intermediate data binning and then apply wavelet thresholding to the binned data for median regression. On the other hand, Hall and Jones (1990), Portnoy (1997) and van de Geer (2003) consider kernels, smoothing splines and more general M-estimation for quantile regression, but they all use global methods for choosing the tuning parameters like cross-validation or penalisation. Here, we develop a generic algorithm to select optimally among local M-estimators. In contrast to classical model selection, we do not only rely on the estimator values themselves to define a data-driven selection criterion. This has significant advantages in the present case of nonlinear base estimators.
Subsequently, we assume that the statistician has chosen the suitable definition of $\rho$ for the problem at hand and we use the corresponding sample versions to construct base estimators for the (generalized) regression function $g$. In the spirit of classical nonparametrics, we assume that $g$ is locally almost constant around a given point $x \in \mathcal{X}$. The statistical challenge is to select adaptively the right neighbourhood $U$ of $x$ where a local $M$-estimator is applied. Let us write

$$m(Y_i, x_i \in U) := \arg\inf_{\mu \in \mathbb{R}} \left\{ \sum_{i : x_i \in U} \rho(Y_i - \mu) \right\} \quad (1.2)$$

for the location estimator on the set $U \subseteq \mathcal{X}$. If the minimizer is not unique, we just select one of them (e.g., a version of the sample median for $|U|$ even). Note that an extension to general local polynomial or more general local-likelihood estimation is straightforward, but this is not the focus of the present work. For each point $x$ let a family of nested neighbourhoods $U_0 \subseteq U_1 \subseteq \cdots \subseteq U_K$ be given and set

$$\hat{\vartheta}_k := m(Y_i, x_i \in U_k).$$

Then the family $(\hat{\vartheta}_k)_{0 \leq k \leq K}$ forms the class of base estimators and we aim at selecting the best estimator of $\vartheta := g(x)$ in this family.

1.1 Example. Let the design space be $\mathcal{X} = [0,1]$ with equidistant design points $x_i = i/n$ and take $\rho(x) = |x|$. Consider the symmetric windows $U_k = [x - h_k, x + h_k]$ generated by some bandwidths $0 \leq h_0 < h_1 < \cdots < h_K$. Then $\hat{\vartheta}_k$ is the classical median filter, see e.g. Truong (1989) or Arias-Castro and Donoho (2006).

Using Lepski’s approach as a starting point, we present our procedure to select optimally among local M-estimators in Section 2. We argue in a multiple testing interpretation that our procedure is usually more powerful. Moreover, it is equally simple to analyze and easy to implement. In Sections 3 and 4 we derive exact and asymptotic error bounds and the latter give optimal minimax rates for Hölder classes. The simulations in Section 5 show that our procedure has convincing finite sample properties. Moreover, they confirm that Lepski’s classical method applied to local median estimators suffers from oversmoothing because changes in the signal are not detected early enough due to the robustness of the median. Finally, the procedure has been implemented to denoise dynamical CT image sequences in Section 6 which is of key interest when assessing tumor therapies. Two more technical proofs are postponed to Section 7.

2 The procedure

2.1 Main ideas

As a starting point let us consider the standard Lepski (1990) method for selecting among $(\hat{\vartheta}_k)_{0 \leq k \leq K}$, given the mean regression model with $E[\varepsilon_i] = 0$ and
Lepski’s method accepts $H_0(k) : g|_{U_k} \equiv \vartheta$ that $g$ is constant on $U_k$ is tested against the alternative of significant deviations. Always assuming that $H_0(0)$ is true, we test sequentially whether $H_0(k + 1)$ is acceptable provided that the hypotheses $H_0(\ell)$ have been accepted for all $\ell \leq k$. Once the test of an $H_0(k + 1)$ is rejected, we select the base estimator $\hat{\vartheta}_k$ corresponding to the last accepted hypothesis. The main point is thus to properly define the single significance tests for $H_0(k + 1)$. Lepski’s method accepts $H_0(k + 1)$ if $|\hat{\vartheta}_{k+1} - \hat{\vartheta}_{k}| \leq z_{\ell}^{(k+1)}$ holds for all $\ell \leq k$ with suitable critical values $z_{\ell}^{(k+1)} > 0$. The wide applicability and success of Lepski’s method is also due to this very simple and intuitive test statistics.

In our nonlinear estimation case it turns out that tests for $H_0(k + 1)$ based on the differences of base estimators are often not optimal. To understand this fact, let us consider a toy model of two neighbourhoods $U_1 \subseteq U_2$ with a piecewise constant median regression function $g$ equal to $\mu_1$ on $U_1$ and to $\mu_2$ on $U_2 \setminus U_1$. The procedure therefore reduces to a simple two-sample location test between the observations in $U_1$ and in $U_2 \setminus U_1$. We proceed by considering abstractly a two-sample location test where the first sample $Y_1, \ldots, Y_n$ is i.i.d. with density $f_1(x) = \frac{1}{2\sigma} \exp(-|x - \mu_1|/\sigma)$ and the second independent sample $Y_{n+1}, \ldots, Y_{2n}$ is i.i.d. with density $f_2(x) = \frac{1}{2\sigma} \exp(-|x - \mu_2|/\sigma)$. Our goal is to test $H_0 : \mu_1 = \mu_2$ for known $\sigma > 0$. Given the Laplace distribution, we follow Lepski’s idea and put $\tilde{m}_1 = \text{med}(Y_i, i = 1, \ldots, n)$, the median over the first sample, and $\tilde{m}_2 = \text{med}(Y_i, i = 1, \ldots, 2n)$, the median over both samples. Then the test rejects if $T_L := 2|\tilde{m}_1 - \tilde{m}_2| > z$ holds for appropriate $z > 0$. A more classical approach, though, relies on a likelihood ratio (LR) test or on a Wald-type test using the maximum likelihood estimator for $\mu_1 - \mu_2$. Since the LR test is not as simple, we focus on the Wald-test statistic which is given by the difference of the medians over the two samples. Hence, we reject $H_0$ if $T_W := |\text{med}(Y_i, i = 1, \ldots, n) - \text{med}(Y_i, i = n + 1, \ldots, 2n)| > z$ holds for appropriate $z > 0$. The following asymptotic result for the two test statistics is proved in Section 7.1

2.1 Proposition. Let $f : \mathbb{R} \to \mathbb{R}^+$ be a symmetric and continuous density with $f(0) > 0$ and let $Y_1, \ldots, Y_n \sim f$, $Y_{n+1}, \ldots, Y_{2n} \sim f(\bullet - \Delta)$ with $\Delta > 0$ be independently distributed. Then with $F$ denoting the cumulative distribution
function of \( f \) we obtain for \( n \to \infty \)

\[
\sqrt{n}\left( \text{med}(Y_i, i = n + 1, \ldots, 2n) - \text{med}(Y_i, i = 1, \ldots, n) - \Delta \right) \Rightarrow N(0, \sigma_W^2)
\]

with \( \sigma_W^2 = \frac{1}{2f^2(0)} \),

\[
\sqrt{n}\left( 2 \left( \text{med}(Y_i, i = 1, \ldots, 2n) - \text{med}(Y_i, i = 1, \ldots, n) \right) - \Delta \right) \Rightarrow N(0, \sigma_L^2)
\]

with \( \sigma_L^2 = \frac{2F(\Delta/2)(1 - F(\Delta/2))}{f^2(\Delta/2)} + \frac{1}{f^2(0)} - \frac{2(1 - F(\Delta/2))}{f(0)f(\Delta/2)} \).

