Local Quantile Regression *

Vladimir Spokoiny [†] Weining Wang [‡] Wolfgang Karl Härdle [§]

July 8, 2012

Abstract

Quantile regression is a technique to estimate conditional quantile curves. It provides a comprehensive picture of a response contingent on explanatory variables. In a flexible modeling framework, a specific form of the conditional quantile curve is not a priori fixed. This motivates a local parametric rather than a global fixed model fitting approach. A nonparametric smoothing estimator of the conditional quantile curve requires to balance between local curvature and stochastic variability. In this paper, we suggest a local model selection technique that provides an adaptive estimator of the conditional quantile regression curve at each design point. Theoretical results claim that the proposed adaptive procedure performs as good as an oracle which would minimize the local estimation risk for the problem at hand. We illustrate the performance of the procedure by an extensive simulation study and consider a couple of applications: to tail dependence analysis for the Hong Kong stock market and to analysis of the distributions of the risk factors of temperature dynamics.

Keywords: local MLE, excess bound, propagation condition, adaptive bandwidth selection.

JEL classification: C00, C14, J01, J31

* The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Economic Risk", Humboldt-Universität zu Berlin is gratefully acknowledged. The first author is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF government grant, ag. 11.G34.31.0073.

† Professor, Weierstrass-Institute, Humboldt University Berlin, Moscow Institute of Physics and Technology, Mohrenstr. 39, 10117 Berlin, Germany. Email:spokoiny@wias-berlin.de

‡ Research associate at Ladislaus von Bortkiewicz Chair, the Institute for Statistics and Econometrics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email:wangwein@cms.hu-berlin.de

§ Professor at Humboldt-Universität zu Berlin and Director of C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email:haerdle@wiwi.hu-berlin.de

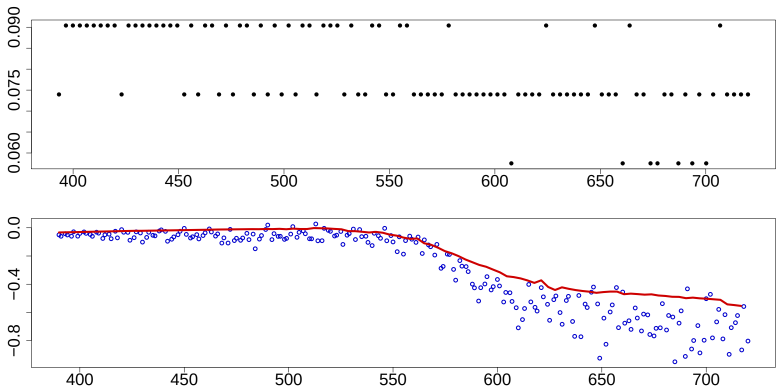


Figure 1: The bandwidth sequence (upper panel), plot of data and the estimated 90% quantile curve (lower panel)

1 Introduction

Quantile regression is gradually developing into a comprehensive approach for the statistical analysis of linear and nonlinear response models. Since the rigorous treatment of linear quantile regression by Koenker and Bassett (1978), richer models have been introduced into the literature, among them are nonparametric, semiparametric and additive approaches. Quantile regression or conditional quantile estimation is a crucial element of analysis in many quantitative problems. In financial risk management, the proper definition of quantile based Value at Risk impacts asset pricing, portfolio hedging and investment evaluation, Engle and Manganelli (2004), Cai and Wang (2008) and Fitzenberger and Wilke (2006). In labor market analysis of wage distributions, education effects and earning inequalities are analyzed via quantile regression. Other applications of conditional quantile studies include, for example, conditional data analysis of children growth and ecology, where it accounts for the unequal variations of response variables, see James et al. (2010).

In applications, the predominantly used linear form of the calibrated models is mainly determined by practical and numerical reasonings. There are many efficient algorithms (like sparse linear algebra and interior point methods) available, Portnoy and Koenker (1989), Portnoy and Koenker (1997), Koenker and Ferreira (1999), and Koenker (2005), etc. However, the assumption of a linear parametric structure can be too restrictive in many applications. This observation spawned a stream of literature on nonparametric modeling of quantile regression, Yu and Jones (1998), Fan et al. (1994), etc. One line of thought concentrated on different smoothing techniques, e.g. splines, kernel smoothing, etc.; see Fan and Gijbels (1996). Another line of literature considers structural semiparametric models to cope with curse of dimensionality, like, partial linear models, Härdle et al. (2012), etc., additive models, Kong et al. (2010), Horowitz and Lee (2005), etc; single index models, Wu et al. (2010), Koenker (2010), etc. Yet another strand of literature has been involved in ultra-high dimensional situations where a careful variable selection technique needs to be implemented, Belloni and Chernozhukov (2010) and Koenker (2010). In most of the aforementioned papers on non and semiparametric quantile regression, a smoothing parameter selection is implicit, and it is mostly a consequence of theoretical assumptions like e.g. rates of convergence, but falls short in practical hints for real data applications. An important exception is the method for local nonparametric kernel smoothing by Yu and Jones (1998) and Cai and Xu (2008). They both propose a data driven choice of tuning parameter.

To address the limitations of the above mentioned literature on local model selection for nonparametric quantile regression, we aim at proposing with theoretical justification an adaptive local quantile regression algorithm that is easy to implement and works for a wide class of applications. The idea of this algorithm is to select tuning parameters locally by a sequence of likelihood ratio tests. The novelty lies in a local model selection technique with computable risk bounds. The main message is that the proposed algorithm is feasible and beneficial for quantile smoothing and helps in proposing alternatives to other models. As an example, consider Figure 1 which presents our results for analyzing the Lidar data set, Ruppert et al. (2003). The presented quantile curve switches smoothness in the middle, and it is naturally reflected by the bandwidth sequence (upper panel) selected. In the presence of changing to sharper slope of the curve, the bandwidths get smaller to attain better approximations. This example shows that the algorithm proposed in this paper can adaptively choose the bandwidth at each design point.

This article is organized as follows: In Section 2, we introduce the local model selection (LMS) procedure and lay down how to simulate critical values. In Section 3, Monte Carlo simulations are conducted to illustrate the proposed methodology. In Section 4, we apply our method on checking the tail dependency among portfolio stocks, and on estimation of quantile curves for temperature risk factors. In Section 5, we explain the main theorem on "Oracle" properties to support the validity of our tests, with the relevant assumptions, definitions and conditions in Appendix. The technical details: 1, exponential risk bounds for conditional quantiles established using the representation of quantiles as Quasi Maximum Likelihood Estimation (qMLE)s of the asymmetric Laplace distribution, 2, theorems for the existence of critical values, 3, proof for "propagation", "stability" and "oracle" property are delegated to the Appendix.

2 Adaptive estimation procedure

This section introduces the considered problem and offers an adaptive estimation procedure.

2.1 Quantile regression model

Given the quantile level $\tau \in (0,1)$, the quantile regression model describes the following relation between the response Y and the regressor X:

$$\boldsymbol{P}(Y > f(x) \,|\, X = x) = \tau,$$

where $f(x)$ is the unknown *quantile regression function.* This function is the target of the analysis and it has to be estimated from independent observations $\{X_i, Y_i\}_{i=1}^{n}$. For the case of a deterministic design, this quantile relation can be represented as

$$Y_i = f(X_i) + \varepsilon_i\,, \tag{1}$$

where the errors ε_i follow $\boldsymbol{P}(\varepsilon_i > 0) = \tau$.

For simplicity of presentation, we consider a univariate regressor $X \in I\!\!R^1$ and a deterministic design in this paper, an extension to the d-dimensional case $X \in I\!\!R^d$ with $d > 1$ is straightforward.

2.2 A qMLE View on Quantile Estimation

The quantile function $f(\cdot)$ in (1) is usually recovered by minimizing the sum

$$\sum_{i=1}^{n} \rho_\tau\{Y_i - f(X_i)\}, \tag{2}$$

over the class of all considered quantile functions $f(\cdot)$, where

$$\rho_\tau(u) \overset{\text{def}}{=} u\{\tau\, 1\!\!\mathrm{I}(u \geq 0) - (1-\tau)\, 1\!\!\mathrm{I}(u < 0)\} = u\bigl\{\tau - 1\!\!\mathrm{I}(u < 0)\bigr\}.$$

Such an approach is reasonable because the true quantile function $f(x)$ minimizes the expected value of the sum in (2). An important special case is given by $\tau = 1/2$. Then an estimator of $f(\cdot)$ is built as minimizer of the least absolute deviations (LAD) contrast $\sum |Y_i - f(X_i)|$.

The minimum contrast approach based on minimization of (2) can also be put in a quasi maximum likelihood framework. Assume that the residuals ε_i from (1) are i.i.d. and $p(x)$ is their negative log-density on $I\!\!R^1$. Then the joint log-density is given by the sum

$$-\sum p\big(Y_i - f(X_i)\big)$$

and its maximization is equivalent to minimization of the contrast (2) with a pdf from the *asymmetric Laplace distribution* ALD_τ:

$$p(u) = p_\tau(u) = \log\big\{\tau(1-\tau)\big\} - \rho_\tau(u), \quad -\infty < u < \infty. \tag{3}$$

The *parametric approach* (PA) additionally assumes that the quantile regression function $f(\cdot)$ belongs to a parametric family of functions $\big\{f_{\boldsymbol{\theta}}(x), \boldsymbol{\theta} \in \Theta\big\}$, where Θ is a subset of the p-dimensional Euclidean space. Equivalently,

$$f(x) = f_{\boldsymbol{\theta}^*}(x),$$

where $\boldsymbol{\theta}^*$ is the true parameter which is usually the target of estimation.

Examples are a constant model:

$$f_{\theta^*}(x) \equiv \theta_0,$$

with $\theta^* = \theta_0$ or a linear model:

$$f_{\boldsymbol{\theta}^*}(x) = \theta_0 + \theta_1 x,$$

with $\boldsymbol{\theta}^* = (\theta_0, \theta_1)^\top$.

Denote by $\boldsymbol{P}_{\boldsymbol{\theta}}$ the parametric measure on the observation space which corresponds to the regression model (1) with $f(\cdot) \equiv f_{\boldsymbol{\theta}}(\cdot)$ and with the i.i.d. errors ε_i following the asymmetric Laplace distribution (3). Then the log-likelihood $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta})$ for $\boldsymbol{P}_{\boldsymbol{\theta}}$ can be written as

$$L(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\big\{\tau(1-\tau)\big\} \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} \rho_\tau\{Y_i - f_{\boldsymbol{\theta}}(X_i)\} \tag{4}$$

and the qMLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$, or, equivalently minimizes the contrast $\sum_{i=1}^{n} \rho_\tau\{Y_i - f_{\boldsymbol{\theta}}(X_i)\}$ over all $\boldsymbol{\theta} \in \Theta$.

The described parametric construction is based on two assumptions: one is about the error distribution (3) and the other one is about the shape of the regression function f. However, it is only used for motivating our approach. Our theoretical study will be done under the true data distribution which follows (1) under mild regularity conditions. The next section explains how a smooth regression function f can be modeled by a flexible local parametric assumption.

2.3 Local polynomial qMLE

This section explains how the restrictive *global PA* $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$ can be relaxed by using a local parametric approach. Let a point x be fixed. The *local PA* at a point $x \in I\!R$ only requires that the quantile regression function $f(\cdot)$ can be approximated by a parametric function $f_{\boldsymbol{\theta}}(\cdot)$ from the given family in a vicinity of x. Below we fix a family of polynomial functions of degree p leading to the usual Taylor approximation:

$$f(u) \approx f_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \theta_0 + \theta_1(u-x) + \ldots + \theta_p(u-x)^p/p! \tag{5}$$

for $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)^\top$. The corresponding parametric model can be written as

$$Y_i = \Psi_i^\top \boldsymbol{\theta} + \varepsilon_i \, , \tag{6}$$

where $\Psi_i = \{1, (X_i - x), (X_i - x)^2/2!, \ldots, (X_i - x)^p/p!\}^\top \in I\!R^{p+1}$.

A *local likelihood approach* at x is specified by a *localizing scheme* W given by a collection of weights w_i for $i = 1, \ldots, n$. The weights w_i vanish for points X_i lying outside a vicinity of the point x. A standard proposal for choosing the weights W is $w_i = K_{\text{loc}}\{(X_i - x)/h\}$, where $K_{\text{loc}}(\cdot)$ is a *kernel function* with a compact support, while h is a *bandwidth* controlling the degree of localization.

