SFB 649 Discussion Paper 2011-029

Pointwise adaptive estimation for quantile regression

Markus Reiß*
Yves Rozenholc**
Charles A. Cuenod***

* Humboldt-Universität zu Berlin, Germany
** Université Paris René Descartes, France
*** George Pompidou Hospital Paris, France

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Pointwise adaptive estimation for quantile regression

Markus Reiß * Yves Rozenholc † Charles A. Cuenod ‡

Abstract

A nonparametric procedure for quantile regression, or more generally nonparametric M-estimation, 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 the basis for denoising CT scans in cancer research.

AMS Classification. Primary 62G08; secondary 62G20, 62G35, 62F05, 62P10.
JEL Classification. C14, C31.

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

*Institut für Mathematik, Humboldt-Universität, Unter den Linden 6, D-10099 Berlin, Germany, mreiss@mathematik.hu-berlin.de (corresponding author) Financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin, is gratefully acknowledged.

†Université Paris René Descartes, yves.rozenholc@univ-paris5.fr

‡George Pompidou Hospital Paris, charles-andre.cuenod@egp.aphp.fr

1 Introduction

1.1 Motivation and background

We consider a generalized regression model

$$Y_i = g(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

with (ε_i) i.i.d., $x_1, \dots, x_n$ in the design space $\mathscr{X}$ and $g : \mathscr{X} \to \mathbb{R}$. We shall work conditionally on the design $x_1, \dots, x_n$ such that it can be deterministic or random. The problems we have in view are those of nonparametric quantile estimation and of nonparametric estimation of g using estimators dealing with heavy-tailed noise (ε_i). Both is highly relevant in applications, our main application here will be median regression for 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, assuming

$$\operatorname{argmin}_{m \in \mathbb{R}} \mathbb{E}[\rho(\varepsilon_i - m)] = 0 \tag{1.1}$$

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 $\operatorname{argmin}_{m \in \mathbb{R}} \mathbb{E}[\rho(Y_i - m)]$, 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 ρ' 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 ($\mathbb{P}(\varepsilon_i \leqslant 0) = \mathbb{P}(\varepsilon_i \geqslant 0) = 1/2$) and the intermediate case $\rho(x) = x^2/2$ for $|x| \leqslant k$ and $\rho(x) = k|x| - k^2/2$ for $|x| \geqslant k$ with some $k > 0$ for the Huber estimator ($\mathbb{E}[\min(\max(\varepsilon_i, -k), k)] = 0$). The quantile regression model is obtained for $\rho(x) = |x| + (2\alpha - 1)x$ ($\mathbb{P}(\varepsilon_i \leqslant 0) = \alpha$ with quantile $\alpha \in (0, 1)$), see e.g. Koenker (2005). Let us mention that the framework also contains local-likelihood location estimation when ρ is taken as the negative log-likelihood with the well-known special cases of mean regression for Gaussian noise and median regression for Laplace noise. Note that the convexity of ρ is in this case equivalent to a log-concave error distribution.

The function g is not supposed to satisfy a global smoothness criterion, but we aim at estimating it locally in each point $x \in \mathscr{X}$ 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. Their drawback is that they do not perform well for spatially inhomogeneous signals like images with edges. In Yu and Jones (1998) local linear quantile regression is proposed with local bandwidth choices based on heuristic Gaussian approximations, yet lacking a theoretical foundation. 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.

1.2 Main results

Subsequently, we assume that ρ is fixed according to the problem at hand and we use the corresponding sample versions to construct base estimators for the (generalized) regression function g. Since we shall care about robustness, we measure the error in pointwise L^r-moment loss for any $r > 0$.
In the spirit of classical nonparametrics, we implicitly assume that g can be approximated by a constant locally around a given point $x \in \mathscr{X}$. The statistical challenge is to select adaptively the right neighbourhood U of x where a local M-estimator is applied. Let us write

$$m(Y_i,\, x_i \in U) := \operatorname{arginf}_{\mu\in\mathbb{R}} \Big\{ \sum_{i:x_i\in U} \rho(Y_i - \mu) \Big\} \tag{1.2}$$

for the location estimator on the set $U \subseteq \mathscr{X}$. 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

$$\tilde\vartheta_k := m(Y_i,\, x_i \in U_k). \tag{1.3}$$

Then the family $(\tilde\vartheta_k)_{0\leqslant k\leqslant K}$ forms the class of base estimators and we aim at selecting the best estimator of $\vartheta := g(x)$ in this family. Note that in general we shall use the design (x_i) and the point x in order to define the (U_k), ensuring that U_0 and all $U_{k+1} \setminus U_k$ contain at least one point x_i. The leading example throughout this paper will be the classical windowed median filter as follows.

1.1 Example. Let the design space be $\mathscr{X} = [0,1]$ 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 $0 \leqslant h_0 < h_1 < \cdots < h_K$. Then $\tilde\vartheta_k$ is the median filter, e.g. studied by Truong (1989) or Arias-Castro and Donoho (2009).

Specified to this example with a geometric grid of bandwidths (h_k), our main non-asymptotic risk bound in Theorem 3.8 yields for our data-driven choice $\hat k$ the following deviation bound from an oracle-type choice k^*, cf. Example 3.10:

$$\mathbb{E}_g[|\tilde\vartheta_{\hat k} - \tilde\vartheta_{k^*}|^r] \leqslant C(\alpha + z_{k^*}^r)\, \mathbb{E}_0[|\tilde\vartheta_{k^*} - 0|^r],$$

where $C > 0$ is some (explicitly computable) constant, $\alpha > 0$ a parameter of the method and z_{k^*} a (known) critical value. This bound reflects the philosophy of the method: In the excess risk bound for the data-driven choice $\hat{k}$ we pay at most a multiple of the oracle stochastic error, i.e. the risk when $g = 0$.
While the non-asymptotic result is very general, it does not concisely convey the size of the error. Therefore, we quantify the asymptotic risk in the setting of equidistant design $x_i = i/n$, a locally s-Hölder continuous function g and some mild assumptions on the possible bandwidths as

$$\mathbb{E}_g[|\tilde{\vartheta}_{\hat{k}} - g(x)|^r]^{1/r} \leqslant C'(\log(n)/n)^{s/(2s+1)}$$

with a constant $C' > 0$, see Theorem 4.7. This is the minimax adaptive rate even for Gaussian errors, but holds true for any zero median noise (ε_i) with finite moments of order $r > 0$. In particular, the so-called payment for pointwise adaptation is logarithmic in n, which is necessary even in the Gaussian setup, and does not increase for heavy-tailed noise. This new and surprising result is due to the robustness of the median and relies technically on moderate deviation bounds. The general approach presented here permits to extend these adaptive minimax results to general nonparametric M-estimation. This is Corollary 4.4 where mainly the moderate deviation bound in Assumption 4.1(c) needs to be checked, see also Section 4.2 for a discussion.