In particular, for \( \Delta = 0 \) we have \( \sigma_L^2 = \sigma_W^2 \) and for \( \Delta \to 0 \) we have the order \( \sigma_L^2 = \sigma_W^2(1 + 2\Delta f(0) + O(\Delta^2 f(0))) \), provided \( f \) is Lipschitz continuous at zero.

Putting \( \Delta = |\mu_1 - \mu_2| \) this result shows that under \( H_0 \), i.e. \( \Delta = 0 \), the test statistics \( T_L \) and \( T_W \) are asymptotically identically distributed, whereas \( T_L \) has a larger asymptotic variance under any alternative \( \Delta > 0 \) than \( T_W \). In the above Laplace model with densities \( f_1, f_2 \) this deterioration is only negligible if the signal-to-noise ratio satisfies \( |\mu_1 - \mu_2|/\sigma \ll 1 \). This is exactly what we see in simulations, see e.g. Example 1 in Section 5 below. Since the Laplace model is Hellinger differentiable, the Wald-type test is (locally) asymptotically efficient for \( n \to \infty \) as is the LR test, see e.g. van der Vaart (1998). Strictly speaking, when considering local alternatives for fixed \( \sigma > 0 \) and \( n \to \infty \), i.e. \( |\mu_1 - \mu_2| = O(n^{-1/2}) \), then the deterioration in using \( T_L \) becomes also negligible.

From a practical perspective, these local asymptotics are often not adequate, e.g. in image denoising, where we face relatively large signal differences \( \Delta \) at borders between objects and do not dispose of a very large number \( n \) of observed pixels.

More generally, two-sample location tests can naturally be based on the difference of the in-sample location estimators. In consequence, we proceed differently in testing the hypotheses \( H_0(k+1) \) of homogeneity: When the hypotheses \( H_0(\ell) \) for \( \ell \leq k \) have been accepted, we ask whether the observations \( Y_i \) in the new points \( x_i \in U_{k+1} \setminus U_k \) are homogeneous with those in \( U_\ell \) for \( \ell \leq k \). This means that our tests reject if the empirical location in the additional data

\[
\tilde{\vartheta}_{(k+1)|k} := m(Y_i, x_i \in U_{k+1} \setminus U_k)
\]

satisfies with certain critical values \( z_{(k+1)}^{(k+1)} > 0 \):

\[
\exists \ell \leq k: |\tilde{\vartheta}_{(k+1)|k} - \tilde{\vartheta}_\ell | > z_{(k+1)}^{(k+1)}.
\]

As in Lepski’s method, it is necessary to perform the testing for all \( \ell \leq k \) and not only with \( \ell = k \) to avoid that the signal slowly drifts away as the neighbourhoods grow. In most cases, though, \( H_0(k + 1) \) will be rejected because the new piece \( \tilde{\vartheta}_{(k+1)|k} \) is not in line with \( \tilde{\vartheta}_k \): due to the smaller variance of \( \tilde{\vartheta}_k \) compared to \( \tilde{\vartheta}_\ell, \ell < k \), this last test is the most powerful. It is then interesting to observe that for linear \( m \) the test statistic \( \tilde{\vartheta}_{(k+1)|k} - \tilde{\vartheta}_k \) is just a multiple of \( \tilde{\vartheta}_{k+1} - \tilde{\vartheta}_k \).
Consequently, for mean regression with linear base estimators our method will not differ much from Lepski’s standard method, whereas the general nonlinear M-estimators are treated in a significantly different way, note also the numerical results in Section 5.

Observe that our approach breaks an ubiquitous paradigm in modern statistics and learning theory (see e.g. Massart (2007) for model selection or Tsybakov (2004) for aggregation): we select the best base learner among \(\tilde{\vartheta}_k\) in a data-driven way not only based on the estimator values themselves, but additionally on the statistics \(\tilde{\vartheta}((k+1)\setminus k)\). Not only in the abstract modeling above, but also in implementations this idea turns out to be very advantageous for nonlinear estimators.

### 2.2 The algorithm

We want to select the best estimator among the family \(\{\tilde{\vartheta}_k \mid k = 0, \ldots, K\}\). Considering the law \(\mathbb{P}_0\) generated by the no-bias setting \(g \equiv 0\), we introduce the stochastic error levels

\[
s_j := \mathbb{E}_0[|\tilde{\vartheta}_j|^r]^{1/r}, \quad s_{kj} := \mathbb{E}_0[|\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j|^r]^{1/r}.
\]

We apply the following sequential procedure for prescribed critical values \((z_j)_{j=0,\ldots,K-1}\) and set \(z_K := 1\):

- **initialize** \(k := 0\);
- **repeat**
  - if for all \(j = 0, \ldots, k\)
    \[
    |\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j| \leq z_js_{kj} + z_{k+1}s_{k+1}
    \]
    then increase \(k\)
    else stop
  until \(k = K\);
- **put** \(\hat{k} := k\) and \(\hat{\vartheta} := \tilde{\vartheta}_{\hat{k}}\).

This algorithm to determine \(\hat{k}\) can be cast in one formula:

\[
\hat{k} := \inf \left\{ k \geq 0 \mid \exists j \leq k : |\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j| > z_js_{kj} + z_{k+1}s_{k+1} \right\} \wedge K.
\]

### 3 Error analysis

#### 3.1 Propagation and stopping late

We need a very natural property of the M-estimator.
3.1 Assumption. The location estimator in (1.2) satisfies for any set $S$ and any partition $S = \bigcup_j S_j$ with pairwise disjoint sets $S_j$:

$$\min_j m(Y_i, x_i \in S_j) \leq m(Y_i, x_i \in S) \leq \max_j m(Y_i, x_i \in S_j).$$

3.2 Lemma. If the function $\rho$ is strictly convex, then Assumption 3.1 is satisfied.

Proof. Let us write $m_T$ as short-hand for $m(Y_i, x_i \in T), T \subseteq \mathcal{X}$. Denoting by $\rho_+^\prime, \rho_-^\prime$ the right- and left-handed derivatives of the convex function $\rho$, the functions $\rho_+^\prime, \rho_-^\prime$ are strictly increasing with $\rho_+^\prime(x) < \rho_-^\prime(y) \leq \rho_+^\prime(y)$ for all $x < y$ and

$$\sum_{x_i \in T} \rho_+^\prime(Y_i - m_T) \leq 0, \quad \sum_{x_i \in T} \rho_-^\prime(Y_i - m_T) \geq 0.$$

If $m_S < m_{S_j}$ were true for all $j$, then

$$\sum_{x_i \in S} \rho_-^\prime(Y_i - m_S) > \sum_j \sum_{x_i \in S_j} \rho_+^\prime(Y_i - m_{S_j}) \geq 0,$$

which contradicts the minimizing property of $m_S$. Hence, $m_S \geq \min_j m_{S_j}$ holds and a symmetric argument shows $m_s \leq \max_j m_{S_j}$.