Define now the local log-likelihood at x by

$$L(W, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \tau(1-\tau) \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i \, . \tag{7}$$

This expression is similar to the global log-likelihood in (4), but each summand in $L(W, \boldsymbol{\theta})$ is multiplied with the weight w_i, so only the points from the local vicinity of x contribute to $L(W, \boldsymbol{\theta})$. Note that this local log-likelihood depends on the central point x via the structure of the basis vectors Ψ_i and via the weights w_i. The corresponding local qMLE at x is defined via maximization of $L(W, \boldsymbol{\theta})$:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}(x) &= \{\widetilde{\theta}_0(x), \widetilde{\theta}_1(x), \ldots, \widetilde{\theta}_p(x)\}^\top \\ &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(W, \boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i \, . \end{aligned} \tag{8}$$

The first component $\widetilde{\theta}_0(x)$ provides an estimator of $f(x)$, while $\widetilde{\theta}_m(x)$ is an estimator of the derivative $f^{(m)}(x)$, $m = 1, \ldots, p$.

2.4 Selection of a Pointwise Bandwidth

The choice of bandwidth h is an important issue in implementing (8). One can reduce the variance of the estimation by increasing the bandwidth, but at a price of possibly inducing more modeling bias measured by the accuracy of approximation in (5); see Figure 2.

A desirable choice of a bandwidth at a fixed point would strike a balance between the variance and the bias depending on the local shape of $f(\cdot)$ in the vicinity of x. Many approaches have been proposed along this line; see e.g. Yu and Jones (1998) and references therein. However, their justification and implementation is based on some asymptotic arguments and require large samples. Here we propose a pointwise bandwidth selection technique based on finite sample theory.

Our basic setup of the algorithm is described as follows. First one fixes a finite ordered set of possible bandwidths $h_1 < h_2 < \ldots < h_K$, where h_1 is very small, while h_K should be a global bandwidth of the order of the design range. The bandwidth sequence can be taken geometrically increasing of the form $h_k = ab^k$ with fixed $a > 0$, $b > 1$, and $n^{-1} < ab^k < 1$ for $k = 1, \ldots, K$ (A.2.). The total number K of the candidate bandwidths is then at most logarithmic in the sample size n. For each $k \le K$, an ordered weighting schemes $W^{(k)} = (w_1^{(k)}, w_2^{(k)}, \ldots, w_n^{(k)})^\top$ is defined via $w_i^{(k)} \stackrel{\text{def}}{=} K_{\text{loc}}\{(x - X_i)/h_k\}$ leading to a local quantile estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ with

$$\widetilde{\boldsymbol{\theta}}_k(x) = \operatorname*{argmax}_{\boldsymbol{\theta}} L(W^{(k)}, \boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i^{(k)} . \tag{9}$$

The proposed selection procedure is similar in spirit to Lepski et al. (1997). If the underlying quantile regression function is smooth, one can expect a good quality of approximation (5) for a large bandwidth values among $\{h_k\}_{k=1}^K$. Moreover, if the approximation is good for one bandwidth, it will be also suitable for all smaller bandwidths. So, if we observe a significant difference between the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ corresponding to the bandwidth h_k and an estimator $\widetilde{\boldsymbol{\theta}}_\ell(x)$ corresponding to a smaller bandwidth h_ℓ, this is an indication that the approximation (5) for the window size h_k becomes too rough. This justifies the following algorithm. Start with the smallest bandwidth h_1. For any $k > 1$, compute the local qMLE $\widetilde{\boldsymbol{\theta}}_k(x)$ and check whether it is consistent with all the previous estimators $\widetilde{\boldsymbol{\theta}}_\ell(x)$ for $\ell < k$. If the consistency check is negative, the procedure terminates and selects the latest accepted estimator.

The most important ingredient of the method is the consistency check. The Lepski method suggests to use the difference $\widetilde{\boldsymbol{\theta}}_k(x) - \widetilde{\boldsymbol{\theta}}_\ell(x)$ as a test statistic; see e.g. Lepski et al. (1997). We follow the suggestion from Polzehl and Spokoiny (2006) and apply the localized likelihood ratio type test. More precisely, the local MLE $\widetilde{\boldsymbol{\theta}}_\ell(x)$ maximizes the log-likelihood $L\bigl(W^{(\ell)}, \boldsymbol{\theta}\bigr)$, and the maximal value of (7) given by $\sup_{\boldsymbol{\theta}} L\bigl(W^{(\ell)}, \boldsymbol{\theta}\bigr) =$

$L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{\ell}(x)\big)$ is compared with the particular log-likelihood value $L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{k}(x)\big)$, where the estimator $\widetilde{\boldsymbol{\theta}}_{k}(x)$ is obtained by maximizing the other local log-likelihood function $L(W^{(k)}, \boldsymbol{\theta})$. The difference $L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{\ell}(x)\big) - L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{k}(x)\big)$ is always non-negative. The check rejects $\widetilde{\boldsymbol{\theta}}_{k}(x)$ if this difference is too large for some $\ell < k$. Equivalently one can say that the test checks whether $\widetilde{\boldsymbol{\theta}}_{k}(x)$ belongs to the confidence sets $\mathcal{E}_{\ell}(\mathfrak{z})$ of $\widetilde{\boldsymbol{\theta}}_{\ell}(x)$:

$$\mathcal{E}_{\ell}(\mathfrak{z}) \stackrel{\text{def}}{=} \big\{\boldsymbol{\theta} : L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{\ell}(x)\big) - L\big(W^{(\ell)}, \boldsymbol{\theta}\big) \le \mathfrak{z}\big\}.$$

A great advantage of the likelihood ratio test is that the critical value $\mathfrak{z}$ can be selected universally. This is justified by the Wilks phenomenon: the likelihood ratio test statistics is nearly χ^2 and its asymptotic distribution depends only on the dimension of the parameter space. Unfortunately, these arguments do not apply for finite samples under possible model misspecification and we therefore offer an alternative way of fixing the critical values $\mathfrak{z}$ which is based on the so called propagation condition. We also allow that the width of the confidence set $\mathcal{E}_{\ell}(\mathfrak{z})$ depends on the index ℓ, that is, $\mathfrak{z} = \mathfrak{z}_{\ell}$. Our adaptation algorithm can be summarized as follows: at each step k, an estimator $\widehat{\boldsymbol{\theta}}_{k}(x)$ is constructed based on the first k estimators $\widetilde{\boldsymbol{\theta}}_{1}(x), \ldots, \widetilde{\boldsymbol{\theta}}_{k}(x)$ by the following rule,

- Start with $\widehat{\boldsymbol{\theta}}_{1}(x) = \widetilde{\boldsymbol{\theta}}_{1}(x)$.
- For $k \ge 2$, $\widetilde{\boldsymbol{\theta}}_{k}(x)$ is accepted and $\widehat{\boldsymbol{\theta}}_{k}(x) \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_{k}(x)$, if $\widetilde{\boldsymbol{\theta}}_{k-1}(x)$ was accepted and

$$L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{\ell}(x)\big) - L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_{k}(x)\big) \le \mathfrak{z}_{\ell}, \qquad \ell = 1, \ldots, k-1. \tag{10}$$

- Otherwise $\widehat{\boldsymbol{\theta}}_{k}(x) = \widehat{\boldsymbol{\theta}}_{k-1}(x)$.

The adaptive estimator $\widehat{\boldsymbol{\theta}}(x)$ is the latest accepted estimator after all K steps:

$$\widehat{\boldsymbol{\theta}}(x) \stackrel{\text{def}}{=} \widehat{\boldsymbol{\theta}}_{K}(x)$$

A visualization of the procedure is presented in Figure 2. The critical values $\mathfrak{z}_{\ell}$'s are selected by an algorithm based on the propagation condition explained in the next section.

2.5 Parameter Tuning by Propagation Condition

The practical implementation requires to fix the critical values of $\mathfrak{z}_{1}, \ldots, \mathfrak{z}_{K-1}$. We apply the *propagation* approach which is an extension of the proposal from Spokoiny (2009); Spokoiny and Vial (2009). The idea is to tune the parameter of the procedure for one artificial parametric situation. Later we show that such defined critical values work well in the general setup and provide a nearly efficient estimation quality. The presented bandwidth selector can be viewed as a multiple testing procedure. This suggests to fix the critical

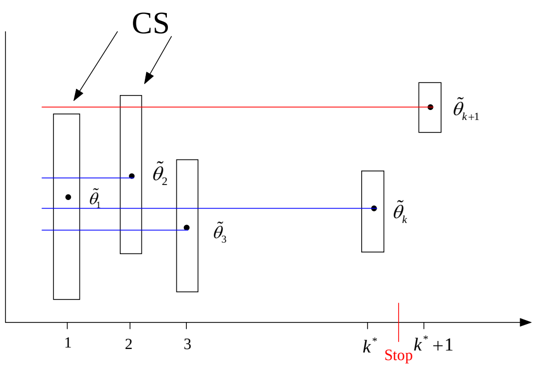


Figure 2: Demonstration of the local adaptive algorithm.

values as in the general testing theory by ensuring a prescribed performance under the null hypothesis. In our case, the null hypothesis corresponds to the pure parametric situation with $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$ in the equation (1). Moreover, we fix some particular distribution of the errors ε_i, our specific choice is ALD_τ with parameter τ. Below in this section we denote by $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ the data distribution under these assumptions.

For this artificial data generating process, all the estimators $\widetilde{\boldsymbol{\theta}}_k(x)$ should be consistent to each other and the procedure should not terminate at any intermediate step $k < K$. We call this effect as *propagation*: in the parametric situation, the degree of locality will be successfully increased until it reaches the largest scale. The critical values are selected to ensure the desired *propagation condition* which effectively means a "no false alarm" property: the selected adaptive estimator coincides in the most of cases with the estimator $\widetilde{\boldsymbol{\theta}}_K(x)$ corresponding to the largest bandwidth. The event $\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\}$ for $k \le K$ is associated with a false alarm and the corresponding loss can be measured by the difference

$$L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \stackrel{\text{def}}{=} L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr) - L\bigl(W^{(k)}, \widehat{\boldsymbol{\theta}}_k(x)\bigr).$$

The *propagation condition* postulates that the risk induced by such false alarms is smaller than the upper bound for the risk of the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ in the pure parametric situation:

$$\boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \le \alpha \mathfrak{R}_r \qquad k = 2, \ldots, K, \tag{11}$$

where the constant $\mathfrak{R}_r$ is such that for all $k \le K$, it holds

$$\boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr) \le \mathfrak{R}_r \,.$$

The values α and r in (11) are two hyper-parameters. The role of α is similar to the significance level of a test, while r denotes the power of the loss function. It is worth mentioning that

$$\boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big) \to \boldsymbol{P}_{\boldsymbol{\theta}^*}\big\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\big\}, \qquad r \to 0.$$

The critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ enter implicitly in the propagation condition: if the false alarm event $\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\}$ happens too often, it is an indication that some of the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ are too small. Note that (11) relies on the artificial parametric model $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ instead of the true model $\boldsymbol{P}$. The point $\boldsymbol{\theta}^*$ here can be selected arbitrarily, e.g. $\boldsymbol{\theta}^* = 0$. This fact relies on linear parametric structure of the model (6) and is justified by the following simple lemma.

Lemma 1. *The distribution of* $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big)$ *and of* $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\big)$ *under* $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ *does not depend on* $\boldsymbol{\theta}^*$*.*

Proof. Under PA $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$, it holds $Y_i - f(X_i) = Y_i - \Psi_i^\top \boldsymbol{\theta}^* = \varepsilon_i$ and hence,

$$L(W^{(k)}, \boldsymbol{\theta}) = \log\{\tau(1-\tau)\} \sum_{i=1}^n w_i^{(k)} + \sum_{i=1}^n \rho_\tau\big(\varepsilon_i - \Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\big) w_i^{(k)}.$$

A simple inspection of this formula yields that the distribution of $L(W^{(k)}, \boldsymbol{\theta})$ only depends on $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. In other words, we can use the free parameter $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ whatever $\boldsymbol{\theta}^*$ is, e.g. $\boldsymbol{\theta}^* \equiv 0$. The same argument applies to the difference $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\big)$ for $\ell < k$. Moreover, $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big)$ is a function of $\big\{L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\big)\big\}_{\ell=1}^k$, so the distribution of $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big)$ does not depend on $\boldsymbol{\theta}^*$. □

A choice of critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ can be implemented in the following way:

- Consider first only $\mathfrak{z}_1$ and fix $\mathfrak{z}_2 = \ldots = \mathfrak{z}_{K-1} = \infty$, leading to the estimators $\widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x)$ for $k = 2, \ldots, K$. The value $\mathfrak{z}_1$ is selected as the minimal one for which

$$\frac{1}{\mathfrak{R}_r} \boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x)\big) \le \frac{\alpha}{K-1}, \qquad k = 2, \ldots, K. \tag{12}$$

- With selected $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$, set $\mathfrak{z}_{k+1} = \ldots = \mathfrak{z}_{K-1} = \infty$. Any particular value of $\mathfrak{z}_k$ would lead to the set of parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_k, \infty, \ldots, \infty$ and the family of estimators $\widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k, x)$ for $m = k+1, \ldots, K$. Select the smallest $\mathfrak{z}_k$ ensuring

$$\frac{1}{\mathfrak{R}_r} \boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\big(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x)\big) \le \frac{k\alpha}{K-1} \tag{13}$$

  for all $m = k+1, \ldots, K$.