1.3 Outline

We present our procedure to select optimally among local M-estimators in Section 2. In Sections 3 and 4 we derive exact and asymptotic error bounds and the latter give optimal minimax rates for adaptive pointwise risk over Hölder classes. Specified to median regression, Theorem 4.7 gives optimal locally adaptive risk bounds under almost no conditions on the noise distribution. 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. The fundamental idea of our approach is developed from a toy model of testing in the Appendix, Section 7, where also a technical proof has been moved.

2 The procedure

2.1 Main ideas

Let us start by considering the Lepski (1990) method for selecting among $(\tilde{\vartheta}_k)_{0 \leqslant k \leqslant K}$, given the mean regression model with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] < \infty$. Note that the base estimators are then ordered according to decreasing variances: $\mathrm{Var}(\tilde{\vartheta}_k) \leqslant \mathrm{Var}(\tilde{\vartheta}_{k-1})$. On the other hand, the bias is usually increasing with increasing neighbourhoods U_k. This is not always the case (for example, think of local means for a linear g), but true in particular for the worst case bias over

smoothness classes like Hölder balls of functions. Lepski's method can be understood as a multiple testing procedure where the hypothesis $H_0(k):\ g|_{U_k} \equiv \vartheta$ 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 $\ell \leqslant k$. Once the test of an $H_0(k+1)$ is rejected, we select the base estimator $\tilde{\vartheta}_k$ 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 $|\tilde{\vartheta}_{k+1} - \tilde{\vartheta}_\ell| \leqslant z_\ell^{(k+1)}$ holds for all $\ell \leqslant k$ with suitable critical values $z_\ell^{(k+1)} > 0$.

In practice, Lepski's method often leads to oversmoothing, which is even more conspicuous for robust M-estimators where the change in the signal is noticed far too late, cf. Section 5. A very intuitive idea is therefore to replace the test for $H_0(k+1):\ g|_{U_{k+1}} \equiv \vartheta$ at stage $k \geqslant 1$ by a test for $H_0(k+1):\ g|_{U_{k+1}\setminus U_k} \equiv \vartheta$. Provided $H_0(\ell)$, $\ell \leqslant k$, have been accepted, both test descriptions amount to the same assertion. The new description, however, leads in the case of M-estimators to more powerful constructions of the decision procedure, based on an estimator for g on the difference $U_{k+1} \setminus U_k$ between consecutive neighbourhoods: 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_ℓ for $\ell \leqslant k$. This means that our tests reject if the empirical location in the additional data

$$\tilde{\vartheta}_{(k+1)\setminus k} := m(Y_i,\, x_i \in U_{k+1} \setminus U_k) \tag{2.1}$$

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

$$\exists \ell \leqslant k:\ |\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_\ell| > z_\ell^{(k+1)}. \tag{2.2}$$

It is necessary to perform the testing for all $\ell \leqslant k$ 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 $\tilde{\vartheta}_{(k+1)\setminus k}$ 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 $\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_k$ 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 method, whereas the generally nonlinear M-estimators are treated in a significantly different way. More insight why our tests for $H_0(k+1)$ are often more powerful and natural is provided by a two-sample toy model which is discussed in Appendix 7.

2.2 The selection algorithm

We aim at selecting the best estimator among the family $\{\tilde{\vartheta}_k \,|\, k = 0, \dots, K\}$. Considering the law $\mathbb{P}_0$ generated by the pure noise setting $g \equiv 0$, we introduce the stochastic error levels

$$s_j := \mathbb{E}_0[|\tilde{\vartheta}_j|^r]^{1/r}, \quad s_{kj} := \mathbb{E}_0[|\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j|^r]^{1/r}. \tag{2.3}$$

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

- `initialize` $k := 0$;
- `repeat`
  `if for all` $j = 0, \dots, k$

$$|\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j| \leqslant z_j s_{kj} + z_{k+1} s_{k+1}$$

  `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:

$$\hat{k} := \inf\left\{k \geqslant 0 \,\middle|\, \exists j \leqslant k : \ |\tilde{\vartheta}_{(k+1)\setminus k} - \tilde{\vartheta}_j| > z_j s_{kj} + z_{k+1} s_{k+1}\right\} \wedge K, \quad \hat{\vartheta} := \tilde{\vartheta}_{\hat{k}}. \tag{2.4}$$

3 Error analysis

We analyse the excess risk between our data-driven estimator $\hat{\vartheta}$ and an optimal oracle-type estimator $\tilde{\vartheta}_{k^*}$. In our sequential procedure the two errors of stopping too early or late are estimated separately and then put together. While stopping late is controlled by the construction of the stopping rule, the error of stopping early is bounded using a prescription of the critical values in the pure noise situation. The asymptotic analysis in the next section will then provide the order of the quantities when $n \to \infty$.

3.1 Propagation and stopping late

We need a very natural monotonicity 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:*

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

Moreover, we have for any $c, c_i \geqslant 0$:

$$m(Y_i + c_i,\, x_i \in S) \geqslant m(Y_i,\, x_i \in S), \quad m(Y_i + c,\, x_i \in S) = m(Y_i,\, x_i \in S) + c.$$

For any reasonable specific choice Assumption 3.1 should be satisfied. In particular, this is true for the sample quantile where we take for N data points Y_i the mean $(Y_{(\lfloor \alpha N \rfloor)} + Y_{(\lfloor \alpha N \rfloor + 1)})/2$ of the order statistics (a malevolent choice would be for example to choose the first order statistic among the two that is irrational). If ρ is strictly convex, then the M-estimator is uniquely defined and the properties always hold:

3.2 Lemma. *If the function ρ is strictly convex, then Assumption 3.1 is satisfied.*

Proof. Starting with the first inequality, let us write m_T as short-hand for $m(Y_i,\, x_i \in T)$, $T \subseteq \mathscr{X}$. Denoting by ρ'_+, ρ'_- the right- and left-handed derivatives of the convex function ρ, the functions ρ'_+, ρ'_- are strictly increasing with $\rho'_+(x) < \rho'_-(y) \leqslant \rho'_+(y)$ for all $x < y$ and

$$\sum_{x_i \in T} \rho'_-(Y_i - m_T) \leqslant 0, \quad \sum_{x_i \in T} \rho'_+(Y_i - m_T) \geqslant 0.$$