3.3 Remark. If $\rho$ is not strictly convex, then we usually impose additional conditions to define $m$ uniquely. For any reasonable specific choice Assumption 3.1 should be satisfied. In particular, this is true for the sample median where we take for an even number $N$ of data points $Y_i$ the mean $(Y_{(N/2)} + Y_{(1+N/2)})/2$ of the order statistics.

3.4 Proposition. Grant Assumption 3.1. Then we have for any $k = 0, \ldots, K - 1$

$$|\hat{\vartheta} - \tilde{\vartheta}_k| 1\{k > k\} \leq \max_{j=k+1, \ldots, K-1} (\sum_{j \in \mathcal{J}_k} s_{jk} + \sum_{j+1 \in \mathcal{J}_k} s_{j+1}).$$

3.5 Remark. This error propagation result is true 'ω-wise', that is, it does not depend on the noise realisation. It is built into the construction of the selection procedure. An analogous result holds for Lepski’s original procedure (Lepski 1990, Lepski et al. 1997).

Proof. From Assumption 3.1 we infer for $\ell > k$

$$|\hat{\vartheta}_\ell - \tilde{\vartheta}_k| \leq \max_{k+1 \leq \ell \leq \ell} |\hat{\vartheta}_{\mathcal{J}_k(j-1)} - \tilde{\vartheta}_k|.$$

We therefore obtain on the event $\{k > k\}$ by construction

$$|\hat{\vartheta} - \tilde{\vartheta}_k| \leq \max_{j=k+1, \ldots, K-1} |\hat{\vartheta}_{\mathcal{J}_k(j-1)} - \tilde{\vartheta}_k| \leq \max_{j=k+1, \ldots, K-1} (\sum_{j \in \mathcal{J}_k} s_{jk} + \sum_{j+1 \in \mathcal{J}_k} s_{j+1}).$$
3.6 Example. For geometrically decreasing stochastic error levels \( s_k \) in (2.1), in particular for the median filter from Example 1.1 with bandwidths \( h_k = h_0 q^k \), we have \( s_{jk} \lesssim s_k \) for \( j > k \), where \( A \lesssim B \) means \( A = \mathcal{O}(B) \) in the \( \mathcal{O} \)-notation. The late stopping error is of order \( z_{r_k} s_{r_k} \), provided the critical values \( (z_k) \) are non-increasing. This will imply that the error due to stopping later than some optimal \( k^* \) is increased by at most the order of \( z_{r_k} s_{r_k} \):

\[
\mathbb{E}_\vartheta[|\tilde{\vartheta} - \vartheta|^r 1(\hat{k} > k^*)] \lesssim \mathbb{E}_\vartheta[|\tilde{\vartheta}_{k^*} - \vartheta|^r] + z_{r_k}^* s_{k^*}^* \leq (1 + z_{r_k}^* s_{k^*}^* ) \mathbb{E}_\vartheta[|\tilde{\vartheta}_{k^*} - \vartheta|^r].
\]

3.2 Critical values and stopping early

As the preceding analysis shows, small critical values \( (z_k) \) lead to small errors caused by stopping late. On the other hand, the \( (z_k) \) should not be too small in order to control the error of stopping early. To this end, we shall require a condition on the critical values \( (z_k) \) in the no-bias situation under \( P_0 \), that is for constant \( g \equiv 0 \). In fact, we face a multiple testing problem, but with an estimation-type loss function. For some confidence parameter \( \alpha > 0 \) we select \( z_k > 0, k = 0, \ldots, K - 1 \), such that the condition

\[
\sum_{j=0}^{K-1} \mathbb{E}_0 \left[ |\tilde{\vartheta}_j|^r 1( \exists \ell \leq j : |\tilde{\vartheta}_{(j+1) \setminus j} - \tilde{\vartheta}_\ell| > z_{\ell s_j} \right] \leq \alpha s_{r_k}^K.
\]

(3.1)

is satisfied. In order to obtain a unique prescription for each \( z_k \) that equilibrates the errors for different stopping times of the algorithm, we can select the \( (z_k) \) sequentially. We choose \( z_0 \) such that

\[
\sum_{j=0}^{K-1} \mathbb{E}_0 \left[ |\tilde{\vartheta}_j|^r 1( |\tilde{\vartheta}_{(j+1) \setminus j} - \tilde{\vartheta}_0| > z_0 s_j \right] \leq \frac{\alpha}{K} s_{r_k}^K.
\]

and then each \( z_k \) for given \( z_0, \ldots, z_{k-1} \) such that

\[
\sum_{j=k}^{K-1} \mathbb{E}_0 \left[ |\tilde{\vartheta}_j|^r 1( |\tilde{\vartheta}_{(j+1) \setminus j} - \tilde{\vartheta}_k| > z_k s_{jk} \right], \forall \ell < k : |\tilde{\vartheta}_{(j+1) \setminus j} - \tilde{\vartheta}_\ell| \leq z_{\ell s_j} \right] \leq \frac{\alpha}{K} s_{r_k}^K.
\]

(3.2)

To determine the \( (z_k) \) in practice, we simulate in Monte Carlo iterations the pure noise case \( g \equiv 0 \) and calculate for each \( k \) the error when the algorithm stops before the (theoretically optimal) index \( K \) due to a rejected test involving \( z_k \). The critical values are determined such that this error is a fraction of the oracle estimation error \( s_{r_k}^* \). For this calibration step the original algorithm of Section 2.2 is taken, only modified by using \( z_j s_{kj} \) instead of \( z_j s_{kj} + z_{k+1} s_{k+1} \) in the testing parts.

The selection rule for the critical values in Lepski’s procedure is the focus in the work by Spokoiny and Vial (2009). Their idea is to transfer properties from the no-bias situation to the general nonparametric specification by bounding the likelihood between the two observation models. This approach, the so-called
small modeling bias condition, could be applied here as well and will give similar results. On a practical level, the difference is that Spokoiny and Vial (2009) enlarge the moment from \( r \) to \( 2r \) in the calibration step, while we add the term \( z_{k+1} s_{k+1} \) to the testing values \( z_{j} s_{k} \) from the calibration. In the asymptotic analysis, however, the method by Spokoiny and Vial (2009) costs us some power in the logarithmic factor and we would thus not attain optimal rates over Hölder balls, cf. Section 4. Moreover, for robustness reasons, we do not want to require higher moment bounds for the error variables and the likelihood.