Few remarks to the proposed algorithm.

1. A value $\mathfrak{z}_1$ ensuring (12) always exists because the choice $\mathfrak{z}_1 = \infty$ yields $\widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x) = \widetilde{\boldsymbol{\theta}}_k(x)$ for all $k \ge 2$.

2. The value $L^r\big(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x)\big)$ from (13) only accumulates the losses associated with the false alarms at the first k steps of the procedure. The other checks at further steps are always accepted because the corresponding critical values $\mathfrak{z}_{k+1}, \ldots \mathfrak{z}_{K-1}$ are set to infinity.

3. The accumulated risk bound $\frac{k\alpha}{K-1}$ grows at each step by $\alpha/(K-1)$. This value can be seen as maximal risk associated with the CV's $\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k$.

4. The value $\mathfrak{z}_k$ ensuring (13) always exists, because the choice $\mathfrak{z}_k = \infty$ yields

$$\widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x) = \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_{k-1}, x)$$

   for all $m \geq k$.

5. All the computed values depend on the considered linear parametric model, the sequence of bandwidths h_k and the quantile level τ. They also depend on the local point x via the basis vectors Ψ_i. However, under common regularity conditions on the design $X_1, \ldots, X_n$, the dependency on x is rather minor. Therefore, the adaptive estimation procedure can be repeated at different points without reiterating the steps of selecting the critical values.

3 Simulations

First, we check the critical values at different quantile levels $\tau = 0.05, 0.5, 0.75, 0.95$ and for different noise distributions: a) ALD, b) normal and c) student $t(3)$. Also, we study how does misidentification of noise distribution affect the critical values.

Second, we compare the performance of our local bandwidth algorithm with two other bandwidth selection techniques. One proposal is from Yu and Jones (1998), in which they consider a rule of thumb bandwidth based on the assumption that the quantiles are parallel, and another comes from Cai and Xu (2008), where an approach based on a nonparametric version of the Akaike information criterion (AIC) is implemented.

3.1 Critical Values

Here we summarize our numerical results on choosing the critical values by the propagation condition as described in Section 2.5. We only consider local constant modeling with $p = 0$ and local linear modeling with $p = 1$ starting with $p = 0$.

Table 1 shows the critical values with several choices of α and r with $\tau = 0.75$, $m = 10000$ Monte Carlo samples, and an bandwidth sequence $(8, 14, 19, 25, 30, 36, 41, 52)*0.001$

Table 1: Critical Values with different r and α

$\alpha = 0.25$,	$r = 0.5$	6.123	2.333	0.987	3.678e-05	0.000
$\alpha = 0.5$,	$r = 0.5$	4.616	1.578	0.357	2.472e-05	0.000
$\alpha = 0.6$,	$r = 0.5$	3.203	0.679	0.025	0.006	7.278e-05
$\alpha = 0.25$,	$r = 0.75$	9.127	3.288	1.031	0.126	5.675e-05
$\alpha = 0.25$,	$r = 1$	12.75	4.280	1.224	1.095e-04	0.000

Table 2: Critical Values with Different τ

$\tau = 0.05$	6.464	2.204	0.620	3.345e-05	0.000
$\tau = 0.5$	7.997	3.089	0.986	0.300e-05	0.000
$\tau = 0.75$	9.203	3.910	1.106	0.123	7.254e-05
$\tau = 0.95$	8.589	5.452	1.904	0.334	1.203e-05

scaled to the interval $[0, 1]$. Critical values decrease when α increases, and increase when r increases.

Table 2 displays critical values for different τ, with $\alpha = 0.25$, $r = 0.5$, $m = 10000$ Monte Carlo samples, a bandwidth sequence $\mathfrak{H}_1 = (8, 14, 19, 25, 30, 36, 41, 52) * 0.001$, and $\mathcal{N}(0, 1)$ noise. Critical values behave similarly for different τ.

Table 3 displays the critical values for three alternative bandwidth sequences:

$$\mathfrak{H}_1 = (8, 14, 19, 25, 30, 36, 41, 52) * 0.001$$

$$\mathfrak{H}_2 = (8, 16, 25, 36, 49, 63, 79, 99) * 0.001$$

$$\mathfrak{H}_3 = (5, 8, 14, 19, 27, 36, 46, 58) * 0.001$$

with $\alpha = 0.25$, $r = 0.5$, and $\tau = 0.85$. Although the critical values differ for different

Table 3: Critical Values with Different Bandwidth Sequences

$\mathfrak{H}_1$	11.33	1.243	6.933e-05	0.000	0.000
$\mathfrak{H}_2$	18.39	6.479	2.230	0.469	8.738e-05
$\mathfrak{H}_3$	6.123	2.333	0.987	3.678e-05	0.000

bandwidth sequences, α, r and τ, they indicate the same patterns (finite and decreasing).

We simulate from different data generating processes, namely the distribution of ε_i (given by the density $p(\cdot)$) does not necessarily coincide with the likelihood (ALD_τ) taken to simulate critical values. Table 4 presents critical values simulated under $t(3)$,

$\mathcal{N}(0,1)$ and ALD_τ. The critical values show the same trend with some differences, so we conclude that a misidentification of error distribution would not significantly contaminate the confidence sets.

Table 4: Critical Values with Different Noise Distributions

$\mathcal{N}(0,1)$	11.50	4.924	2.514	1.313	2.765e-05
ALD_τ	14.05	6.554	3.304	1.443	5.879e-05
$t(3)$	15.42	8.707	2.370	0.342	3.898e-05

In Table 5, critical values are shown in the same circumstances as in Table 4 for the local linear case. Since introducing one more variable (trend), critical values doubled or tripled compared to the local constant case. The behavior with respect to tail functions stays the same.

Table 5: Critical Values with Different Noise Distributions in Local Linear Case

$\mathcal{N}(0,1)$	29.97	58.64	43.21	33.41	19.43	07.40
$ALD(0.5)$	45.28	74.51	66.43	50.42	31.42	13.50
$t(3)$	51.77	84.94	59.28	44.99	29.07	11.57

3.2 Comparison of Different Bandwidth Selection Techniques

We illustrate our proposal by considering $x \in [0,1]$, $\tau = 0.75$. The sample with ($n = 1000$) are simulated under three scenarios:

$$f^{[1]}(x) = \begin{cases} 0 & \text{if } x \in [0, 0.333]\,; \\ 8 & \text{if } x \in (0.333, 0666]\,; \\ -1 & \text{if } x \in (0.666, 1] \end{cases}$$

$$f^{[2]}(x) = 2x(1+x),$$

$$f^{[3]}(x) = \sin(k_1 x) + \cos(k_2 x)\,\mathbb{I}\{x \in (0.333, 0.666)\} + \sin(k_2 x)$$

The noise distributions are: $\mathcal{N}(0, 0.03), ALD_\tau, t(3)$.

Figure 3 presents pictures on comparisons of different estimators in the local constant case. Figure 4 and 5 show in the local linear case the estimators of the functions ($\widehat{f}(x)$) and its first derivatives as well. Our technique provides closer fits to the true curve ($f(x)$) than methods with a global fixed bandwidth, especially in the presence of jump. Table 6, which shows the averaged absolute errors for the four methods, further confirms our conclusion.

Table 6: Comparison of Monte Carlo errors, averaged over 1000 samples

	Fixed bandw	Local constant	Local linear	Fixed bandw (Cai)
$f^{[1]}(x)$	0.654	0.172	0.169	0.378
$f^{[2]}(x)$	0.206	0.008	0.008	0.245
$f^{[3]}(x)$	0.137	0.021	0.019	0.123

Table 7 offers further an analysis for misspecified error distributions. Specifically, to evaluate the accuracy of our estimation for error distributions generated differently than the ALD density. Table 7 gives L_1 errors between $\widehat{f}(\cdot)$ (with critical values simulated from ALD_τ) and $f(\cdot)$, from which we conclude that mis-specification of error distributions would not contaminate our results significantly.

Table 7: Comparison of error mis-specification, errors are calculated averaged over 1000 samples

	Local constant $\{\mathcal{N}(0,1)\}$	Local constant $\{t(3)\}$	Local linear $\{\mathcal{N}(0,1)\}$
$f^{[1]}(x)$	0.252	0.220	0.169
$f^{[2]}(x)$	0.070	0.016	0.043
$f^{[3]}(x)$	0.009	0.021	0.019

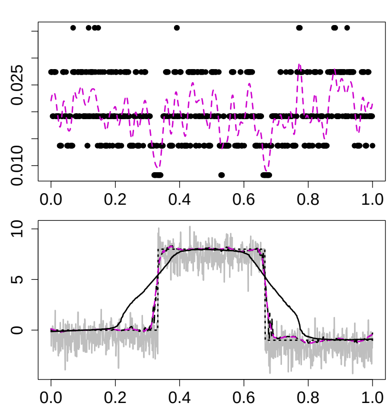
4 Applications

In the study of financial products, it is very important to detect and understand tail dependence among underlyings such as stocks. In particular, the tail dependence structure represents the degree of dependence in the corner of the lower-left quadrant or upper-right quadrant of a bivariate distribution. Hauksson et al. (2001) and Embrechts and Straumann (1999) provide a good access to the literature on tail dependence and Value at Risk. With the adaptive quantile technique, we provide an alternative approach to study tail dependence.

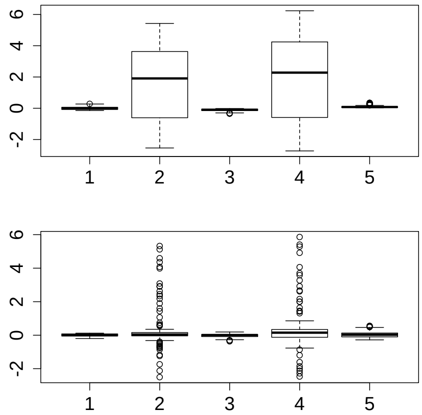
The correlation is calibrated from real data as given in Figure 6, where X is standardized return from stock “clpholdings” from Hong Kong Hangseng Index, and Y is return from stock “cheung kong”. The conditional quantile function is linear, for example, $X_1 \in \mathcal{N}(u_1, \sigma_1)$ and $X_2 \in \mathcal{N}(u_2, \sigma_2)$, the conditional quantile function α is:

$$f(x) = \varphi^{-1}(\alpha)(\sigma_2 - \sigma_{12}^2/\sigma_1) + u_i + \sigma_{12}\sigma_2^{-1}(x - u_2).$$

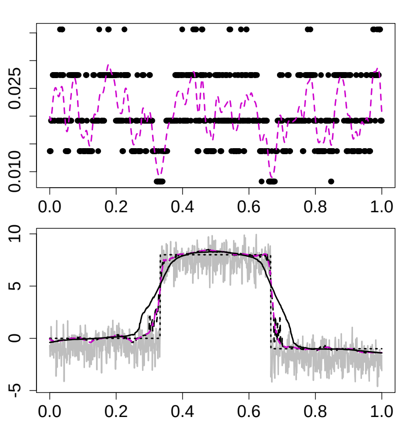
Figure 6 and Figure 7 show the empirical conditional quantile curves actually deviate