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

$$\sum_{x_i \in S} \rho'_-(Y_i - m_S) > \sum_j \sum_{x_i \in S_j} \rho'_+(Y_i - m_{S_j}) \geqslant 0,$$

which contradicts the minimizing property of m_S. Hence, $m_S \geqslant \min_j m_{S_j}$ holds and a symmetric argument shows $m_s \leqslant \max_j m_{S_j}$.
For the inequality in the second display we have by definition and by strict monotonicity of ρ'_+

$$\sum_{x_i \in S} \rho'_+(Y_i + c_i - m_S) \geqslant \sum_{x_i \in S} \rho'_+(Y_i - m_S) \geqslant 0.$$

This implies $m(Y_i + c_i,\, x_i \in S) \geqslant m_S$. The shift identity in this display follows by uniqueness immediately from the definition. □

3.3 Proposition. *Grant Assumption 3.1 and consider $\hat{\vartheta}$ from* (2.4)*. Then we have for any $k = 0, \dots, K-1$*

$$|\hat{\vartheta} - \tilde{\vartheta}_k| \mathbf{1}(\hat{k} > k) \leqslant \max_{j=k+1,\dots,K-1} \big(z_k s_{jk} + z_{j+1} s_{j+1}\big).$$

3.4 Remark. This 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, see also Lepski (1990). It is intuitive because when we stop later than at k^*, then only because the local M-estimators have not differed significantly from $\tilde{\vartheta}_{k^*}$; the propagation of the error remains under control.

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

$$|\tilde{\vartheta}_\ell - \tilde{\vartheta}_k| \leqslant \max_{k+1 \leqslant j \leqslant \ell} |\tilde{\vartheta}_{j \setminus (j-1)} - \tilde{\vartheta}_k|.$$

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

$$|\hat{\vartheta} - \tilde{\vartheta}_k| \leqslant \max_{j=k+1,\dots,\hat{k}} |\tilde{\vartheta}_{j \setminus (j-1)} - \tilde{\vartheta}_k| \leqslant \max_{j=k+1,\dots,K-1} \big(z_k s_{jk} + z_{j+1} s_{j+1}\big).$$

□

3.5 Example. For geometrically decreasing stochastic error levels s_k in (2.3), 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 $A(p) \lesssim B(p)$ means $A(p) \leqslant cB(p)$ with a constant $c > 0$ independent of the parameters p involved. The late stopping error is of order $z_k^r s_k^r$, 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 $z_{k^*}^r$:

$$\mathbb{E}_g[|\hat{\vartheta} - \vartheta|^r \mathbf{1}(\hat{k} > k^*)] \lesssim \mathbb{E}_g[|\tilde{\vartheta}_{k^*} - \vartheta|^r] + z_{k^*}^r s_{k^*}^r \leqslant (1 + z_{k^*}^r)\, \mathbb{E}_g[|\tilde{\vartheta}_{k^*} - \vartheta|^r].$$

The inequality $\mathbb{E}_0[|\tilde{\vartheta}_{k^*}|^r] \leqslant \mathbb{E}_g[|\tilde{\vartheta}_{k^*} - \vartheta|^r]$ used here means that the stochastic error ($g = 0$) is less than the overall error. Its lengthy proof is based on the explicit median distribution for $(g(x_i) + \varepsilon_i)_{x_i \in U_{k^*}}$, but is omitted.

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 pure noise situation under $\mathbb{P}_0$, 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 $z_k > 0$, $k = 0, \dots, K-1$, such that the condition

$$\sum_{j=0}^{K-1} \mathbb{E}_0 \left[|\tilde{\vartheta}_j|^r \mathbf{1}\big(\exists \ell \leqslant j : \ |\tilde{\vartheta}_{(j+1)\setminus j} - \tilde{\vartheta}_\ell| > z_\ell s_{j\ell}\big) \right] \leqslant \alpha s_K^r \tag{3.1}$$

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

$$\sum_{j=0}^{K-1} \mathbb{E}_0 \left[|\tilde{\vartheta}_j|^r \mathbf{1}\big(|\tilde{\vartheta}_{(j+1)\setminus j} - \tilde{\vartheta}_0| > z_0 s_{j0}\big) \right] \leqslant \tfrac{\alpha}{K} s_K^r$$

and then each z_k for given $z_0, \dots, z_{k-1}$ such that

$$\sum_{j=k}^{K-1} \mathbb{E}_0 \left[|\tilde{\vartheta}_j|^r \mathbf{1}\Big(|\tilde{\vartheta}_{(j+1)\setminus j} - \tilde{\vartheta}_k| > z_k s_{jk},\, \forall \ell < k : \ |\tilde{\vartheta}_{(j+1)\setminus j} - \tilde{\vartheta}_\ell| \leqslant z_\ell s_{j\ell}\Big) \right]$$
$$\leqslant \tfrac{\alpha}{K} s_K^r. \tag{3.2}$$

Summing the left-hand sides over $k = 0, \dots, K-1$, we obtain exactly (3.1). 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 equals (at most) $\frac{\alpha}{K} s_K^r$. 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). The idea is to transfer properties from the pure noise 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 there the moment order is enlarged from r to $2r$ in the calibration step, while here the term $z_{k+1}s_{k+1}$ is added to the testing values $z_j s_{kj}$ from the calibration. In the asymptotic analysis we do not lose powers in the logarithmic factor, as is the case for the small modeling bias ansatz, and we 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.
Let us also mention that in the case of a misspecified noise distribution (which is at the origin of robust statistics) Condition (3.1) holds with some other parameter α. The non-asymptotic risk bounds will show a linear dependence in α so that the error can still be controlled. The asymptotic analysis in Proposition 4.3 will provide admissible critical values $z_k \sim \sqrt{\log n}$ under moment and moderate deviation bounds on the M-estimator over $(\varepsilon_i)_{1 \leqslant i \leqslant N}$ for large N which for bounded influence functions is a very weak requirement, see Section 4.2 below.

3.6 Definition. Given the regression function g, introduce its variation on U_k

$$\mathrm{V}_k(g) := \sup_{y_1, y_2 \in U_k} |g(y_1) - g(y_2)|$$

and consider the oracle-type index

$$k^* := \min\{k = 0, \dots, K-1 \mid \mathrm{V}_{k+1}(g) > z_{k+1}s_{k+1}\} \wedge K.$$

This definition implies that for all $k \leqslant k^*$ the maximal bias $V_k(g)$ of $\tilde{\vartheta}_k$ is less than its stochastic error level s_k from (2.3) 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 $k = K - 1$.