3.7 Definition. Given the regression function \( g \), introduce its variation on \( U_k \)

\[ V_k(g) := \sup_{y_1, y_2 \in U_k} |g(y_1) - g(y_2)| \]

and consider the oracle-type index

\[ k^* := \min\{k = 0, \ldots, K - 1 \mid V_{k+1}(g) > z_{k+1} s_{k+1}\} \wedge K. \]

This definition implies that for all \( k \leq k^* \) the maximal bias \( V_k(g) \) of \( \tilde{\vartheta}_k \) is less than its stochastic error level \( s_k \) from (2.1) times the critical value \( z_k \). The next result, when specialised to \( k = k^* \), means intuitively that the error due to stopping before \( k^* \) can be bounded in terms of the stochastic error of \( \tilde{\vartheta}_{k^*} \), involving the critical value \( z_{k^*} \) as a factor. Let us also mention here that the rationale for the choice \( z_K = 1 \) in the algorithm of Section 2.2 is to equilibrate maximal bias and stochastic error at step \( k = K - 1 \).

3.8 Proposition. We have for any \( k = 0, \ldots, k^* \)

\[ \mathbb{E} \left[ |	ilde{\vartheta} - \tilde{\vartheta}_k|^r 1(\hat{k} < k) \right] \leq (3^{r-1} \vee 1)(z_k^r + 1 + \alpha) s_k^r. \]

Proof. We shall write \( \hat{k}(g), \tilde{\vartheta}_k(g) \) etc. to indicate that \( \hat{k}, \tilde{\vartheta}_k \) etc. depend on the underlying regression function \( g \). We shall need the inequality

\[ |\tilde{\vartheta}_j(g) - \tilde{\vartheta}_k(g)| \leq |\tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0)| + V_k(g) \text{ for } j < k \quad (3.3) \]

which follows from

\[ \tilde{\vartheta}_j(g) - \tilde{\vartheta}_k(g) = m(g(x_i) + \varepsilon_i, x_i \in U_j) - m(g(x_i) + \varepsilon_i, x_i \in U_k) \leq m(\varepsilon_i, x_i \in U_j) + \sup_{x \in U_j} g(x) - m(\varepsilon_i, x_i \in U_k) - \inf_{x \in U_k} g(x) \leq \tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0) + V_k(g) \]

and by a symmetric argument for \( \tilde{\vartheta}_k(g) - \tilde{\vartheta}_j(g) \).

By definition of \( k^* \) and using the condition on the \( (z_k) \) as well as (3.3) for \( \tilde{\vartheta}_j \)
and \( \tilde{\vartheta}_{(j+1)\backslash j} \), we obtain for all \( k \leq k^* \)

\[
\mathbb{E} \left[ |\hat{\vartheta}(g) - \tilde{\vartheta}_k(g)|^r 1(\hat{k}(g) < k) \right] \\
= \sum_{j=0}^{k-1} \mathbb{E} \left[ |\tilde{\vartheta}_j(g) - \tilde{\vartheta}_k(g)|^r 1(\hat{k}(g) = j) \right] \\
\leq \sum_{j=0}^{k-1} \mathbb{E} \left[ (V_k(g) + |\tilde{\vartheta}_j(0)| + |\tilde{\vartheta}_k(0)|)^r 1(\hat{k}(g) = j) \right] \\
\leq (3^{-1} \lor 1) \left( V_k(g)^r + \mathbb{E} [|\tilde{\vartheta}_k(0)|^r] + \sum_{j=0}^{k-1} \mathbb{E} \left[ |\tilde{\vartheta}_j(0)|^r \left( \exists \ell \leq j : |\tilde{\vartheta}_{(j+1)\backslash j}(g) - \tilde{\vartheta}_{\ell}(g)| > z_{j\ell}s_{j\ell} + z_{j+1}s_{j+1} \right) \right] \right) \\
\leq (3^{-1} \lor 1) \left( z_k^r s_k^r + s_k^r + \sum_{j=0}^{k-1} \mathbb{E} \left[ |\tilde{\vartheta}_j(0)|^r \left( \exists \ell \leq j : |\tilde{\vartheta}_{(j+1)\backslash j}(0) - \tilde{\vartheta}_{\ell}(0)| + V_{j+1}(g) > z_{j\ell}s_{j\ell} + z_{j+1}s_{j+1} \right) \right] \right) \\
\leq (3^{-1} \lor 1) \left( z_k^r s_k^r + s_k^r + \alpha s_K^r \right).
\]

The result follows from the isotonic decay of \((s_k)\).

### 3.3 Total risk bound

#### 3.9 Theorem

Assume that \((z_k s_k)\) is non-increasing in \(k\). Then under Assumption 3.2 the following excess risk estimate holds for all \(k \leq k^*\):

\[
\mathbb{E} \left[ |\hat{\vartheta} - \tilde{\vartheta}_k^*| \right] \leq (3^{-1} \lor 1) \left( (2z_k^r + 1 + \alpha)s_k^r + z_k^r \max_{j=k+1 \ldots K-1} s_{j_k}^r \right).
\]

**Proof.** For the late-stopping error Proposition 3.4 and the decay of \((z_k s_k)\) give

\[
|\hat{\vartheta} - \tilde{\vartheta}_k^*|^r 1(\hat{k} > k) \leq (2^{-1} \lor 1) \max_{j > k} (z_{j_k}^r s_{j_k}^r + z_{j+1}^r s_{j+1}^r) \leq (2^{-1} \lor 1) z_k^r \max_{j > k} s_{j_k}^r.
\]

Add the early-stopping error from Proposition 3.8.

#### 3.10 Example (continued)

For geometrically increasing bandwidths \((h_k)\) we obtain \(s_{j_k} \lesssim s_k\) for \(j > k\) and thus

\[
\mathbb{E} \left[ |\hat{\vartheta} - \tilde{\vartheta}_k^*|^r \right] \lesssim (\alpha + z_k^r) s_k^r.
\]

The factor \(\alpha + z_k^r\) is the term we pay for adaptation.
4 Asymptotic risk

4.1 General result

We shall derive convergence rates for \( n \to \infty \) of the critical values \( (z_k) \). All quantities in the procedure may depend on \( n \), but we still write \( U_k, K, z_k \) instead of \( U_k(n), K(n), z_k(n) \). The notation \( A \lesssim B \) will always mean \( A(n) \leq cB(n) \) with some \( c > 0 \) independent of \( n \) and \( A \sim B \) is short for \( A \lesssim B \) and \( B \lesssim A \). We work under the following assumption whose validity under mild conditions will be derived in the next subsection.

4.1 Assumption.

(a) The cardinalities \( N_k \) of the neighbourhoods \( U_k \) grow with geometric order:

\[
q_1 N_k \leq N_{k+1} \leq q_2 N_k \quad \text{for all } k = 0, \ldots, K - 1
\]

for some fixed \( q_2 \geq q_1 > 1 \) and with \( N_1/\log(N_K) \to \infty \), \( N_K \sim n \) as \( n \to \infty \).