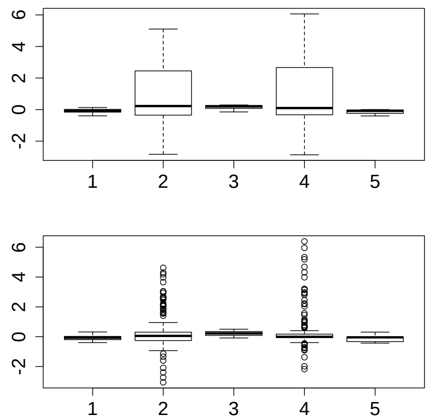
(a) Normal



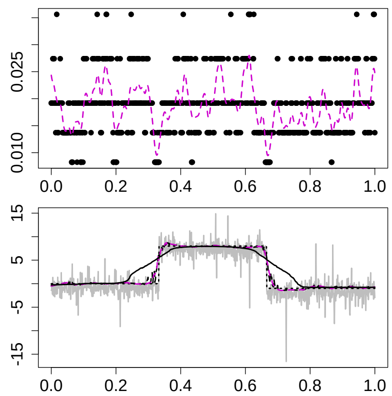
(b) Normal



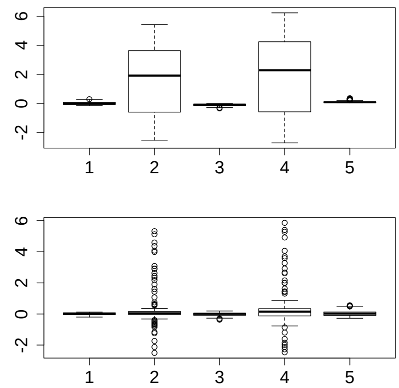
(c) ALD(0.5)



(d) ALD(0.5)

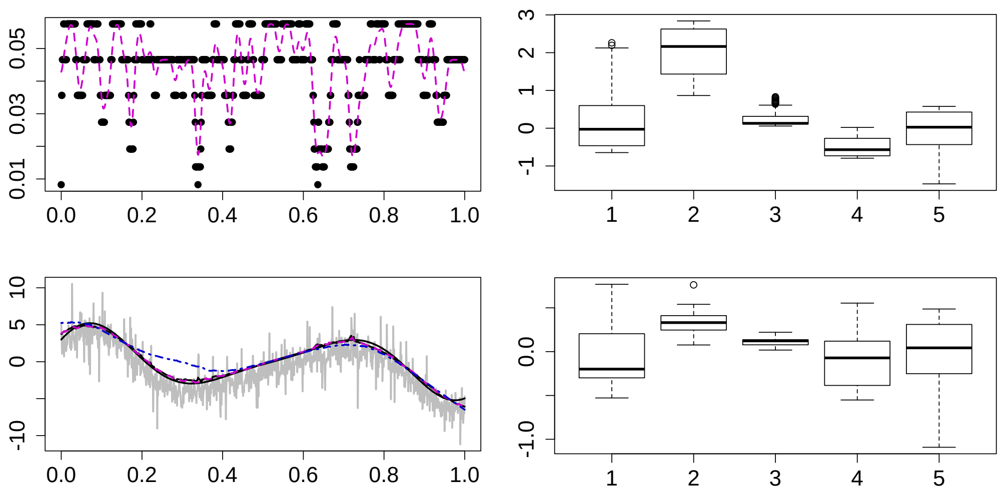


(e) t(3)



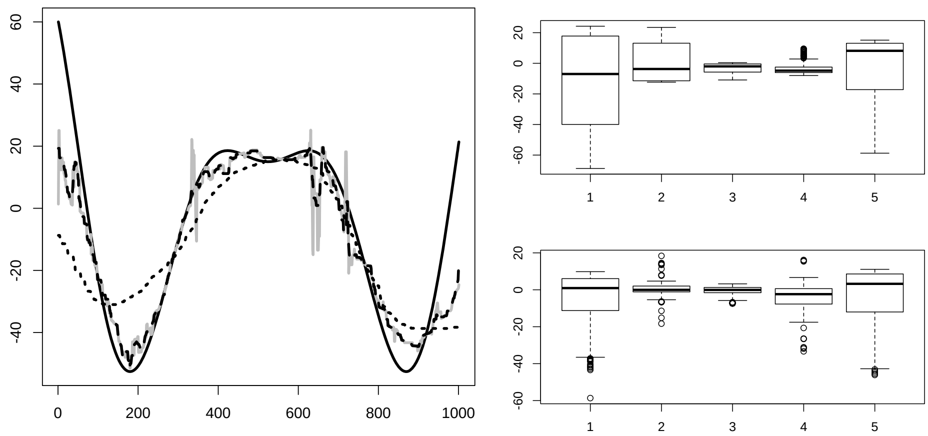
(f) t(3)

Figure 3: The bandwidth sequence (upper left panel), the smoothed bandwidth (magenta dashed); the data with noise (grey, lower left panel), the adaptive estimation of 0.75 quantile (dashed black), the quantile smoother with fixed optimal bandwidth $= 0.06$ (solid black), the estimation with smoothed bandwidth (dashed magenta); boxplot of block residuals fixed bandwidth (upper right), adaptive bandwidth (lower right)



(a) (b)

Figure 4: The bandwidth sequence (upper left panel), the smoothed bandwidth sequence (dashed magenta); the observations (grey, lower left panel), the adaptive estimation of 0.75 quantile (dotted black), the true curve (solid black), the quantile smoother with fixed optimal bandwidth $= 0.063$ (dashed dotted blue), the estimation with adaptively smoothed bandwidth (dashed magenta); the blocked error of the adaptive estimator (lower right); the fixed estimator (upper right).



(a) (b)

Figure 5: The adaptive estimation of first derivative of the above quantile function (left panel grey), the true curve (solid black), the estimation with smoothed bandwidth (dashed black), the quantile smoother with fixed optimal bandwidth $= 0.045$ (dotted black); the blocked error of the adaptive estimator (lower right); the fixed estimator (upper right).

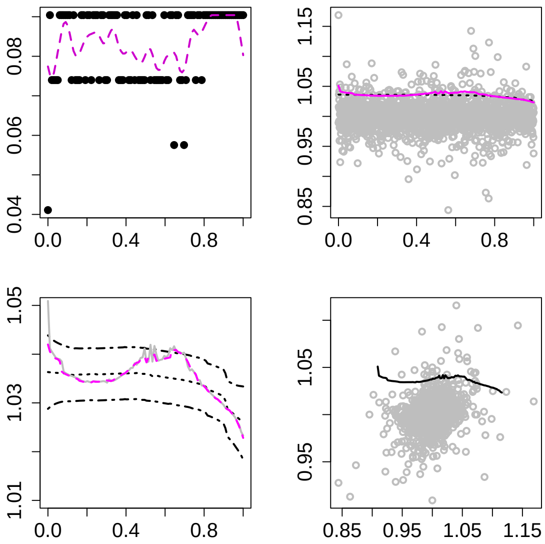


Figure 6: The bandwidth sequence with smoothed bandwidth curve(upper left panel), the smoothed bandwidth (dashed magenta); Scatter plot of stock returns (upper right panel), the adaptive estimation of 0.90 quantile (solid magenta), the quantile smoother with fixed optimal bandwidth $= 0.15$ (dotted black); fixed bandwidth curve (dotted black), adaptive bandwidth curve (grey), the estimation with smoothed bandwidth (dashed magenta), confidence band (dashed black) (lower left panel); adaptive bandwidth with normal scale (lower right).

from the one calculated from normal distributions, which implies non normality. The motivation of adaptive bandwidth selection is clear to see from Figure 6 and Figure 7, the dependency structure change is more obvious compared with the fixed bandwidth curve. Moreover, the flexible adaptive curve is not likely to be a consequence of overfitting since it mostly lies in the confidence bands produced by fixed bandwidth estimation, see Härdle and Song (2010).

Figure 8 shows the first derivative curve for the above example. The curve gets more volatile while x increases until a drastic change, then it turns flat.

We measure the deviation from normality by accumulated L_1 distance to the normal fitting and examine different combination of stocks from Hong Kong Hangseng Index. The results is summarized in Table 8.

Another application of quantile function estimation is in temperature data analysis, which is of key interest for pricing temperature derivatives. Quantile regression can provide a more flexible and comprehensive approach to understand the temperature risk drivers defined in (14).

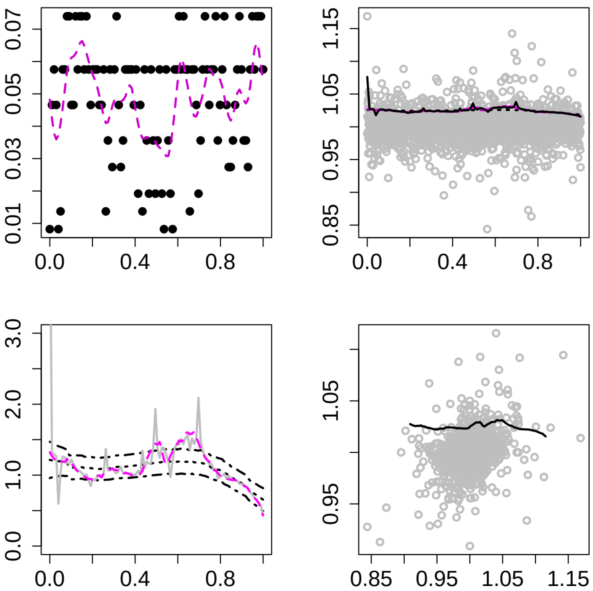


Figure 7: The bandwidth sequence with smoothed bandwidth curve (upper left panel); Scatter plot of stock returns (upper right panel), the adaptive estimation of 0.90 quantile (red), the quantile smoother with fixed optimal bandwidth $= 0.19$ (dotted black); fixed bandwidth curve (dotted black), adaptive bandwidth curve (grey), confidence bands (dotted dashed black) (lower left panel); adaptive bandwidth with normal scale (lower right panel)

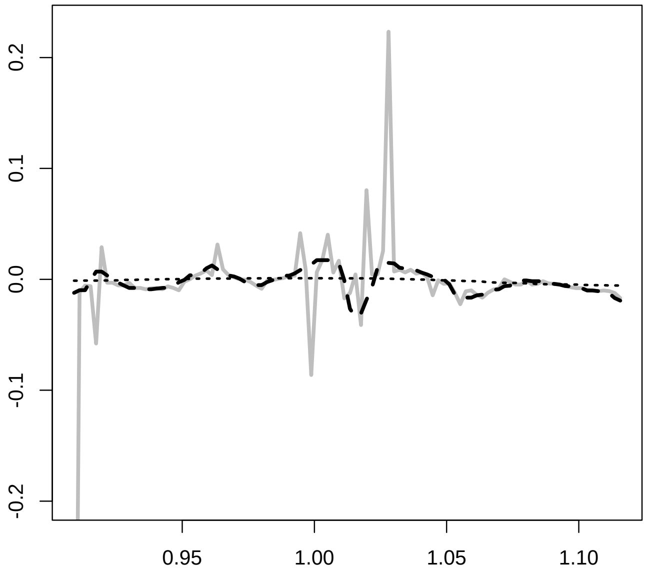


Figure 8: The adaptive trend curve (grey), smoothed adaptive curve (dashed black), estimation with fixed bandwidth (dotted black). $\tau = 0.90$

Table 8: Summary of deviation from normality

	Chalco	Cosco pacific	Bank of China
New world devo	0.252	0.220	0.169
Sino land	0.070	0.016	0.043
Swire pacific A	0.009	0.021	0.019

Denote daily temperature as $T \mapsto (t, j)$, with $t = 1, \cdots, \tau = 365$ days, $j = 0, \cdots, J$ years. The time series decomposition for $T_{t,j}$ is given as:

$$
\begin{aligned}
X_{t,j} &= T_{t,j} - \Lambda_t \\
X_{t,j} &= \sum_{l=1}^{L} \beta_l X_{t-l,j} + \sigma_t \eta_{t,j} \\
\eta_{t,j} &\sim \mathcal{N}(0,1), \\
\varepsilon_{t,j} &\stackrel{\text{def}}{=} \sigma_t \varepsilon_{t,j} \\
\widehat{\varepsilon}_{t,j} &\stackrel{\text{def}}{=} X_{365j+t} - \sum_{l=1}^{L} \widehat{\beta}_l X_{365j+t-l} \qquad (14)
\end{aligned}
$$

where $T_{t,j}$ is the temperature at day t in year j, Λ_t denotes the seasonality effect and σ_t the seasonal volatility.