3.7 Proposition. *Grant Assumption 3.1 and use the definition in* (2.4) *and the condition in* (3.1) *for some* $\alpha > 0$. *Then we have for any* $k = 0, \dots, k^*$ *and* $r > 0$ *such that all error levels in* (2.3) *are finite:*

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

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

$$|\tilde{\vartheta}_j(g) - \tilde{\vartheta}_k(g)| \leqslant |\tilde{\vartheta}_j(0) - \tilde{\vartheta}_k(0)| + V_k(g) \text{ for } j < k \tag{3.3}$$

which follows by Assumption 3.1 from

$$\begin{aligned}\tilde\vartheta_j(g) - \tilde\vartheta_k(g) &= m(g(x_i)+\varepsilon_i,\, x_i \in U_j) - m(g(x_i)+\varepsilon_i,\, x_i \in U_k)\\ &\leqslant m\Big(\varepsilon_i + \sup_{x\in U_j} g(x),\, x_i \in U_j\Big) - m\Big(\varepsilon_i + \inf_{x\in U_k} g(x),\, x_i \in U_k\Big)\\ &\leqslant m(\varepsilon_i,\, x_i \in U_j) + \sup_{x\in U_j} g(x) - m(\varepsilon_i,\, x_i \in U_k) - \inf_{x\in U_k} g(x)\\ &\leqslant \tilde\vartheta_j(0) - \tilde\vartheta_k(0) + V_k(g)\end{aligned}$$

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 $\tilde\vartheta_{(j+1)\setminus j}$, we obtain for all $k \leqslant k^*$

$$\begin{aligned}&\mathbb{E}\left[|\hat\vartheta(g) - \tilde\vartheta_k(g)|^r \mathbf{1}(\hat k(g) < k)\right]\\ &= \sum_{j=0}^{k-1} \mathbb{E}\left[|\tilde\vartheta_j(g) - \tilde\vartheta_k(g)|^r \mathbf{1}(\hat k(g) = j)\right]\\ &\leqslant \sum_{j=0}^{k-1} \mathbb{E}\left[(\mathrm{V}_k(g) + |\tilde\vartheta_j(0)| + |\tilde\vartheta_k(0)|)^r \mathbf{1}(\hat k(g) = j)\right]\\ &\leqslant (3^{r-1}\vee 1)\Big(\mathrm{V}_k(g)^r + \mathbb{E}[|\tilde\vartheta_k(0)|^r] +\\ &\quad \sum_{j=0}^{k-1} \mathbb{E}\left[|\tilde\vartheta_j(0)|^r \mathbf{1}\big(\exists \ell \leqslant j : |\tilde\vartheta_{(j+1)\setminus j}(g) - \tilde\vartheta_\ell(g)| > z_\ell s_{j\ell} + z_{j+1}s_{j+1}\big)\right]\Big)\\ &\leqslant (3^{r-1}\vee 1)\Big(z_k^r s_k^r + s_k^r +\\ &\quad \sum_{j=0}^{k-1} \mathbb{E}\left[|\tilde\vartheta_j(0)|^r \mathbf{1}\big(\exists \ell \leqslant j : |\tilde\vartheta_{(j+1)\setminus j}(0) - \tilde\vartheta_\ell(0)| + V_{j+1}(g) > z_\ell s_{j\ell} + z_{j+1}s_{j+1}\big)\right]\Big)\\ &\leqslant (3^{r-1}\vee 1)\Big(z_k^r s_k^r + s_k^r + \alpha s_K^r\Big).\end{aligned}$$

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

3.3 Total risk bound

3.8 Theorem. *Consider $\hat\vartheta$ from (2.4) and critical values (z_k) such that Condition (3.1) is satisfied and $(z_k s_k)$ is non-increasing in k. Then under Assumption 3.1 the following excess risk estimate holds for all $k \leqslant k^*$ and $r > 0$ such that all error levels in (2.3) are finite:*

$$\mathbb{E}_g[|\hat\vartheta - \tilde\vartheta_k|^r] \leqslant (3^{r-1}\vee 1)\Big((2z_k^r + 1 + \alpha)s_k^r + z_k^r \max_{j=k+1,\dots,K-1} s_{jk}^r\Big).$$

Proof. For the late-stopping error Proposition 3.3 and the decay of $(z_k s_k)$ give

$$\begin{aligned}|\hat\vartheta - \tilde\vartheta_k|^r \mathbf{1}(\hat k > k) &\leqslant (2^{r-1}\vee 1)\max_{j>k}(z_k^r s_{jk}^r + z_{j+1}^r s_{j+1}^r)\\ &\leqslant (2^{r-1}\vee 1) z_k^r\big(s_k^r + \max_{j>k} s_{jk}^r\big).\end{aligned}$$

Add the early-stopping error from Proposition 3.7. □

3.9 Remarks.

(a) The decay of $(z_k s_k)$ is imposed to facilitate the result. Moreover, it is quite natural since s_k is of the order $|U_k|^{-1/2}$ ($|U_k|$: number of observations in U_k) while the critical values z_k usually decrease as well, compare Proposition 4.3 and the discussion thereafter.

(b) Concerning the moment $r > 0$ required, we remark that r is in general much larger than the maximal finite moment of the noise (ε_i). In the proof of Proposition 4.6 for example we derive that the median has finite moments three times larger than (ε_i) for sample size at least five, which for the Cauchy distribution would already give $r \approx 3$!

3.10 Example (continued)**.** For geometrically increasing bandwidths (h_k) we obtain $s_{jk} \lesssim s_k$ for $j > k$ and thus

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

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 and z_k instead of $U_k(n)$, $K(n)$, $z_k(n)$. The notation $A \lesssim B$ will always mean $A(n) \leqslant cB(n)$ 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:

$$q_1 N_k \leqslant N_{k+1} \leqslant q_2 N_k \quad \text{for all } k = 0, \dots K-1$$

for some fixed $q_2 \geqslant q_1 > 1$ 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

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

(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

$$\limsup_{N \to \infty} e^{c\tau_N^2}\, \mathbb{P}\left(N^{1/2} |m(\varepsilon_i,\, i = 1, \dots, N)| > \tau_N\right) < \infty.$$

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* $k \geqslant j$

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

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*

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

with $\zeta > 0$ *a sufficiently large constant in order to satisfy Condition* (3.2). *For* $K \sim \log n$ *this yields asymptotically* $z_k \sim \sqrt{\log n}$.