(b) For all sufficiently large \( N \) we have

\[
\mathbb{E}[|m(\varepsilon_i, i = 1, \ldots, N)|^{r}]^{1/r} \sim \mathbb{E}[|m(\varepsilon_i, i = 1, \ldots, N)|^{2r}]^{1/2r} \sim N^{-1/2}.
\]

(c) For all \( \tau_N \to \infty \) with \( \tau_N N^{-1/2} \to 0 \) a moderate deviations bound applies: there is some \( c > 0 \) such that

\[
\limsup_{N \to \infty} e^{c\tau_N^2} \mathbb{P}(N^{1/2}|m(\varepsilon_i, i = 1, \ldots, N)| > \tau_N) < \infty.
\]

The following asymptotic bounds follow directly from the definitions:

4.2 Lemma. Assumption 4.1(b) implies \( s_j \sim N_j^{-1/2} \) and \( N_j^{-1/2} \wedge (N_{k+1} - N_k)^{-1/2} \lesssim s_{kj} \lesssim N_j^{-1/2} \vee (N_{k+1} - N_k)^{-1/2} \). Assumption 4.1(a) then yields for \( k \geq j \)

\[
s_j \sim s_{kj} \sim N_j^{-1/2}.
\]

Under Assumption 4.1 critical values of the same order as in the Gaussian case suffice.

4.3 Proposition. Grant Assumption 4.1 and suppose \( \alpha \in (0, 1) \). We can choose

\[
z_k^2 = \zeta (2r \log(s_k/s_K) + \log(\alpha^{-1}) + \log(K)), \quad k = 0, \ldots, K - 1,
\]

with \( \zeta > 0 \) a sufficiently large constant in order to satisfy Condition (3.2). For \( K \sim \log n \) this yields asymptotically \( z_k \sim \sqrt{\log n} \).
Note that the chosen critical values \( z_k \) are decreasing in \( k \), which has the desirable effect that we do not permit stopping at an early stage with the same probability as stopping at higher indices \( k \). Moreover, this guarantees that \( z_k s_k \) is non-increasing in \( k \), the hypothesis in Theorem 3.9. From Theorem 3.9 we therefore obtain the following asymptotic risk bound.

4.4 Corollary. Grant Assumptions 3.1 and 4.1 and let \( K \sim \log n \). Choosing the critical values as in Proposition 4.3 gives

\[
\mathbb{E}[|\tilde{\theta} - \theta|^r] \lesssim (\log n)^{r/2} \mathbb{E}[|\tilde{\theta}_{k^*} - \theta|^r].
\]

4.5 Example (continued). Let us specify to \( s \)-Hölder continuous \( g : [0, 1] \rightarrow \mathbb{R} \), equidistant design and kernel estimators with geometrically increasing bandwidths \( h_k = h_0 q^k \), \( K \sim \log(n) \). Then we can choose \( z_k \sim \sqrt{\log(n)} \) and the index \( k^* \) satisfies \( V_{k^*} f^2 \sim h_{k^*}^{2s} \sim (n h_{k^*})^{-1} \log(n) \), that is \( h_{k^*} \sim (\log(n)/n)^{1/(2s+1)} \) and \( z_k s_k \sim (\log(n)/n)^{s/(2s+1)} \). This is the classical minimax rate for pointwise adaptive estimation in the one-dimensional \( s \)-Hölder continuous case, see Lepski et al. (1997) for the Gaussian case. Here, we have derived the same rate for pointwise adaptive \( M \)-estimation under very weak conditions on the error distribution, compare the discussion on specific models below. Let us also mention that Truong (1989) obtains the same rate result, but without logarithmic factor, for the non-adaptive median regression case.

Proof of Proposition 4.3. Let \( j \geq k \). For \( n \) sufficiently large Assumption 4.1(c) together with the asymptotics \( z_k s_{jk} \lesssim (\log(N_k)N_k^{-1})^{1/2} \sim 0 \) (using Assumption 4.1(a,b) and Lemma 4.2) yields

\[
\mathbb{P}_0(|\tilde{\theta}_{j+1}\setminus j - \tilde{\theta}_k| > z_k s_{jk}) \\
\leq \mathbb{P}_0(|\tilde{\theta}_{j+1}\setminus j| > z_k s_{jk}/2) + \mathbb{P}_0(|\tilde{\theta}_k| > z_k s_{jk}/2) \\
\lesssim \exp(-cz_k^2 s_{jk}^2 (N_{j+1} - N_j)/4) + \exp(-cz_k^2 s_{jk}^2 N_k/4).
\]

By Lemma 4.2 there is another constant \( c' > 0 \) such that for large \( z_k \)

\[
\mathbb{P}_0(|\tilde{\theta}_{j+1}\setminus j - \tilde{\theta}_k| > z_k s_{jk}) \lesssim \exp(-c'z_k^2).
\]

Our choice of \( z_k \) with \( \zeta \) sufficiently large guarantees \( \exp(-c'z_k^2/2) = o(\alpha(s_K/s_k)^r K^{-2}) \) for large \( K \). We therefore more than satisfy (3.1) and the construction in (3.2) provided \( n \) is sufficiently large:

\[
\sum_{j=k}^{K-1} \mathbb{E}
\left[
|\tilde{\theta}_j|\mathbf{1}
\left(
|\tilde{\theta}_{j+1}\setminus j - \tilde{\theta}_k| > z_k s_{jk}
\right)
\right]
\leq \sum_{j=k}^{K-1} \mathbb{E}
\left[
|\tilde{\theta}_j|^r\mathbf{1}
\left(
|\tilde{\theta}_{j+1}\setminus j - \tilde{\theta}_k| > z_k s_{jk}
\right)
\right]^{1/2}
\leq \sum_{j=k}^{K-1} s_j^{r} \exp(-c'z_k'^2/2)
\leq o\left(\frac{K^r s_k^r}{K}\right).
\]
For $K \sim \log N$ we obtain \( \log(N_K/N_k) \leq (K - k) \log q_2 \leq \log N \) and thus \( z_k^2 \sim \log n \).

### 4.2 Specific models

The preceding asymptotic analysis was based on Assumption 4.1 where part (a) can be ensured by construction whereas parts (b) and (c) depend on the noise model and the choice of M-estimator. The most severe restriction will usually be the moderate deviation property of Assumption 4.1(c). In the case where the law of the error variable \( \varepsilon_i \) is absolutely continuous, this property holds by Corollary 2.1 in Arcones (2002) under the following conditions:

(a) \( \mathbb{E}[\rho(\varepsilon_i + h) - \rho(\varepsilon_i)] = Vh^2 + o(h^2) \) for some \( V > 0 \) and \( |h| \to 0 \);
(b) \( \rho \) is Lebesgue-almost everywhere differentiable with derivative \( \rho' \);
(c) there are \( \lambda, \delta > 0 \) such that \( \mathbb{E}[\exp(\lambda|\rho'(\varepsilon_i)\!|)] \) and \( \mathbb{E}[\exp(\lambda \sup_{|h|\leq \delta} |\rho(\varepsilon_i + h) - \rho(\varepsilon_i) - h\rho'(\varepsilon_i)/h|)] \) are finite.