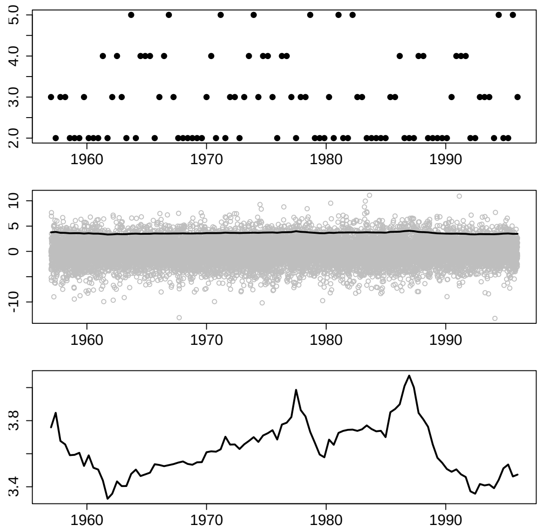


Figure 9: Plot of quantile curve for standardized weather residuals over 40 years at Berlin, 95% quantile, $1967-2006$. Selected bandwidths (upper), observations with estimated the quantile function (middle), the estimated the quantile function (lower).

We are interested specifically in the stochastic risk drivers $\varepsilon_{t,j}$, Figure 9 presents a time series plot of $\widehat{\varepsilon}_{t,j}/\widehat{\sigma}_t$, and the estimated 90% quantile function. By zooming in the curve, we observe a very interesting phenomena: an changing of trend of the standardized residual over years.

To further understand the risk factors, we analyze the quantile functions of $\widehat{\varepsilon}^2_{t,j}$ over 12 years, and average over 4 years for comparison, see Figure 10 and Figure 11. The differences between Berlin and Kaoshiung are easy to see, the variance function has high value from Jan-Feb, while for Berlin the peaks come more in summer. Moreover, there is a tendency for Kaoshiung to be more volatile over time, but this phenomenon does not appear in Berlin.

In addition, our technique can also be used for estimating the function σ_t. We propose four methods: 1, Estimate the median curve of $\widehat{\varepsilon}_{t,j}$ using adaptive technique. 2, Take $\{\widehat{f}_{\varepsilon,0.75}-\widehat{f}_{\varepsilon,0.25}\}/1.34$ (1.34 is the inter quartile range of a standard normal distribution), where $\widehat{f}_{\varepsilon,0.75}$, $\widehat{f}_{\varepsilon,0.25}$ are the adaptive quantile estimators. 3, Estimate the mean curve of $\widehat{\varepsilon}_{t,j}$ using adaptive bandwidth. 4, Estimate the mean function of $\widehat{\varepsilon}_{t,j}$ with fixed bandwidth. The aforementioned methods are compared by testing the normality of $\widehat{\eta}_{t,j}=\widehat{\varepsilon}_{t,j}/\widehat{\sigma}_t$. As according to our normal assumption on $\eta_{t,j}$, a good estimation for σ_t leads to normal standardized residuals $\widehat{\eta}_{t,j}$. Table 9 and 10 summarize statistics from the normality test of standardized residuals from three methods in Berlin and Kaoshiung. It can be seen that Berlin has more normal residuals than Kaoshiung. Method three is al-

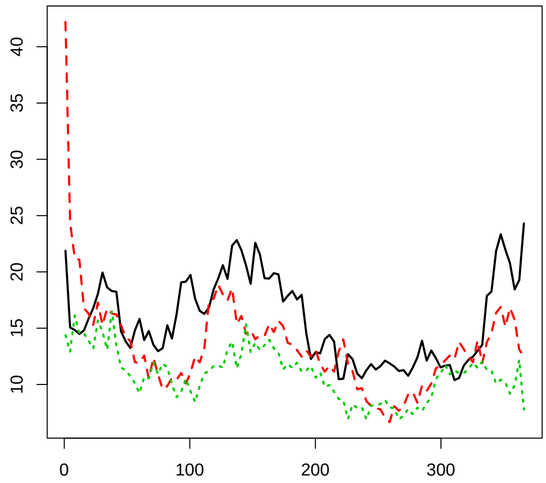


Figure 10: Estimated 90% quantile of variance functions, Berlin, average over $1995-1998$, $1999-2002$ (red), $2003-2006$ (green)

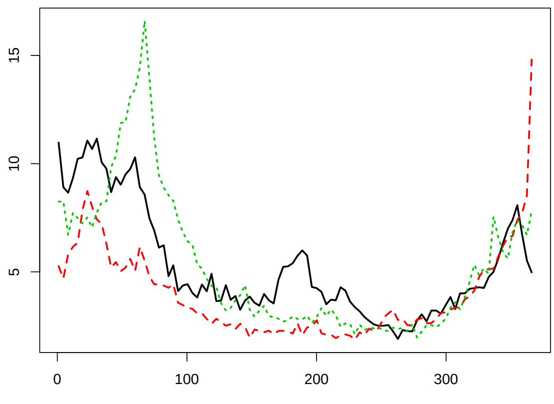


Figure 11: Estimated 90% quantile of variance functions, Kaoshiung, average over $1995-1998$, $1999-2002$ (red), $2003-2006$ (green)

ways better in getting more normal residuals, and method two is compatible with method three. It may be due to the fact that quantiles at higher or lower levels are better to explain the extremes of the volatility function. Method four performs not so well as it is with a fixed bandwidth. Therefore we conclude that our adaptive technique is useful in modeling temperature residuals.

Table 9: P-values of Normality Tests: Berlin

	AD	JB	KS
1	0.000	0.010	0.060
2	0.062	0.000	0.020
3	0.054	0.487	0.171
4	0.009	0.000	0.002

Table 10: P-values of Normality Tests:Kaoshiung

	AD	JB	KS
1	0.000	0.000	0.000
2	1.03e-05	0.077	0.043
3	2.37e-06	0.742	0.674
4	0.000	0.021	0.019

5 Finite Sample Theory

This section discusses some theoretical properties of the proposed estimator $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)$ under a general data distribution. Here $\widehat{k} = \widehat{k}(x)$ is the index selected by the pointwise procedure from Section 2.4. The main "oracle" result shows that $\widehat{\boldsymbol{\theta}}(x)$ is *adaptive* in the sense that it provides nearly the same quality of estimation as the *oracle* estimator $\widetilde{\boldsymbol{\theta}}_{k^*}(x)$ which is the best in the family $\{\widetilde{\boldsymbol{\theta}}_k(x)\}_{k=1}^{K}$. A precise definition of k^* will be given below in term of the *modeling bias.*

5.1 Modeling Bias

The proposed approach for the bandwidth selection suggests to take larger and larger bandwidth until the linear parametric assumption is not significantly violated on the considered interval. The likelihood ratio test statistics $L\big(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \widetilde{\boldsymbol{\theta}}_k(x)\big)$ from (10) are used for this check. The formal definition of the best or oracle choice requires to introduce a measure for the deviation of the function $f(\cdot)$ from its best linear approximation of the form $\Psi^\top \boldsymbol{\theta}$ on the interval of radius h_k considered at step k of the procedure. We follow Spokoiny (2009) who introduced the *modeling bias* for measuring the deviation from the linear parametric structure. Define P_i as the distribution of the observation Y_i. Let also $P_{i,s}$ be a shift of P_i by s, that is, the distribution of $Y_i - s$. Also denote $f_i = f(X_i)$ and $f_i(\boldsymbol{\theta}) = \Psi_i^\top \boldsymbol{\theta}$. In particular, P_{i,f_i} is the distribution of $\varepsilon_i \stackrel{\text{def}}{=} Y_i - f(X_i)$, so that its τ-quantile is zero. The underlying measure $\boldsymbol{P}$ is the product of the measures P_{i,f_i}. Under the linear PA $f(X_i) = f_{\boldsymbol{\theta}}(X_i)$, the corresponding measure $\boldsymbol{P}_{\boldsymbol{\theta}}$ is the product of the $P_{i,f_i(\boldsymbol{\theta})}$:

$$\boldsymbol{P} = \prod_{i=1}^n P_{i,f_i}, \qquad \boldsymbol{P}_{\boldsymbol{\theta}} = \prod_{i=1}^n P_{i,f_i(\boldsymbol{\theta})}.$$

The modeling bias at step k measures the deviation of the true quantile function f from the linear parametric one and it is defined as

$$\begin{aligned} \Delta_k &\stackrel{\text{def}}{=} \inf_{\boldsymbol{\theta}} \Delta_k(\boldsymbol{\theta}), \\ \Delta_k(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} \sum_{i=1}^n \mathcal{K}\bigl(P_{i,f_i}, P_{i,f_i(\boldsymbol{\theta})}\bigr)\, \mathbb{I}\{w_i^{(k)} > 0\}. \end{aligned}$$

Here $\mathcal{K}(P, Q)$ is the Kullback-Leibler divergence between two measures P and Q. The quantity $\Delta_k(\boldsymbol{\theta})$ can be viewed as weighted Kullback-Leibler divergence between $\boldsymbol{P}$ and $\boldsymbol{P}_{\boldsymbol{\theta}}$ localized to the observations in the interval of radius h_k around x. The value Δ_k describes the quality of the best linear approximation on this interval. The *small modeling bias* (SMB) condition manifests that the value Δ_k does not exceed a prescribed quantity $\Delta > 0$, and the oracle choice of the bandwidth h_k is defined as the largest bandwidth among h_k for which the SMB condition is satisfied:

$$k^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{k \le K} \{\Delta_k \le \Delta\}. \tag{15}$$

All the introduced quantities depend on the central point x. Therefore, the parameter $\boldsymbol{\theta}^*$ of the best parametric fit and the oracle bandwidth k^* also depend on x: our approach allows for specifying the best bandwidth for each point separately. Under the measure $\boldsymbol{P}_{\boldsymbol{\theta}^*}$, the estimate $\widetilde{\boldsymbol{\theta}}_k(x)$ is close to $\boldsymbol{\theta}^*$ in the sense that the confidence set $\mathcal{E}_k(\mathfrak{z}_k)$ covers $\boldsymbol{\theta}^*$ with a high probability and the risk $\boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr)$ remains bounded by a fixed constant $\mathcal{R}_r$ for all $k \le K$. The definition of the modeling bias based on the Kullback-Leibler divergence allows to translate this properties to the general case at cost of the additional factor e^{Δ}. More precisely, the following bound holds.

Theorem 5.1. *Let* $\boldsymbol{\theta}^*$ *and* $k^* \le K$ *be such that* $\Delta_{k^*}(\boldsymbol{\theta}^*) \le \Delta$ *. Then for each* $k \le k^*$

$$\boldsymbol{E} \log\left\{1 + \frac{L^r(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*)}{\mathfrak{R}_r}\right\} \le \Delta + 1$$

So, if Δ is small all the confidence or risk bounds continue to apply even in the local nonparametric situation.

5.2 "Oracle" Property

This section presents our main result called the oracle risk bound. The main message of this result is that the adaptive estimator $\widehat{\boldsymbol{\theta}}(x)$ performs nearly as well as the best (oracle) estimator does. Let the bandwidth index k^* be defined by the SMB condition (15) leading to the oracle estimator $\widetilde{\boldsymbol{\theta}}_{k^*}(x)$. The next result claims that for the final estimator $\widehat{\boldsymbol{\theta}}(x)$, the difference

$$L\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}(x)\bigr) = L\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x)\bigr) - L\bigl(W^{(k^*)}, \widehat{\boldsymbol{\theta}}(x)\bigr)$$

is not larger in order than $\mathfrak{z}_{k^*} \mathrm{e}^{\Delta}$. Later we show that the critical value $\mathfrak{z}_{k^*}$ is at most logarithmic in the sample size n. Therefore, the oracle result means that the adaptive estimator $\widehat{\boldsymbol{\theta}}(x)$ belongs with a dominating probability to a confidence set of the oracle.

Theorem 5.2. *Suppose* A.*1*–A.*5. Let* $k^* \le K$ *be such that* $\Delta_{k^*}(\boldsymbol{\theta}) \le \Delta$ *. Then*

$$\boldsymbol{E} \log\left\{1 + \frac{L^r\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}(x)\bigr)}{\mathfrak{R}_r}\right\} \le \alpha + \Delta + \log\bigl(1 + \frac{\mathfrak{z}_{k^*}}{\mathfrak{R}_r}\bigr).$$

6 Appendix

The appendix collects the conditions, technical results, and the proofs. First we fix our assumptions. We assume independent observations $Y_1, \ldots, Y_n$. The results are stated for a deterministic design $X_1, \ldots, X_n$ under mild regularity conditions. The case of a random design can be considered by the usual conditioning argument. Given τ, the quantile function $f(\cdot)$ is defined by the relation $\boldsymbol{P}\bigl\{Y_i > f(X_i)\bigr\} = \tau$. To avoid ambiguous notation, we suppose that this equation has an unique solution for each i. The general case can be easily reduced to this one by standard arguments; see e.g. Koenker (2005). We also denote by P_i the distribution of the residual $\varepsilon_i = Y_i - f(X_i)$ and by $p_i(\cdot)$ its density. Below a point x is fixed and the target of estimation is the quantile $f(x)$. The local parametric approach requires to fix a localizing weighting scheme $W = (w_1, \ldots, w_n)$ and linear parametric family $f_{\boldsymbol{\theta}}(\cdot)$ with $f_{\boldsymbol{\theta}}(X_i) = \Psi_i^\top \boldsymbol{\theta}$, where $\Psi_{i,m} = (X_i - x)^m / m!$ for $m = 0, 1, \ldots, p$.