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

$$\begin{aligned}
&\mathbb{P}_0(|\tilde\vartheta_{(j+1)\setminus j} - \tilde\vartheta_k| > z_k s_{jk}) \\
&\leqslant \mathbb{P}_0(|\tilde\vartheta_{(j+1)\setminus j}| > z_k s_{jk}/2) + \mathbb{P}_0(|\tilde\vartheta_k| > z_k s_{jk}/2) \\
&\lesssim \exp(-cz_k^2 s_{jk}^2 (N_{j+1} - N_j)/4) + \exp(-cz_k^2 s_{jk}^2 N_k/4).
\end{aligned}$$

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

$$\mathbb{P}_0(|\tilde\vartheta_{(j+1)\setminus j} - \tilde\vartheta_k| > z_k s_{jk}) \lesssim \exp(-c' z_k^2).$$

Our choice of z_k with ζ 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:

$$\begin{aligned}
&\sum_{j=k}^{K-1} \mathbb{E}_0\left[|\tilde\vartheta_j|^r \mathbf{1}(|\tilde\vartheta_{(j+1)\setminus j} - \tilde\vartheta_k| > z_k s_{jk})\right] \\
&\qquad \leqslant \sum_{j=k}^{K-1} \mathbb{E}_0\left[|\tilde\vartheta_j|^{2r}\right]^{1/2} \mathbb{P}_0(|\tilde\vartheta_{(j+1)\setminus j} - \tilde\vartheta_k| > z_k s_{jk})^{1/2} \\
&\qquad \lesssim \sum_{j=k}^{K-1} s_j^r \exp(-c' z_k^2/2) \\
&\qquad = o\big((K-k) s_k^r \alpha (s_K/s_k)^r K^{-2}\big) \\
&\qquad = o\Big(\frac{\alpha s_K^r}{K}\Big).
\end{aligned}$$

For $K \sim \log N$ we obtain $\log(N_K/N_k) \leqslant (K-k)\log q_2 \lesssim \log N$ and thus $z_k^2 \sim \log n$. □

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 3.8. From Theorem 3.8 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*

$$\mathbb{E}_g[|\hat{\vartheta} - g(x)|^r] \lesssim (\log n)^{r/2}\, \mathbb{E}_g[|\tilde{\vartheta}_{k^*} - \vartheta|^r].$$

4.5 Example (continued)**.** Let us specify to s-Hölder continuous $g : [0,1] \to \mathbb{R}$, 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 $\mathrm{V}_{k^*}(f)^2 \sim h_{k^*}^{2s} \sim (nh_{k^*})^{-1} \log(n)$, that is $h_{k^*} \sim (\log(n)/n)^{1/(2s+1)}$ and $z_{k^*} s_{k^*} \sim (\log(n)/n)^{s/(2s+1)}$.

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 ε_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) ρ is Lebesgue-almost everywhere differentiable with derivative ρ';

(c) there are $\lambda, \delta > 0$ such that $\mathbb{E}[\exp(\lambda|\rho'(\varepsilon_i)|)]$ and $\mathbb{E}[\exp(\lambda \sup_{|h|\leqslant\delta}|\rho(\varepsilon_i + h) - \rho(\varepsilon_i) - h\rho'(\varepsilon_i)|/h)]$ 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 ε_i is required. For median regression the result applies with $V = f_\varepsilon(0)/2$ and $\rho'(\varepsilon_i) = \operatorname{sgn}(\varepsilon_i)$ and because of $||\varepsilon_i + h| - |\varepsilon_i| - h \operatorname{sgn}(\varepsilon_i)| \leqslant 2h$ 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 base 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_ε continuously differentiable around zero, $f_\varepsilon(0) > 0$, $r \in \mathbb{N}$ and $Z \sim N(0,1)$:

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

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 ε_i have a Lebesgue density f_ε 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 $\mathrm{med}(\varepsilon) := \mathrm{med}(\varepsilon_1, \ldots, \varepsilon_N)$, N odd, we have*

$$\forall N \geqslant 5 : \ \mathbb{E}[|\mathrm{med}(\varepsilon)|^r] \sim N^{-r/2} \ \textit{and} \ \mathbb{E}[|\mathrm{med}(\varepsilon)|^{2r}] \sim N^{-r}$$

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

$$\limsup_{N\to\infty} \mathbb{P}\left(2N^{1/2} f_\varepsilon(0)|\mathrm{med}(\varepsilon)| > \tau_N\right) \exp(\tau_N^2/8) \leqslant 2.$$

Specifying to the median regression case, we gather the results obtained in Corollary 4.4, Example 4.5 and Proposition 4.6 to obtain the following asymptotic optimality result.

4.7 Theorem. *Suppose that $g : [0,1] \to \mathbb{R}$ is C^s-Hölder continuous with $s \in (0,1]$ in a neighbourhood of $x \in (0,1)$, that the design is equidistant $x_i = i/n$ and that the error variables ε_i have median zero (median regression model). Choose a geometrical range of bandwidths $h_k = h_0 q^k$ with some $h_0 > 0$, $q > 1$, $k = 0, \ldots, K$ and $K \sim \log n$ and consider the adaptive choice $\hat{\vartheta}_n$ among the corresponding local median filters $\tilde{\vartheta}_k$, applied in windows of size h_k around x. Then for any $r > 0$ such that $\mathbb{E}[|\varepsilon_i|^r] < \infty$ we have*

$$\mathbb{E}_g[|\hat{\vartheta}_n - g(x)|^r]^{1/r} \lesssim (\log(n)/n)^{s/(2s+1)}.$$

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. It should be stressed that also for heavy-tailed distributions we only have to pay a logarithmic factor to achieve local adaptivity. This is due to the fact that the sample median as robust statistics concentrates well around the true location.

5 Simulation results

We illustrate our procedure by an implementation for median regression on $\mathscr{X} = [-1, 1]$ and the estimation of the regression function at $x = 0$. We simulate $n = 200$ equidistant observations (Y_i) with standardized errors (ε_i) ($\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = 1$) 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 $\lfloor 5^k/4^{k-1} \rfloor$ 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

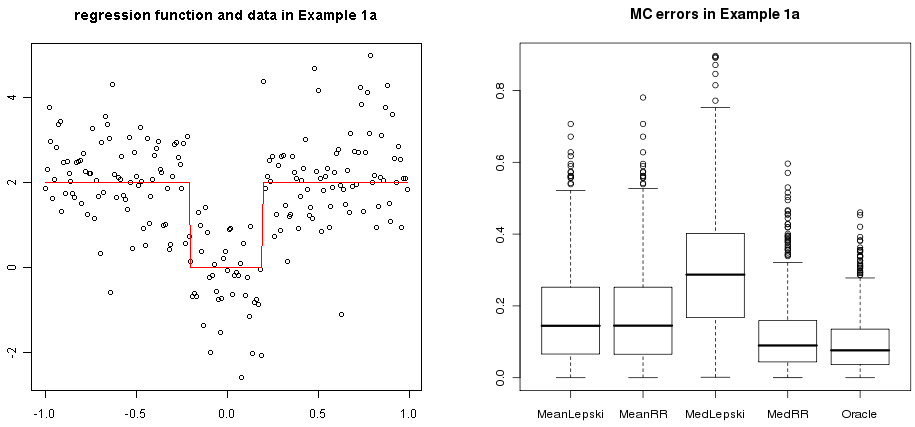

Figure 1: Example 1 with Laplace noise: A typical realisation and a box plot of the sample errors in 1000 Monte Carlo runs.