For mean regression \( \rho(x) = x^2 \) we have \( V = 1 \) and \( \rho'(\varepsilon_i) = 2\varepsilon_i \) such that a finite exponential moment for \( \varepsilon_i \) is required. For median regression the result applies with \( V = f_{\varepsilon}(0) / 2 \) and \( \rho'(\varepsilon_i) = \text{sgn}(\varepsilon_i) \) and because of \( ||\varepsilon_i + h| - |\varepsilon_i| - h\text{sgn}(\varepsilon_i)|| \leq 2|h| \) no moment bound is required. The same is true for any robust statistic with bounded influence function, in particular for the Huber estimator and general quantile estimators. Arcones (2002) discusses that an exponential tail estimate for \( \rho'(\varepsilon_i) \) is also necessary to obtain a moderate deviation bound, which might be a serious drawback when using Lepski’s method with linear non-robust estimators.

For the median the requirements are not difficult to verify directly. Assumption 4.1(b) is for example established by Chu and Hotelling (1955), who show that for \( f_{\varepsilon} \) continuously differentiable around zero, \( f_{\varepsilon}(0) > 0, r \in \mathbb{N} \) and \( Z \sim N(0, 1) \):

\[
\lim_{N \to \infty} N^r \mathbb{E}[\text{med}(\varepsilon_1, \ldots, \varepsilon_N)^{2r}] = (2f_{\varepsilon}(0))^{-r} \mathbb{E}[Z^{2r}].
\]

Using a coupling result, we can establish Assumption 4.1(b,c) under even more general conditions, see Section 7.2 for a proof:

**4.6 Proposition.** Assume that the \( \varepsilon_i \) have a Lebesgue density \( f_{\varepsilon} \) which is Lipschitz continuous at zero and satisfies \( \int_{-\infty}^{0} f_{\varepsilon}(x) dx = 1/2, f_{\varepsilon}(0) > 0, \mathbb{E}[|\varepsilon_i|^r] < \infty \). Noting \( \text{med}(\varepsilon) := \text{med}(\varepsilon_1, \ldots, \varepsilon_N), N \text{ odd}, \) we have

\[
\forall N \geq 5: \mathbb{E}[|\text{med}(\varepsilon)|^r] \sim N^{-r/2} \text{ and } \mathbb{E}[|\text{med}(\varepsilon)|^{2r}] \sim N^{-r}
\]

as well as for \( \tau_N \to \infty \) with \( \tau_N = o(N^{1/2}) \)

\[
\limsup_{N \to \infty} \mathbb{P}\left(2N^{1/2}g_{\varepsilon}(0)|\text{med}(\varepsilon)| > \tau_N\right) \exp(\tau_N^2/8) \leq 2.
\]
5 Simulation results

We illustrate our procedure by an implementation for median regression on \( \mathcal{X} = [-1, 1] \) and the estimation of the regression function at \( x = 0 \). We simulate \( n = 200 \) equidistant observations \( (Y_i) \) with standardized errors \( (\varepsilon_i) \) \((\mathbb{E}[\varepsilon_i] = 0, \text{Var}(\varepsilon_i) = 1)\) that are (a) Laplace, (b) normal and (c) Student t-distributed with three degrees of freedom. The location is each time estimated by local sample means as well as by local sample medians. As neighbourhoods we take symmetric intervals \( U_k \) around zero containing \( \lfloor 5k/4 \rfloor \) data points. This gives \( K = 17 \) different base estimators.

The calibration of the procedure is performed for Laplace distributed errors with \( r = 2 \) and \( \alpha = 1 \). The variances \( s_j, s_{jk} \) of the sample means are calculated exactly and those of the sample medians are approximated by their asymptotic values (which are quite close to Monte Carlo values). The critical values \( (z_k) \) are chosen according to the prescription in (3.1). This is achieved in both cases, mean and median estimators, by using the choice in Proposition 4.3 with values \( \zeta \) that are calibrated by 10000 Monte Carlo runs for the pure noise situation. It turned out that this gives almost equally sized error contributions for the different values \( z_k \), as postulated in (3.2). The same calibration principle was applied for the original Lepski procedure with mean and median estimators.

As a first example we take a simple change point problem by considering the regression function \( g(x) = 0 \) for \( |x| \leq 0.2 \) and \( g(x) = 2 \) for \( |x| > 0.2 \), which can be considered as a toy model for edge detection in image restoration or for structural breaks in econometrics. In Figure 1 we show a typical data set in the Laplace case (a) together with box plots for the absolute error of the different methods in 1000 Monte Carlo repetitions: local means with Lepski’s and with

Figure 1: Example 1 with Laplace noise: A typical realisation and a box plot of the sample errors in 1000 Monte Carlo runs.
Figure 2: Box plot of the sample errors in 1000 Monte Carlo runs for Gaussian (left) and Student t(3) noise (right).

our method, local medians with Lepski’s and with our method and the oracle method, which is just the sample median over \([-0.2, 0.2] = \{x : g(x) = 0\}\). For exactly the same methods, especially still calibrated to Laplace errors, Figure 2 presents the results for Gaussian and heavy-tailed Student t(3) errors.

It is obvious that in all cases Lepski’s method applied to sample medians as base estimators works quite badly. This is due to the fact that this method stops far too late: the sample median over the complete intervals \(U_k\) does not really ‘notice’ the jump in the data. In fact, in the Laplace simulation study the oracle \(k = 10\) is selected by this method in less than 1% of the cases while most often (65%) the selection is \(k = 12\) which yields the 1.5 times larger window \(U_{12} = [-0.29, 0.29]\). The methods using the sample mean estimators perform reasonably well and especially both very similarly. Still, they are clearly beaten by our median based procedure in cases (a) and (c) where the median is the more efficient location estimator. It is remarkable here that we nearly achieve the risk of the oracle median estimator. Even in the Gaussian case (b) the linear procedures have only minor advantages. Finally, we notice the robustness property that the calibration with the wrong error distribution in Figure 2 does not seriously affect the results.