Our theoretical study can be separated into two parts. An essential and the most difficult part is done under the linear parametric assumption $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$. Then we extend the results to the case when this assumption is approximately fulfilled in a local vicinity of the central point x.

Below a family of localizing weighting schemes $W^{(k)} = \{w_i^{(k)}\}_{i=1}^n$ for $k = 1, \dots, K$ is supposed to be fixed. Our standard proposal is $w_i^{(k)} = K_{\mathrm{loc}}\{(X_i - x)/h_k\}$ for a given kernel $K_{\mathrm{loc}}(\cdot)$ and a sequence of bandwidths $h_1 < h_2 < \ldots < h_K$. Define

$$D_k^2 \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \Psi_i \Psi_i^\top p_i(0) w_i^{(k)} \tag{16}$$

$$V_k^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\bigl\{\nabla L(W^{(k)}, \boldsymbol{\theta}^*)\bigr\} = \tau(1-\tau) \sum_{i=1}^n \Psi_i \Psi_i^\top \bigl|w_i^{(k)}\bigr|^2, \tag{17}$$

$$N_k^{-1/2} \stackrel{\mathrm{def}}{=} \max_{i \le n} \sup_{\boldsymbol{\gamma} \in I\!\!R^p} \frac{|\boldsymbol{\gamma}^\top \Psi_i| \, \mathbb{I}\bigl(w_i^{(k)} > 0\bigr)}{\|V_k \boldsymbol{\gamma}\|} \sqrt{\tau(1-\tau)}. \tag{18}$$

Here D_k^2 and V_k^2 are symmetric $p \times p$ matrices: D_k^2 can be defined similarly to the Fisher information matrix $D_k^2 = -\nabla^2 \boldsymbol{E} L(W^{(k)}, \boldsymbol{\theta}^*)$, while V_k^2 is the covariance matrix of the score $\nabla L(W^{(k)}, \boldsymbol{\theta}^*)$ under the parametric assumption $f \equiv f_{\boldsymbol{\theta}^*}$. In the global parametric situation, and if the likelihood ALD corresponds to the true error distribution, these two matrices coincide. The value N_k can be treated as the local sample size corresponding to the localizing scheme $W^{(k)}$.

The following conditions will be assumed for our results.

A.1 $\{Y_i\}_{i=1}^n$ are independent.

A.2 For some constants $0 < \mathfrak{u}_0 < \mathfrak{u} < 1$,

$$0 < \mathfrak{u}_0 \le \|D_k^{-1} D_{k-1}^2 D_k^{-1}\|_\infty \le \mathfrak{u} < 1.$$

A.3 For a constant $\mathfrak{a} > 0$ and all $k = 1, \dots, K$, it holds

$$V_k^2 \le \mathfrak{a}^2 D_k^2.$$

A.4 For some fixed $\delta < 1/2$ and $\rho > 0$,

$$\bigl|p_i(u)/p_i(0) - 1\bigr| \le \delta, \qquad |u| \le \rho.$$

A.5 The kernel function $K_{\mathrm{loc}}(\cdot)$ is symmetric, $K(0) = 1$, $K(u)$ decreases in $u \ge 0$ and $K(u) = 0$ for $|u| \ge 1$.

The condition A.2 effectively requires that the bandwidth sequence h_k grows geometrically with k. Condition A.3 is the local identifiability condition and it ensures that the local variability of the process $L(W^{(k)}, \boldsymbol{\theta})$ measured by the matrix V_k^2 is not significantly larger than the local information measured by the matrix D_k^2. A.4 only requires that the density functions $p_i(\cdot)$ are uniformly continuous in a vicinity of zero. In particular, the residuals can be unequally distributed. All the results below tacitly assume that the conditions A.1 – A.5 hold.

Below we use generic notation $C = C(\mathrm{A})$ to indicate that a constant C only depends on the constants from conditions A.1 – A.5 like $\mathfrak{a}$, ρ, δ, $\mathfrak{u}_0$, $\mathfrak{u}$, etc. We will also use conditions $(E\mathtt{r})$, $(\mathcal{L}\mathtt{r})$ etc. defined later in section 6.2.1.

6.1 Uniform concentration of the MLEs $\widetilde{\boldsymbol{\theta}}_k(x)$ under $\boldsymbol{P}_{\boldsymbol{\theta}^*}$

The first result explains the localization property of the estimators $\widetilde{\boldsymbol{\theta}}_k(x)$ from (9) under the linear parametric structure of the quantile function, that is, $f(X_i) = \Psi_i^\top \boldsymbol{\theta}^*$. With some value $\mathtt{r}_0$ fixed, define for each $k \le K$ a local elliptic set

$$\Theta_k(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} \bigl\{\boldsymbol{\theta} : \|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}_0\bigr\}$$

with V_k^2 from (17). The question under study is a proper choice of the radius $\mathtt{r}_0$ which ensures a prescribed small deviation probability for the event $\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)$ uniformly in $k \le K$.

Theorem 6.1. *Suppose* $(E\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$, *and there exist constants* $C_1 = C_1(\mathrm{A})$ *and* $C_2 = C_2(\mathrm{A})$ *such that the conditions*

$$\mathtt{r}_0^2 \ge C_1(\mathtt{x} + p), \qquad \rho^2 N_k \ge C_2(\mathtt{x} + p) \tag{19}$$

ensure for $k \le K$

$$\boldsymbol{P}_{\boldsymbol{\theta}^*}\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)\bigr\} \le 2\mathrm{e}^{-\mathtt{x}},$$

$$\boldsymbol{E}_{\boldsymbol{\theta}^*}\bigl[L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr)\, \mathrm{I\!I}\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)\bigr\}\bigr] \le C(\mathrm{A})\mathrm{e}^{-\mathtt{x}}.$$

In particular, a choice $\mathtt{x} = \log(K) + \mathtt{x}_0$ and then $\mathtt{r}_0^2 \ge C_1(\mathtt{x} + p)$ ensures a dominating probability $1 - 2\mathrm{e}^{-\mathtt{x}_0}$ for the joint concentration event

$$\mathcal{A}_1 = \bigcap_{k=1}^{K} \bigl\{\widetilde{\boldsymbol{\theta}}_k(x) \in \Theta_k(\mathtt{r}_0)\bigr\}.$$

In what follows we suppose that the values $\mathtt{x} = \log(K) + \mathtt{x}_0$ and $\mathtt{r}_0$ are fixed in a way that the probability of the set $\mathcal{A}_1$ is sufficiently close to 1. This allows to restrict ourselves to the case when each estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ belongs to the local vicinity $\Theta_k(\mathtt{r}_0)$. The conditions in (19) require that $\mathtt{r}_0^2$ is of order $\log(K)+p$, and the local sample size N_k should be at least of the same order.

6.2 Uniform quadratic approximation of the local excess

The previous subsection stated that the chance for any of the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ lying outside the neighborhood $\Theta_k(\mathtt{r}_0)$ is small, therefore in this subsection, we focus on the stochastic behavior of $\widetilde{\boldsymbol{\theta}}_k$ in $\Theta_k(\mathtt{r}_0)$. The proposed estimation procedure is likelihood-based: all quantities are defined in terms of the quasi log-likelihood functions $L(W^{(k)}, \boldsymbol{\theta})$. Particularly, the properties of the *excess* $L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr) \stackrel{\text{def}}{=} L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr) - L(W^{(k)}, \boldsymbol{\theta}^*)$ plays a very important role in the whole method. The famous Wilks result claims that the excess is asymptotically χ_p^2. Unfortunately the local parametric approach for a narrow local neighborhood of the point x leads to a relatively small effective sample size N_k, and the asymptotic results cannot be validated. The general parametric approach of Spokoiny (2011) though allows to operate with finite samples and it can be directly applied to a local parametric analysis.

It holds

$$
\begin{aligned}
\nabla L(W^{(k)}, \boldsymbol{\theta}^*) &= -\sum_{i=1}^{n} \rho_\tau'(Y_i - \Psi_i^\top \boldsymbol{\theta}^*) w_i^{(k)} \\
&= \sum_{i=1}^{n} \bigl\{ -\tau + \mathbb{I}(Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0) \bigr\} \, \Psi_i \, w_i^{(k)}.
\end{aligned}
$$

Further, for $\boldsymbol{\epsilon} = (\delta, \varrho)$ and D_k^2 from (16), define

$$
\begin{aligned}
D_{\boldsymbol{\epsilon},k}^2 &= D_k^2(1-\delta) - \varrho V_k^2, \\
\boldsymbol{\xi}_{\boldsymbol{\epsilon},k} &\stackrel{\text{def}}{=} D_{\boldsymbol{\epsilon},k}^{-1} \nabla L(W^{(k)}, \boldsymbol{\theta}^*),
\end{aligned}
$$

and similarly for $\underline{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} -\boldsymbol{\epsilon} = (-\delta, -\varrho)$. The values δ, ϱ are assumed to be small enough to ensure that $D_{\boldsymbol{\epsilon},k}^2$ is positive and the value

$$
\alpha_{\boldsymbol{\epsilon},k} \stackrel{\text{def}}{=} \lambda_{\max}\bigl(I_p - D_{\boldsymbol{\epsilon},k} D_{\underline{\boldsymbol{\epsilon}},k}^{-2} D_{\boldsymbol{\epsilon},k}\bigr) \tag{20}
$$

is small as well. Finally, define

$$
\begin{aligned}
\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(W^{(k)}, \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 \\
&= \boldsymbol{\xi}_{\boldsymbol{\epsilon},k}^\top D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2
\end{aligned}
$$

and a similar definition for $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Theorem 6.2. *Under the conditions* (ED_0), (ED_1), $(\mathcal{L}_0)$, *it holds for all* $k \le K$ *and all* $\boldsymbol{\theta} \in \Theta_k(\mathtt{r}_0)$

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\boldsymbol{\epsilon},k} \le L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon},k}\,. \tag{21}$$

Here $\diamondsuit_{\boldsymbol{\epsilon},k}$ *are the random error terms which fulfill with some* $C_1(\mathrm{A})$ *and* $C_2(\mathrm{A})$ *the following conditions: for any* $\mathtt{x} > 0$ *with* $C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A}) \le \mathtt{y}_c$

$$\boldsymbol{P}_{\boldsymbol{\theta}^*}\bigl(\diamondsuit_{\boldsymbol{\epsilon},k}/\varrho > C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p\bigr) \le C(\mathrm{A})\mathrm{e}^{-\mathtt{x}},$$

$$\boldsymbol{E}_{\boldsymbol{\theta}^*}\bigl|\diamondsuit_{\boldsymbol{\epsilon},k}/\varrho\bigr|^r \le C_r(\mathrm{A}),$$

where $\mathtt{y}_c$ *is a constant of order* N_k.

The sandwiching result (21) for each k follows from Theorem 3.1 of Spokoiny (2011). It is only worth mentioning that the local sets $\Theta_k(\mathtt{r}_0)$ are embedded: $\Theta_1(\mathtt{r}_0) \supset \Theta_2(\mathtt{r}_0) \supset \ldots \supset \Theta_K(\mathtt{r}_0)$, so it suffices to check the bound (21) on $\Theta_1(\mathtt{r}_0)$ for each $k \le K$.

The majorization bound (21) yields that the maximum of the process $L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is also sandwiched between the maximum of $\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and of $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ up to a small random error term. Moreover, $\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is quadratic in $\boldsymbol{\theta}$, and its maximum is given by a quadratic form $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2$; similarly for $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$. The next result presents a probabilistic bound for such quadratic forms.