ζ 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 $|x| \leqslant 0.2$ and $g(x) = 2$ for $|x| > 0.2$, which can be considered as a toy model for edge detection in image restauration 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 our method (RR), local medians with Lepski's and with our method (RR) 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

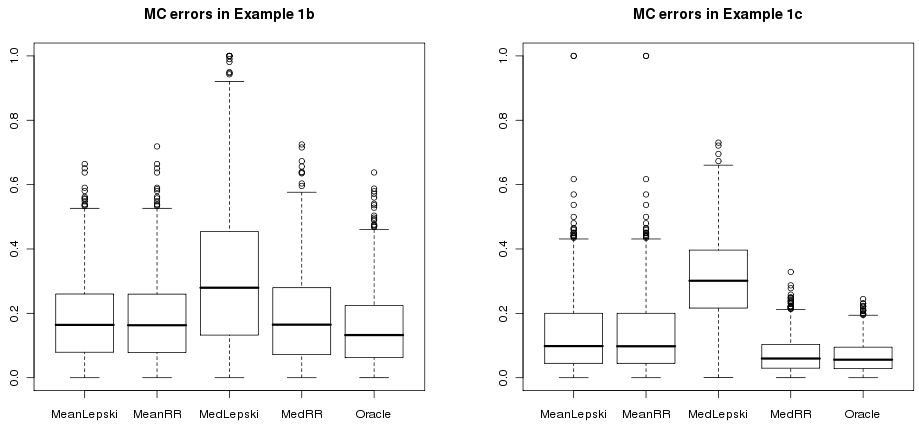

Figure 2: Box plot of the sample errors in 1000 Monte Carlo runs for Gaussian (left) and Student t(3) noise (right).

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, 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 (our method in capitals), each time stating the Monte Carlo median of the absolute error:

Ex.	Mean Lepski	MEAN	Median Lepski	MEDIAN	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

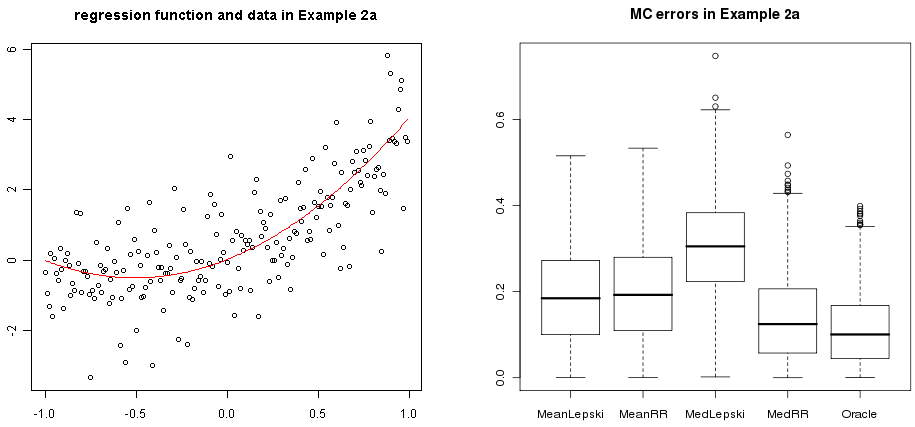

Figure 3: Example 2 with Laplace noise. A typical realisation and a box plot of the sample errors in 1000 Monte Carlo runs.

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 $g : \mathscr{X} \to \mathbb{R}^K$ is observed under vector-valued noise ε_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 signif-

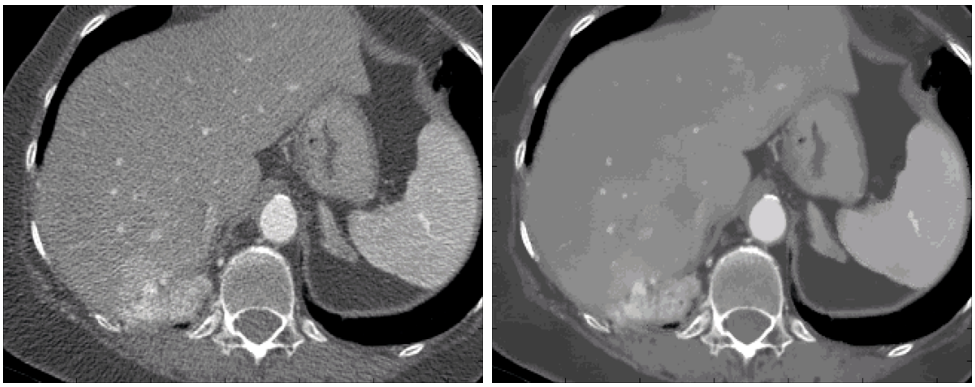

Figure 4: Axial CT image of the abdomen centered on a liver metastasis. This image is part of a dynamic series, it has been acquired after 25 seconds, during the arterial phase. Left: original. Right: denoised version.

icantly 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 512×512-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 Analysis of a toy model

To provide some insight why our tests for $H_0(k+1)$ are often more powerful and natural, let us consider a toy model of two neighbourhoods $U_1 \subseteq U_2$ with a piecewise constant quantile regression function g equal to μ_1 on U_1 and to μ_2 on $U_2 \backslash U_1$. The procedure therefore reduces to a test of the hypothesis $H_0 : \mu_2 = \mu_1$. Putting $\Delta = \mu_2 - \mu_1$ we face a two-sample location test on $H_0 : \Delta = 0$ where the first sample $Y_1, \ldots, Y_n$ (from U_1) is i.i.d. with density f and the