In a second example we consider the smooth regression function \(g(x) = 2x(x + 1)\). Because we are estimating locally around \(x = 0\), this is a caricature of a \(C^2\)-function with \(g'(0) = 2\) and \(g''(0) = 4\). Figure 3 shows again a typical data set and boxplots for the different methods in 1000 Monte Carlo runs under Laplace errors. This time the oracle choice is the window \([-0.39, 0.39]\). Our median based procedure outperforms the others where the advantage over the mean-based approaches is again mainly due to the relative efficiency gain of size \(1/\sqrt{2}\) induced by the base estimators in the Laplace model. This gain, though,
Figure 3: Example 2 with Laplace noise. A typical realisation and a box plot of the sample errors in 1000 Monte Carlo runs.

is not at all visible when using Lepski’s method for selecting among the sample medians. The results for the error distributions (b) and (c) resemble those of the first example, we confine ourselves to summarizing the numerical results for all examples in the following table, each time stating the Monte Carlo median of the absolute error:

| Ex. | Mean Lepski | Mean RR | Median Lepski | Median RR | Median Oracle |
|-----|-------------|---------|---------------|-----------|---------------|
| 1a  | 0.1446      | 0.1450  | 0.2871        | 0.0897    | 0.0763        |
| 1b  | 0.1640      | 0.1630  | 0.2795        | 0.1647    | 0.1325        |
| 1c  | 0.0982      | 0.0978  | 0.3012        | 0.0596    | 0.0560        |
| 2a  | 0.1846      | 0.1924  | 0.3051        | 0.1246    | 0.1005        |
| 2b  | 0.1808      | 0.1886  | 0.3430        | 0.1586    | 0.1241        |
| 2c  | 0.2102      | 0.2126  | 0.2455        | 0.1047    | 0.0822        |

Further simulation experiments confirm this picture. Especially for lower values of the moment $r$ our median-based procedure is very efficient, while sometimes for $r = 2$ the mean-based procedures profit from less severe outliers in the Monte Carlo runs. In all these experiments the location is equally described by mean and median and we mainly see the efficiency gain of the sample median for non-Gaussian noise. For general quantile regression, however, linear methods do not apply and the standard Lepski procedures based on the nonlinear base estimators will perform badly. Our approach gives significantly better results. The error reductions by a factor of two and more, achieved in the median procedures above, confirm this very clearly.
6 Application

The proposed procedure is applied to denoise images used in the surveillance of cancer therapies. In Dynamic Contrast Enhanced Computer Tomography (DCE-CT) a contrast agent is injected in the human body and its diffusion over time is observed which is specific for different kinds of cell tissues and allows thus the surveillance of cancer therapies. For medical reasons the dose of contrast agent is kept small which leads to a poor signal-to-noise ratio. An analysis of residuals shows that the observational noise is well modeled by the Laplace distribution. Moreover, sometimes human movements produce significant outliers. Therefore local median estimation is employed. Especially for dynamical image sequences, the denoising is remarkably successful when the same spatial neighbourhoods are used over the whole observation period. This means that at each voxel location $x_i$ a vector-valued intensity function $g: \mathcal{X} \rightarrow \mathbb{R}^K$ is observed under vector-valued noise $\varepsilon_i$. The vector $g(x_i)$ encodes the intensity at time points $(t_1, \ldots, t_K)$ recorded at spatial location $x_i$. Our previously developed procedure perfectly applies to this situation, we just need a testing procedure between vector-valued local M-estimators.

Details of the experimental setup and the estimation procedure are discussed in Rozenholc et al. (2009) and we merely give a rough description of the setting. A multiresolution test procedure is applied to compare different vector estimates. In a first pre-selection step for each voxel $x_i$ we disregard voxels that are significantly different from $x_i$ and construct then circular neighbourhoods around $x_i$ consisting only of non-rejected voxels. This allows geometrically richer neighborhood structures that in practice adapt well to the structure. Mathematically, the analysis of the algorithm remains the same when conditioning on the result of this first pre-selection.

For the present example we dispose of a DCE-CT sequence of $K = 53$ recordings of $512 \times 512$-pixel images in the upper abdomen of a cancer patient. In Figure 4 the original image at time step 23 is depicted together with the result of our
denoising procedure. The noise reduction is remarkable while fine structures like edges are well preserved and not smoothed out. The residuals in Figure 5(left) show some artefacts due to human body movements and CT radial artefacts, which our procedure removed as well. In Figure 5(right) a zoom into Figure 4(right) is shown together with the sequence of neighbourhoods constructed for one voxel inside the cancerogeneous tissue. The effect of the pre-selection step is clearly visible by the geometrically adaptive form of the neighbourhoods. Further results, in particular the denoised dynamics in certain voxels and an application to automatic clustering of cell tissues are reported in Rozenholc et al. (2009). The generality of our procedure has the potential to provide statistical solutions in many further applications where spatial inhomogeneity and robustness are key issues.

7 Appendix

7.1 Proof of Proposition 2.1

The asymptotic normality of the sample median \( \sqrt{n} \text{med}(Y_1, \ldots, Y_n) \Rightarrow N(0, 1/(4f^2(0))) \) is well known (van der Vaart 1998, Corollary 21.5) and implies by independence the first asymptotic result.

Since the sample medians in the second case are not independent, we consider their joint distribution using empirical processes. Let us write \( F_\Delta \) for the cumulative distribution function of \( f(\bullet - \Delta) \) and denote by \( B^1, B^2 \) two independent standard Brownian bridges. Then empirical process theory yields by independence

\[
\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} 1([Y_i, \infty)) - F, \frac{1}{n} \sum_{i=n+1}^{2n} 1([Y_i, \infty)) - F_\Delta \right) \Rightarrow (B^1 \circ F, B^2 \circ F_\Delta).4g
\]
The joint median med\((Y_i, i = 1, \ldots, 2n)\) satisfies in terms of the empirical distribution functions \(F^n\) and \(F^n_\Delta\) of the two samples

\[
F^n(\text{med}(Y_i, i = 1, \ldots, 2n)) + F^n_\Delta(\text{med}(Y_i, i = 1, \ldots, 2n)) = 1.
\]

Hence, it can be expressed as the functional \((F^n + F^n_\Delta)^{-1}(1)\) of \((F^n, F^n_\Delta)\), assuming that the inverse is defined properly (e.g. giving the mean of all admissible values). Combining two-dimensional versions of Theorem 20.8 and Lemma 21.4 of van der Vaart (1998), we infer

\[
\sqrt{n}\left(\text{med}(Y_i, i = 1, \ldots, 2n), \text{med}(Y_i, i = 1, \ldots, 2n) - \Delta/2\right)
\]

\[
\Rightarrow \left(-(B^1 \circ f/f) \circ F^{-1}(1/2), -(B^1 \circ f + B^2 \circ f_\Delta)/(f + f(\bullet - \Delta)) \circ (F + f_\Delta)^{-1}(1)\right).
\]

By symmetry of \(f\) the right-hand side simplifies to

\[
\left(-(B^1(1/2)/f(0), -(B^1(F(\Delta/2)) + B^2(F_\Delta(\Delta/2)))/(2f(\Delta/2))\right).
\]

Consequently, \(\sqrt{n}(2(\text{med}(Y_i, i = 1, \ldots, 2n) - \text{med}(Y_i, i = 1, \ldots, n)) - \Delta)\) is asymptotically normal with mean zero and variance

\[
\sigma^2_L = 4\mathbb{E}\left[\left(-B^1(F(\Delta/2)) + B^2(F(-\Delta/2))/(2f(\Delta/2)) + B^1(1/2)/f(0)\right)^2\right]
\]

\[
= 4\frac{F(-\Delta/2)(1 - F(-\Delta/2))}{4f^2(\Delta/2)} + \frac{F(\Delta/2)(1 - F(\Delta/2))}{4f^2(\Delta/2)} + \frac{1}{4f^2(0)} - \frac{1 - F(\Delta/2)}{2f(0)f(\Delta/2)}
\]

\[
= \frac{2F(\Delta/2)(1 - F(\Delta/2))}{f^2(\Delta/2)} + \frac{1}{f^2(0)} - \frac{2(1 - F(\Delta/2))}{f(0)f(\Delta/2)}.
\]