Theorem 6.3. *Assume* A.1 *through* A.5. *There exist* $C_1(\mathrm{A})$ *and* $C_2(\mathrm{A})$ *such that for each* $\mathtt{x}$ *with* $C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p \le \mathtt{y}_c$ *and* $k \le K$, *it holds*

$$\boldsymbol{P}_{\boldsymbol{\theta}^*}\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 > C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p\bigr\} \le 2\mathrm{e}^{-\mathtt{x}}.$$

Furthermore, for $r > 0$ *and* $k \le K$, *it holds*

$$\boldsymbol{E}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2r} \le C_r(\mathrm{A})\,.$$

Consider the random set

$$\mathcal{A}_2 = \bigcap_{k=1}^{K}\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\| \le \mathtt{r}_0\bigr\}.$$

Due to the bound of Theorem 6.3, the choice $\mathtt{r}_0^2 = C_1(\mathrm{A})(\mathtt{x} + \log K) + C_2(\mathrm{A})p$ ensures that the probability of the set $\mathcal{A}_2$ is at most $2\mathrm{e}^{-\mathtt{x}}$.

Below we restrict ourselves to the set $\mathcal{A}$ with $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2$. By construction

$$\boldsymbol{P}(\mathcal{A}) \ge 1 - 4\mathrm{e}^{-\mathtt{x}}$$

and on this set $\widetilde{\boldsymbol{\theta}}_k \in \Theta_k(\mathtt{r}_0)$ and $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\| \le \mathtt{r}_0$ for all $k \le K$.

The results of Theorem 6.2 and 6.3 have a number of important corollaries; cf. Spokoiny (2011).

Corollary 1. *It holds on* $\mathcal{A}$ *for every* $k \le K$

$$\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 - \diamondsuit_{\boldsymbol{\epsilon},k} \le L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*) \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 + \diamondsuit_{\boldsymbol{\epsilon},k}. \tag{22}$$

Corollary 2. *It holds on* $\mathcal{A}$ *for every* $k \le K$

$$\begin{aligned} \bigl\| D_{\boldsymbol{\epsilon},k}\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}^*\bigr\} - \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \bigr\|^2 &\le 4\diamondsuit_{\boldsymbol{\epsilon},k} + \alpha_{\boldsymbol{\epsilon},k}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2, \\ \bigl\| D_{\boldsymbol{\epsilon},k}\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}^*\bigr\} \bigr\| &\le 2\diamondsuit_{\boldsymbol{\epsilon},k}^{1/2} + \bigl(1 + \alpha_{\boldsymbol{\epsilon},k}^{1/2}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|. \end{aligned} \tag{23}$$

The result of Corollary 1 can be viewed as a non-asymptotic version of Wilks Theorem. It claims that the twice excess $2L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*)$ can be approximated by the quadratic form $\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2$. Moreover, the vector $\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}$ is asymptotically normal under usual assumptions by the central limit theorem, thus the twice excess is asymptotically χ^2_p.

One can summarize the obtained general results as follows. On the set $\mathcal{A}$ of dominating probability, each estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ belongs to the local vicinity $\Theta_k(\mathtt{r}_0)$ which yields the bounds (22), (23). Moreover, the random quantities $\diamondsuit_{\boldsymbol{\epsilon},k}$ and $\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}$ obey the deviation and moment bounds of Theorem 6.2 and Theorem 6.3.

6.2.1 Conditions from Spokoiny (2011)

Here we list the conditions from Spokoiny (2011) which are assumed to be fulfilled for each local likelihood $L(W^{(k)}, \boldsymbol{\theta})$, $k \le K$. Some value $\mathtt{r}_0$ is assumed to be fixed for all conditions. It separates the local zone of local quadratic approximation and the large deviation zone. The assumption are stated under the true data distribution $\boldsymbol{P}$. However, we apply the assumptions only in the case of linear parametric structure with $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$. Define

$$\begin{aligned} \zeta_k(\boldsymbol{\theta}) &\stackrel{\text{def}}{=} L(W^{(k)}, \boldsymbol{\theta}) - \boldsymbol{E}L(W^{(k)}, \boldsymbol{\theta}) \\ &= -\sum_{i=1}^{n}\Bigl\{\rho_\tau(Y_i - \Psi_i^\top\boldsymbol{\theta}) - \boldsymbol{E}\{\rho_\tau(Y_i - \Psi_i^\top\boldsymbol{\theta})\Bigr\} w_i^{(k)}. \end{aligned}$$

Also denote $\nabla\zeta_k(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}}\zeta(\boldsymbol{\theta})$. The majorization bound (21) of Theorem 6.2 is stated in Spokoiny (2011) under the following conditions.

$(\boldsymbol{ED_0})$ There exists a positive symmetric matrix V_k^2, and constants $\mathtt{g} > 0$ and $\nu_0 \ge 1$ such that $\operatorname{Var}\bigl\{\nabla\zeta_k(\boldsymbol{\theta})\bigr\} \le V_k^2$ and for all λ with $|\lambda| \le \mathtt{g}$,

$$\sup_{\gamma \in I\!\!R^{p+1}} \log \boldsymbol{E}\exp\Bigl\{\lambda\frac{\gamma^\top\nabla\zeta_k(\boldsymbol{\theta}^*)}{\|V_k\gamma\|}\Bigr\} \le \nu_0^2\lambda^2/2.$$

With this matrix V_k, define the local set

$$\Theta_k(\mathtt{r}_0) = \{\boldsymbol{\theta} : \|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}_0\}.$$

($\boldsymbol{ED_1}$) For each $\mathtt{r} \le \mathtt{r}_0$, there exists a constant $\varrho(\mathtt{r}) \le 1/2$ such that it holds for all $\boldsymbol{\theta} \in \Theta_k(\mathtt{r}_0)$ and $|\lambda| \le \mathtt{g}$:

$$\sup_{\boldsymbol{\gamma} \in I\!\!R^{p+1}} \log \boldsymbol{E} \exp\biggl\{ \lambda \frac{\boldsymbol{\gamma}^\top \{\nabla\zeta_k(\boldsymbol{\theta}) - \nabla\zeta_k(\boldsymbol{\theta}^*)\}}{\varrho(\mathtt{r})\|V_k\boldsymbol{\gamma}\|} \biggr\} \le \nu_0^2\lambda^2/2;$$

($\boldsymbol{\mathcal{L}_0}$) *There are a positive matrix D_k and for each $\mathtt{r} \le \mathtt{r}_0$ and a constant $\delta(\mathtt{r}) \le 1/2$, such that it holds for all $\boldsymbol{\theta} \in \Theta_k$,*

$$\left| \frac{-2\boldsymbol{E}L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} - 1 \right| \le \delta(\mathtt{r});$$

($\boldsymbol{E}\mathtt{r}$) For any $\mathtt{r} \ge \mathtt{r}_0$, there exist a value $\mathtt{g}(\mathtt{r}) > 0$ and a constant ν_0 such that for all $|\lambda| \le \mathtt{g}(\mathtt{r})$,

$$\sup_{\boldsymbol{\gamma} \in I\!\!R^{p+1}} \sup_{\boldsymbol{\theta} \in \Theta_k(\mathtt{r})} \log \boldsymbol{E} \exp\biggl\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta_k(\boldsymbol{\theta})}{\|V_k\boldsymbol{\gamma}\|} \biggr\} \le \nu_0^2\lambda^2/2.$$

($\boldsymbol{\mathcal{L}}\mathtt{r}$) For each $\mathtt{r} \ge \mathtt{r}_0$ and any $\boldsymbol{\theta}$ with $\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$,

$$\frac{-\boldsymbol{E}L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \ge \mathtt{b}(\mathtt{r}) > 0.$$

All these conditions are assumed to be fulfilled for each $k \le K$. Conditions $(ED_0), (ED_1)$, $(\mathcal{L}_0)$ are local conditions which should be applied on the local set $\Theta_k(\mathtt{r}_0)$, while $(\mathcal{L}\mathtt{r}), (E\mathtt{r})$ are global conditions which we apply on the complement of $\Theta_k(\mathtt{r}_0)$. Also $(ED_0), (ED_1), (E\mathtt{r})$ are smoothness or moment assumptions on the log likelihood process, and the conditions $(\mathcal{L}_0), (\mathcal{L}\mathtt{r})$ ensure the identifiability properties.

6.2.2 Proof of $(\boldsymbol{E}\mathtt{r})$, $(\boldsymbol{ED_0})$ and $(\boldsymbol{ED_1})$.

Let us fix some $k \le K$. Let N_k be the local sample size for the weighting scheme $W^{(k)}$; see (18). Let also $\mathtt{r}_0$ by fixed in a way that $\mathtt{r}_0|\Psi_i| \le \rho N_k$ for all i with $w_i^{(k)} > 0$, that is, for all X_i with $|X_i - x| \le h_k$.

First we check $(E\mathtt{r})$. It holds by definition

$$\begin{aligned}
\nabla\zeta_k(\boldsymbol{\theta}) &= \sum_{i=1}^n \Psi_i \bigl[1\!\!\mathrm{I}(Y_i - \Psi_i^\top\boldsymbol{\theta} < 0) - \boldsymbol{P}(Y_i - \Psi_i^\top\boldsymbol{\theta} < 0) \bigr] w_i^{(k)} \\
&= \sum_{i=1}^n \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)}
\end{aligned}$$

with $\varepsilon_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{1}(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0) - \boldsymbol{P}(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$. Obviously $\mathbb{1}(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$ is a Bernoulli random variable with the parameter $q_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \boldsymbol{P}(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$ and

$$\log \boldsymbol{E} \exp\{\delta \varepsilon_i(\boldsymbol{\theta})\} = \log\bigl\{1 - q_i(\boldsymbol{\theta}) + q_i(\boldsymbol{\theta}) \mathrm{e}^{\delta}\bigr\} - \delta q_i(\boldsymbol{\theta}).$$

The function $f(\delta) \stackrel{\text{def}}{=} \log\bigl(1 - q + q\mathrm{e}^{\delta}\bigr) - q\delta$ fulfills for any $q < 1$

$$f(0) = 0, \qquad f'(0) = 0, \qquad f''(\delta) \le q(1-q)\mathrm{e}^{\delta}.$$

This implies

$$\log \boldsymbol{E} \exp\{\delta \varepsilon_i(\boldsymbol{\theta})\} \le q_i(\boldsymbol{\theta})\{1 - q_i(\boldsymbol{\theta})\}\nu_0^2 \delta^2/2, \qquad |\delta| \le \mathtt{g}_1$$

for a constant $\nu_0 \ge 1$ depending on $\mathtt{g}_1$ only. Therefore, it holds for any $\boldsymbol{\gamma} \in I\!\!R^{p+1}$ and $\rho > 0$ with $\rho|\boldsymbol{\gamma}^\top \Psi_i| \le \mathtt{g}_1$ that,

$$\begin{aligned}
\log \boldsymbol{E} \exp\{\rho \boldsymbol{\gamma}^\top \nabla \zeta_k(\boldsymbol{\theta})\} &\le \log \boldsymbol{E} \exp\left\{\rho \sum_{i=1}^n \boldsymbol{\gamma}^\top \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)}\right\} \\
&\le \sum_{i=1}^n \log \boldsymbol{E} \exp\bigl\{\rho \boldsymbol{\gamma}^\top \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)}\bigr\} \\
&\le \sum_{i=1}^n \rho^2 \bigl|\boldsymbol{\gamma}^\top \Psi_i w_i^{(k)}\bigr|^2 q_i(\boldsymbol{\theta})\{1 - q_i(\boldsymbol{\theta})\}\nu_0^2/2 \\
&\le \nu_0^2 \rho^2 \|V_k(\boldsymbol{\theta})\boldsymbol{\gamma}\|^2/2,
\end{aligned}$$

where

$$V_k^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n q_i(\boldsymbol{\theta})\{1 - q_i(\boldsymbol{\theta})\}\Psi_i \Psi_i^\top \bigl|w_i^{(k)}\bigr|^2.$$

This yields (ED_0) with $V_k^2 \stackrel{\text{def}}{=} V_k^2(\boldsymbol{\theta}^*)$ and $\mathtt{g} = \mathtt{g}_1 N_k^{1/2}$; see (18). Furthermore, the linear PA $f \equiv f_{\boldsymbol{\theta}^*}$ yields $q_i(\boldsymbol{\theta}^*) = \tau$ and hence

$$V_k^2(\boldsymbol{\theta}^*) = 4\tau(1-\tau)\overline{V}_k^2$$

for

$$\overline{V}_k^2 \stackrel{\text{def}}{=} \frac{1}{4} \sum_{i=1}^n \Psi_i \Psi_i^\top \bigl|w_i^{(k)}\bigr|^2.$$

For any $\boldsymbol{\theta} \in \Theta$, it obviously holds $V_k^2(\boldsymbol{\theta}) \le \overline{V}_k^2 \le \bigl\{4\tau(1-\tau)\bigr\}^{-1} V_k^2$, and thus $(E\mathtt{r})$ is fulfilled with $\mathtt{g}^2(\mathtt{r}) \equiv 4\tau(1-\tau)N_k \mathtt{g}_1^2$.