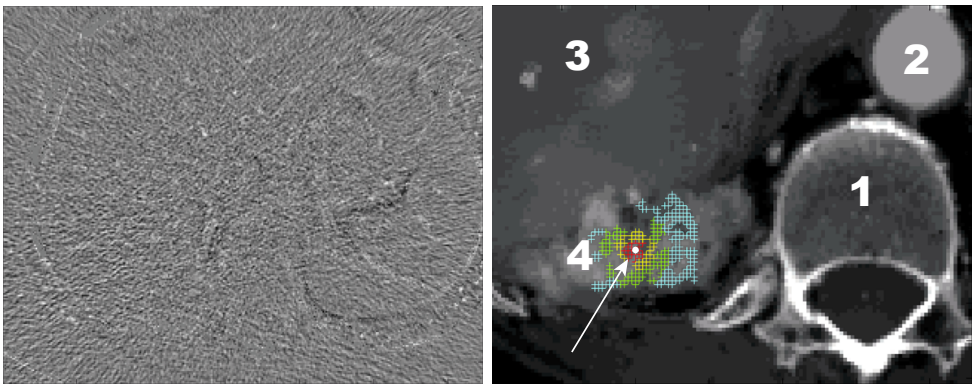

Figure 5: Left: residuals as difference between the original and the denoised version (see Figure 4). Right: zoom into the denoised version showing a vertebra (1), the aorta (2), the liver (3) and a tumor (4). Around a selected voxel, marked by the white dot/arrow, the first four data-driven neigbourhoods (in order: red, yellow, green and blue) are shown.

second independent sample $Y_{n+1}, \dots, Y_{2n}$ (from $U_2 \setminus U_1$) is i.i.d. with density $f(\bullet - \Delta)$. We introduce $\tilde{m}_1 := \hat{q}_\alpha(Y_i,\, i = 1, \dots, n)$, $\tilde{m}_2 := \hat{q}_\alpha(Y_i,\, i = 1, \dots, 2n)$ and $\tilde{m}_{2\setminus 1} := \hat{q}_\alpha(Y_i,\, i = n+1, \dots, 2n)$, the sample quantiles over the first, joint and second sample, respectively. The test in Lepski's method rejects if $T_L := 2|\tilde{m}_1 - \tilde{m}_2|$ is large, while we propose to reject if $T_W := |\tilde{m}_1 - \tilde{m}_{2\setminus 1}|$ is large.
The rationale for our approach is that the sample α-quantile is the maximum likelihood estimator for the location family with $f(x) = 2\alpha(1-\alpha)e^{-|x|+(2\alpha-1)x}$. Classical testing theory then reveals our test as the Wald test using the maximum likelihood estimator for the location and establishes its asymptotic efficiency (a similarly efficient likelihood ratio test is less explicit here). For general noise distributions this is only a pseudo-likelihood approach, but we are still able to derive its superiority.

7.1 Proposition. *Let $f : \mathbb{R} \to \mathbb{R}^+$ be a Lipschitz continuous density, F its distribution function and consider the α-quantile q_α, $\alpha \in (0,1)$. Assume that $f(q_\alpha) > 0$ and $f(q_\alpha + h) + f(q_\alpha - h) - 2f(q_\alpha) = o(h)$ hold for $h \to 0$ (e.g. f differentiable or just symmetric at q_α). If $Y_1, \dots, Y_n \sim f$, $Y_{n+1}, \dots, Y_{2n} \sim f(\bullet - \Delta)$ are independently distributed, then we obtain for $n \to \infty$*

$$\sqrt{n}\big(\tilde{m}_{2\setminus 1} - \tilde{m}_1 - \Delta\big) \Rightarrow N(0, \sigma_W^2), \quad \sqrt{n}\big(2\big(\tilde{m}_2 - \tilde{m}_1\big) - \Delta\big) \Rightarrow N(0, \sigma_L^2)$$

with variances $\sigma_W^2 = \frac{2\alpha(1-\alpha)}{f^2(q_\alpha)}$ and $\sigma_L^2 = \sigma_W^2 + \frac{2\alpha}{f(q_\alpha)}|\Delta| + o(\Delta)$ for $\Delta \downarrow 0$, $\sigma_L^2 = \sigma_W^2 + \frac{2(1-\alpha)}{f(q_\alpha)}|\Delta| + o(\Delta)$ for $\Delta \uparrow 0$.

Proof. The asymptotic normality of the sample quantile $\sqrt{n}(\hat{q}_\alpha(Y_1, \dots, Y_n) - q_\alpha) \Rightarrow N(0, \alpha(1-\alpha)/f^2(q_\alpha))$ is well known (van der Vaart 1998, Corollary 21.5) and implies by independence the first asymptotic result.
Since the sample quantiles in the second case are not independent, we consider their joint distribution using empirical processes. Let us write F_Δ for the cumulative distribution function of $f_\Delta = f(\bullet - \Delta)$ and denote by B^1, B^2 two

independent standard Brownian bridges. Then empirical process theory yields by independence

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}([Y_i,\infty)) - F, \frac{1}{n}\sum_{i=n+1}^{2n}\mathbf{1}([Y_i,\infty)) - F_\Delta\Big) \Rightarrow (B^1\circ F, B^2\circ F_\Delta).$$

The joint sample quantile $\hat{q}_\alpha(Y_i, i=1,\dots,2n)$ satisfies in terms of the empirical distribution functions F^n and F^n_Δ of the two samples

$$\frac{F^n + F^n_\Delta}{2}\Big(\hat{q}_\alpha(Y_i, i=1,\dots,2n)\Big) = \alpha.$$

Hence, it can be expressed as the functional $(F^n + F^n_\Delta)^{-1}(2\alpha)$ of (F^n, F^n_Δ), 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

$$\begin{aligned}&\sqrt{n}\Big(\hat{q}_\alpha(Y_i,\, i=1,\dots,n) - q_\alpha, \hat{q}_\alpha(Y_i,\, i=1,\dots,2n) - q_\alpha - \Delta/2\Big) \Rightarrow \\ &\Big(- (B^1\circ F/f)\circ F^{-1}(\alpha), -(B^1\circ F + B^2\circ F_\Delta)/(f+f_\Delta)\circ(F+F_\Delta)^{-1}(2\alpha)\Big).\end{aligned}$$

Consequently, $\sqrt{n}(2(\hat{q}_\alpha(Y_i,\, i=1,\dots,2n) - \hat{q}_\alpha(Y_i,\, i=1,\dots,n)) - \Delta)$ is asymptotically normal with mean zero and variance (put $q^\Delta_\alpha := (F+F_\Delta)^{-1}(2\alpha)$)

$$\begin{aligned}\sigma_L^2 &= 4\,\mathbb{E}\Big[\Big(- \frac{B^1(F(q^\Delta_\alpha)) + B^2(F(q^\Delta_\alpha - \Delta))}{f(q^\Delta_\alpha) + f(q^\Delta_\alpha - \Delta)} + \frac{B^1(\alpha)}{f(q_\alpha)}\Big)^2\Big] \\ &= \frac{4\alpha(1-\alpha)}{f^2(q_\alpha)} + \frac{4F(q^\Delta_\alpha)(1-F(q^\Delta_\alpha)) + 4F(q^\Delta_\alpha - \Delta)(1-F(q^\Delta_\alpha - \Delta))}{(f(q^\Delta_\alpha) + f(q^\Delta_\alpha - \Delta))^2} \\ &\quad - \frac{8(\alpha \wedge F(q^\Delta_\alpha) - \alpha F(q^\Delta_\alpha))}{(f(q^\Delta_\alpha) + f(q^\Delta_\alpha - \Delta))f(q_\alpha)}\end{aligned}$$