While \(\sigma^2_L = \sigma^2_W\) for \(\Delta = 0\) is straight-forward, we rewrite \(\sigma^2_L\) in terms of \(R = F/f\) to study the behaviour as \(\Delta \to 0\):

\[
\sigma^2_L = 2R(\Delta/2)R(-\Delta/2) + 4R^2(0) - 4R(0)R(-\Delta/2).
\]

Because of \(R(\Delta/2) - R(-\Delta/2) = \Delta + O(\Delta^2)\) by the Lipschitz property of \(f\), we obtain asymptotically

\[
\sigma^2_L = \sigma^2_W + 2\left((R(-\Delta/2) - R(0))^2 + R(-\Delta/2)(R(\Delta/2) - R(-\Delta/2))\right) = \sigma^2_W + O(\Delta^2).
\]

This gives \(\sigma^2_L = \sigma^2_W(1 + 2\Delta f(0) + O(\Delta^2f(0)))\).

### 7.2 Proof of Proposition 4.6

We shall only consider the case of odd \(N = 2m + 1\). Under the conditions of the proposition Brown et al. (2008) show the following result.
7.1 Theorem. For all $m \geq 0$ the sample $\varepsilon_1, \ldots, \varepsilon_{2m+1}$ can be realised on the same probability space as a standard normal random variable $Z$ such that $\text{med}(\varepsilon) := \text{med}(\varepsilon_i, i = 1, \ldots, 2m + 1)$ satisfies

$$\left| \text{med}(\varepsilon) - \frac{Z}{\sqrt{4(2m+1)f_\varepsilon(0)}} \right| \leq \frac{C}{2m+1} \left( 1 + Z^2 \right) \text{ if } |Z| \leq \delta \sqrt{2m + 1},$$

where $\delta, C > 0$ are constants depending on $f_\varepsilon$, but independent of $m$.

The construction and the inequality of the theorem yield with some constant $C' > 0$

$$E[|\text{med}(\varepsilon)|^{2r}] \leq (2^{-r} \lor 1) \left( (4(2m+1)f_\varepsilon(0))^2 - r |Z|^{2r} + \frac{C^{2r}}{(2m+1)^{2r}} \left( 1 + Z^2 \right)^{2r} \right) \leq C'(2m+1)^{-r}.$$

On the other hand, because of $\varepsilon_i \in L^r$ we have for $z \to \infty$ that the cdf satisfies $F_\varepsilon(-z) \lesssim |z|^{-r}$ and $1 - F_\varepsilon(z) \lesssim |z|^{-r}$. From the formula for the density of $\text{med}(\varepsilon)$

$$f_m(z) = \left( \frac{2m+1}{m+1} \right)^m (m+1)f_\varepsilon(z)F_\varepsilon(z)^m(1 - F_\varepsilon(z))^m$$

we therefore infer that $\|\text{med}(\varepsilon)\|_{L^{3r}}$ for $m \geq 2$ is finite and uniformly bounded. Hence, the Hölder inequality gives

$$E[|\text{med}(\varepsilon)|^{2r}1(|Z| \leq \delta \sqrt{2m + 1})] \leq E[|\text{med}(\varepsilon)|^{3r}2/3 |Z| > \delta \sqrt{2m + 1}]^{1/3},$$

which by Gaussian tail estimates is of order $\exp(-\delta^2(2m + 1)/6)$ and thus for $m \to \infty$ asymptotically negligible. This gives the upper moment bound for $\text{med}(\varepsilon)$, the lower bound follows symmetrically. The $r$-th moment is bounded by even simpler arguments.

The second assertion follows via quantile coupling from

$$P\left( \sqrt{4(2m+1)f_\varepsilon(0)}|\text{med}(\varepsilon)| > \tau_m \right) \leq P\left( |Z| + \frac{C}{\sqrt{2m+1}}(1 + Z^2) > \tau_m \right) + P\left( |Z| > \delta \sqrt{2m + 1} \right) \leq P(2|Z| > \tau_m) + P(|Z| > \delta \sqrt{2m + 1}) \leq 2 \exp(-\tau_m^2/8).$$

\[\square\]

References

Arcones, M. A. (2002). Moderate Deviations for $M$-estimators. Test 11(2), 465-500.
ARIA-CASTRO, E. and D. DONoho (2006). Does median filtering truly preserve edges better than linear filtering?, Preprint in Math arXive math/0612422v1.

BROWN, L. D., T. CAI and H. ZHOU (2008). Robust nonparametric estimation via wavelet median regression. Annals of Statistics 36(5), 2055–2084.

CHU, J. T. and H. HOTELLING (1955). The moments of the sample median. Annals of Math. Statistics 26(4), 593–606.

HALL, P. and M. C. JONES (1990). Adaptive $M$-estimation in nonparametric regression. Annals of Statistics 18(4), 1712–1728.

HUBER, P. J. (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics 35(1), 73–101.

KATKOVIUK, V. and V. SPOKOINY (2008). Spatially adaptive estimation via fitted local likelihood techniques. IEEE Transactions on Signal Processing 56(3), 873–886.

KOENKER, R. (2005). Quantile Regression. Econometric Society Monographs 38, Cambridge University Press.

LEPSKII, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. Theory Probab. Appl. 35(3), 454–466. Translated from Teor. Veroyatnost. i Primenen. 35(3) (1990), 459–470.

LEPSKI, O., E. MAMMEN and V. SPOKOINY (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. Annals of Statistics 25(3), 929–947.

MASSART, P. (2005). Concentration inequalities and model selection. Ecole d’Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics 1896, Springer, Berlin.

POLZEH, J. and V. SPOKOINY (2003). Image denoising: pointwise adaptive approach. Annals of Statistics 31, 30–57.

PORTNOY, S. (1997). Local asymptotics for quantile smoothing splines. Annals of Statistics 25(1), 414–434.

ROZENHOLC, Y., M. REISS, D. BALVAY and C.-A. CUENOD(2009). Growing time-homogeneous neighbourhods for denoising and clustering dynamic contrast enhanced-CT sequences, Preprint Université Paris V.

SPOKOINY, V. and C. VIAL (2009). Parameter tuning in pointwise adaptation using a propagation approach, Annals of Statistics, to appear.

TRUONG, Y.K. (1989). Asymptotic Properties of Kernel Estimators Based on Local Medians, Annals of Statistics 17(2), 606–617.

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. Annals of Statistics 32(1) 135–166.
VAN DE GEER, S. (2003). Adaptive quantile regression, in Recent Trends in Nonparametric Statistics (Eds. M.G. Akritas and D.N. Politis), Elsevier Science, 235–250.

VAN DER VAART, A. (1998). Asymptotic Statistics, Cambridge University Press.