While $\sigma_L^2 = \sigma_W^2$ for $\Delta = 0$ is straight-forward, the behaviour as $\Delta \to 0$ is analysed using $q^\Delta_\alpha = q_\alpha + \Delta/2 + O(\Delta^2)$, obtained from the Lipschitz continuity of f, and a fortiori

$$f(q^\Delta_\alpha) + f(q^\Delta_\alpha - \Delta) = 2f(q_\alpha) + o(\Delta),$$

obtained from the second difference bound on f. This gives

$$\sigma_L^2 = \sigma_W^2 - \frac{8((\Delta/2)f(q_\alpha)\mathbf{1}(\Delta<0) - \alpha(\Delta/2)f(q_\alpha))}{2f(q_\alpha)^2} + o(\Delta)$$

and hence the assertion. □

This result shows that under $H_0 : \Delta = 0$, the test statistics T_L and T_W are asymptotically identically distributed, whereas T_L has a larger asymptotic variance under the alternative than T_W. The relative deterioration $\frac{\sigma_L^2 - \sigma_W^2}{\sigma_W^2}$ is of order $f(q_\alpha)|\Delta|$ which can be interpreted as a signal-to-noise ratio. The deterioration is thus only negligible if the signal-to-noise ratio is small. Especially in

image denoising we face significant signal differences Δ at edges while we do not dispose of a very large number n of observed pixels such that this deterioration is clearly visible, see also Example 1 in Section 5. In addition to this improvement of results, the new tests are also easier to analyse and calibrate by using the independence of the noise on disjoint domains.

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.

7.2 Theorem. *For all $m \geqslant 0$ the sample $\varepsilon_1, \dots, \varepsilon_{2m+1}$ can be realised on the same probability space as a standard normal random variable Z such that* $\operatorname{med}(\varepsilon) := \operatorname{med}(\varepsilon_i,\, i = 1, \dots, 2m + 1)$ *satisfies*

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

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

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

$$\begin{aligned}
&\mathbb{E}[|\operatorname{med}(\varepsilon)|^{2r}\mathbf{1}(|Z| \leqslant \delta\sqrt{2m+1})] \\
&\quad\leqslant (2^{r-1} \vee 1)\,\mathbb{E}\left[\big(4(2m+1)f_\varepsilon(0)^2\big)^{-r}|Z|^{2r} + \tfrac{C^{2r}}{(2m+1)^{2r}}\Big(1 + Z^2\Big)^{2r}\right] \\
&\quad\leqslant C'(2m+1)^{-r}.
\end{aligned}$$

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 $\operatorname{med}(\varepsilon)$

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

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

$$\mathbb{E}[|\operatorname{med}(\varepsilon)|^{2r}\mathbf{1}(|Z| > \delta\sqrt{2m+1})] \leqslant \mathbb{E}[|\operatorname{med}(\varepsilon)|^{3r}]^{2/3}\,\mathbb{P}(|Z| > \delta\sqrt{2m+1})^{1/3},$$

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 $\operatorname{med}(\varepsilon)$, the lower bound follows symmetrically. The r-th moment is bounded by even simpler arguments.
The second assertion follows via quantile coupling from

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{4(2m+1)}f_\varepsilon(0)|\operatorname{med}(\varepsilon)| > \tau_m\right) \\
&\leqslant \mathbb{P}\left(|Z| + \tfrac{C}{\sqrt{2m+1}}(1 + Z^2) > \tau_m\right) + \mathbb{P}\left(|Z| > \delta\sqrt{2m+1}\right) \\
&\leqslant \mathbb{P}(2|Z| > \tau_m) + \mathbb{P}(|Z| > \delta\sqrt{2m+1}) \\
&\leqslant 2\exp(-\tau_m^2/8).
\end{aligned}$$

□

Acknowledgement

We are grateful for constructive remarks and questions by two referees and one associate editor that helped to improve the quality of the paper.

References

Arcones, M. A. (2002). Moderate Deviations for M-estimators. *Test* **11**(2), 465-500.

Arias-Castro, E. and D. Donoho (2009). Does median filtering truly preserve edges better than linear filtering?, *Annals of Statistics* **37**(3), 1172–1206.

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.

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.

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 neighbourhoods for denoising and clustering dynamic contrast enhanced-CT sequences, *Preprint Université Paris V*, available under `http://www.math-info.univ-paris5.fr/~rozen/RRBC-preprint/last-version.pdf`.

Spokoiny, V. and C. Vial (2009). Parameter tuning in pointwise adaptation using a propagation approach, *Annals of Statistics* **37**(5b), 2783–2807.

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.

Yu, K. and M.C. Jones (1998). Local linear quantile regression. *J. Amer. Stat. Assoc.* **93**(441), 228–237.

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.
002 "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.
003 "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.
004 "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.
005 "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.
006 "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.
007 "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.
008 "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.
009 "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.
010 "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.
011 "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.
012 "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.
013 "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.
014 "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.
015 "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.
016 "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.
017 "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.
018 "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.
019 "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.
020 "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.
021 "Customer Reactions in Out-of-Stock Situations – Do promotion-induced phantom positions alleviate the similarity substitution hypothesis?" by Jana Luisa Diels and Nicole Wiebach, May 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Extreme value models in a conditional duration intensity framework" by Rodrigo Herrera and Bernhard Schipp, May 2011.
023 "Forecasting Corporate Distress in the Asian and Pacific Region" by Russ Moro, Wolfgang Härdle, Saeideh Aliakbari and Linda Hoffmann, May 2011.
024 "Identifying the Effect of Temporal Work Flexibility on Parental Time with Children" by Juliane Scheffel, May 2011.
025 "How do Unusual Working Schedules Affect Social Life?" by Juliane Scheffel, May 2011.
026 "Compensation of Unusual Working Schedules" by Juliane Scheffel, May 2011.
027 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappus and Markus Reiß, May 2011.
028 "Asymptotic equivalence and sufficiency for volatility estimation under microstructure noise" by Markus Reiß, May 2011.
029 "Pointwise adaptive estimation for quantile regression" by Markus Reiß, Yves Rozenholc and Charles A. Cuenod, May 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".