Next we check the local condition (ED_1). For $\mathtt{r} \le \mathtt{r}_0$ and $\boldsymbol{\theta} \in \Theta_k(\mathtt{r})$, it holds

$$\nabla\zeta_k(\boldsymbol{\theta}) - \nabla\zeta_k(\boldsymbol{\theta}^*) = \sum_{i=1}^n \Psi_i\bigl\{\varepsilon_i(\boldsymbol{\theta}) - \varepsilon_i(\boldsymbol{\theta}^*)\bigr\} w_i^{(k)}.$$

Similarly to the above, the identity $\boldsymbol{E}\bigl\{\varepsilon_i(\boldsymbol{\theta}) - \varepsilon_i(\boldsymbol{\theta}^*)\bigr\} = q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)$ implies

$$\begin{aligned}
&\log \boldsymbol{E} \exp\bigl[\lambda\boldsymbol{\gamma}^\top\bigl\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\bigr\}\bigr] \\
&\quad \le 2\nu_0^2\lambda^2\|\overline{V}_k\boldsymbol{\gamma}\|^2 \max_{i\le n}\bigl|q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)\bigr| \, \mathrm{I\!I}\bigl(w_i^{(k)} > 0\bigr) \\
&\quad \le \omega(\mathtt{r})\nu_0^2\lambda^2\|\overline{V}_k\boldsymbol{\gamma}\|^2/2
\end{aligned}$$

with

$$\omega(\mathtt{r}) \stackrel{\mathrm{def}}{=} 4\max_{i\le n} \sup_{\boldsymbol{\theta}\in\Theta_k(\mathtt{r}_0)} \bigl\{q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)\bigr\} \, \mathrm{I\!I}\bigl(w_i^{(k)} > 0\bigr).$$

Further, for any $\boldsymbol{\theta} \in \Theta_k(\mathtt{r})$, it holds $\bigl|\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| w_i^{(k)} \le \mathtt{r}/N_k$

$$\begin{aligned}
|q_i(\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}^*)| \, \mathrm{I\!I}\bigl(w_i^{(k)} > 0\bigr) &\le C|\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \, \mathrm{I\!I}\bigl(w_i^{(k)} > 0\bigr) \\
&\le C N_k^{-1/2}\|\overline{V}_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le C N_k^{-1/2}\mathtt{r},
\end{aligned}$$

and (ED_1) holds with $\varrho(\mathtt{r}) = N_k^{-1/2}\mathtt{r}$.

6.2.3 The $(\mathcal{L}_0)$ and $(\mathcal{L}r)$ conditions

These identifiability conditions will be checked under the measure $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ corresponding to the linear quantile function $f(\cdot) = f_{\boldsymbol{\theta}^*}(\cdot)$. It holds

$$\nabla \boldsymbol{E}_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}) = -\sum_{i=1}^n \Psi_i\Bigl\{\tau - \boldsymbol{P}\bigl(Y_i - \Psi_i^\top\boldsymbol{\theta} < 0\bigr)\Bigr\} w_i^{(k)}$$

and

$$-\nabla^2 \boldsymbol{E}_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}) = \sum_{i=1}^n \Psi_i\Psi_i^\top p_i\bigl(\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr) w_i^{(k)} \stackrel{\mathrm{def}}{=} D_k^2(\boldsymbol{\theta}).$$

Now the Taylor expansion of $-\boldsymbol{E}_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta})$ at the extreme point $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ implies

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}) - \boldsymbol{E}_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}^*) &= -\sum_{i=1}^n |\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 p_i\bigl(\Psi_i^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\bigr) w_i^{(k)}/2 \\
&= -(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D_k^2(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2.
\end{aligned}$$

for some $\boldsymbol{\theta}^{\circ} \in [\boldsymbol{\theta}, \boldsymbol{\theta}^{*}]$. Further, for any $\boldsymbol{\theta} \in \Theta_k(\mathtt{r}_0)$, it holds by A.4

$$\Bigl| \frac{p_i\bigl(\Psi_i^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr)}{p_i(0)} - 1 \Bigr| \, \mathbb{I}\bigl(w_i^{(k)} > 0\bigr) \le \delta,$$

and $(\mathcal{L}_0)$ follows. The global identifiability condition $(\mathcal{L}r)$ is fulfilled if $\mathtt{r}^2 \ge C_1(\mathtt{x} + p)$ for some fixed constants C_1; see Spokoiny (2011), Section 5.3, for more details.

6.3 Theorem for critical values

The theorem below assures an upper bound for the critical values $\mathfrak{z}_k$ constructed in Section 2.5. To avoid technical burdens, we restrict the analysis to the random set $\mathcal{A}$ and discard the large deviation probability part on its complement. The notation $\boldsymbol{P}'(B)$ for a set B means $\boldsymbol{P}(B \cap \mathcal{A})$.

Theorem 6.4. *Suppose that $r > 0, \alpha > 0$. There exist constants a_0, a_1 s.t. the propagation condition is fulfilled with the choice of*

$$\mathfrak{z}_k = a_0 + \log(\alpha^{-1}) + a_1 r(K - k) + r \log(p) \tag{24}$$

Proof. First we bound the quantity $L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr)$ on the random set $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2$. The majorization (21) and its corollary (22) yield on $\mathcal{A}$ with $\boldsymbol{u}_{\ell k} \stackrel{\text{def}}{=} D_{\underline{\boldsymbol{\epsilon}},k}\bigl(\widetilde{\boldsymbol{\theta}}_\ell(x) - \boldsymbol{\theta}^*\bigr)$

$$\begin{aligned}
L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) &= L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr) - L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \boldsymbol{\theta}^*\bigr). \\
&\le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 - \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \boldsymbol{\theta}^*\bigr) + \diamondsuit_{\boldsymbol{\epsilon},k} \\
&= \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 - \boldsymbol{u}_{\ell k}^{\top}\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},k} + \|\boldsymbol{u}_{\ell k}\|^2/2 + 2\diamondsuit_{\boldsymbol{\epsilon},k} \\
&\le \bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\| + \|\boldsymbol{u}_{\ell k}\|\bigr)^2/2 + 2\diamondsuit_{\boldsymbol{\epsilon},k} \\
&\le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 + \|\boldsymbol{u}_{\ell k}\|^2 + 2\diamondsuit_{\boldsymbol{\epsilon},k},
\end{aligned} \tag{25}$$

where we used the fact that $\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},k}\| \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|$. It is not difficult to see that

$$\|\boldsymbol{u}_{\ell k}\|^2 = \|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-1} D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*\bigr)\|^2 \le \|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-2} D_{\underline{\boldsymbol{\epsilon}},k}\|_\infty \, \|D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*\bigr)\|^2.$$

By construction $D_{\boldsymbol{\epsilon},k}^2 \le D_k^2 \le D_{\underline{\boldsymbol{\epsilon}},k}^2$ and the definition (20) implies by $\alpha_{\boldsymbol{\epsilon},k} \le 1/2$

$$D_{\underline{\boldsymbol{\epsilon}},k}^2 \le (1 - \alpha_{\boldsymbol{\epsilon},k})^{-1} D_{\boldsymbol{\epsilon},k}^2 \le 2 D_{\boldsymbol{\epsilon},k}^2 .$$

Now it follows from condition A.2 that

$$\|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-2} D_{\underline{\boldsymbol{\epsilon}},k}\|_\infty \le 2\|D_k D_\ell^{-2} D_k\|_\infty \le \begin{cases} 2/\mathrm{u}_0^{k-\ell}, & k > \ell, \\ 2\mathrm{u}^{\ell-k}, & k < \ell. \end{cases} \tag{26}$$

Corollary 2 implies

$$\|D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_{\ell}(x)-\boldsymbol{\theta}^{*}\bigr)\| \le 2\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \bigl(1+\alpha_{\boldsymbol{\epsilon},\ell}^{1/2}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\| \le 2\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + 2\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|. \tag{27}$$

We also use that $\boldsymbol{E}_{\boldsymbol{\theta}^{*}}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2r} \le p^{r}C_{r}(\mathrm{A})$ for all $k \le K$. Now it holds from (25), (26), and (27) for $k > \ell$

$$\begin{aligned}\boldsymbol{E}'_{\boldsymbol{\theta}^{*}} L^{r}\bigl(W^{(k)},\widetilde{\boldsymbol{\theta}}_{k}(x),\widetilde{\boldsymbol{\theta}}_{\ell}(x)\bigr) \;&\le\; \boldsymbol{E}'_{\boldsymbol{\theta}^{*}}\Bigl[\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2} + 8\mathfrak{u}_{0}^{-k+\ell}\bigl(\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|\bigr)^{2} + 2\diamondsuit_{\boldsymbol{\epsilon},k}\Bigr]^{r} \\ &\le\; C(\mathrm{A})p^{r}\mathfrak{u}_{0}^{-r(k-\ell)}. \end{aligned}\tag{28}$$

Similarly one can show that for $k < \ell$ by $\mathfrak{u} < 1$

$$\begin{aligned}\boldsymbol{E}'_{\boldsymbol{\theta}^{*}} L^{r}\bigl(W^{(k)},\widetilde{\boldsymbol{\theta}}_{k}(x),\widetilde{\boldsymbol{\theta}}_{\ell}(x)\bigr) \;&\le\; \boldsymbol{E}'_{\boldsymbol{\theta}^{*}}\bigl[\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2} + 8\bigl(\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|\bigr)^{2} + 2\diamondsuit_{\boldsymbol{\epsilon},k}\bigr]^{r} \\ &\le\; C(\mathrm{A})p^{r}. \end{aligned}$$

Also by Theorem 6.3 for $\mathtt{x} > 0$

$$\boldsymbol{P}_{\boldsymbol{\theta}^{*}}\bigl\{L\bigl(W^{(k)},\widetilde{\boldsymbol{\theta}}_{k}(x),\widetilde{\boldsymbol{\theta}}_{\ell}(x)\bigr) > C_{1}p + C_{2}\mathtt{x}\bigr\} \le 2\mathrm{e}^{-\mathtt{x}}. \tag{29}$$

These bounds can be used to check that the critical value $\mathfrak{z}_{k}$ which is selected in the form (24) to ensure the propagation condition in (11). Consider a random set $\mathcal{B}_{\ell} \stackrel{\mathrm{def}}{=} \{\widehat{k}(x) = \ell\}$, By definition of $\widehat{k}$, when $\mathcal{B}_{\ell}$ happens, at least one of the estimator $\widetilde{\boldsymbol{\theta}}_{\ell+1}(x)$ must be not accepted, that is,

$$\mathcal{B}_{\ell} \subseteq \bigcup_{m=1}^{\ell}\Bigl\{L\bigl(W^{(m)},\widetilde{\boldsymbol{\theta}}_{m}(x),\widetilde{\boldsymbol{\theta}}_{\ell+1}(x)\bigr) > \mathfrak{z}_{m}\Bigr\}.$$

The bounds (28) and (29) yield by the Cauchy-Schwarz inequality

$$\begin{aligned}&\boldsymbol{E}'_{\boldsymbol{\theta}^{*}} L^{r}\bigl(W^{(k)},\widetilde{\boldsymbol{\theta}}_{k}(x),\widehat{\boldsymbol{\theta}}_{k}(x)\bigr) \\ &\quad\le \sum_{\ell=1}^{k}\bigl[\boldsymbol{E}'_{\boldsymbol{\theta}^{*}} L^{2r}\bigl(W^{(k)},\widetilde{\boldsymbol{\theta}}_{k}(x),\widetilde{\boldsymbol{\theta}}_{\ell}(x)\bigr)\bigr]^{1/2}\bigl[\boldsymbol{P}'_{\boldsymbol{\theta}^{*}}(\mathcal{B}_{\ell})\bigr]^{1/2} \\ &\quad\le C(\mathrm{A})p^{2r}\sum_{\ell=1}^{k}\mathfrak{u}_{0}^{-2r(k-\ell)}\bigl[\boldsymbol{P}'_{\boldsymbol{\theta}^{*}}(\mathcal{B}_{\ell})\bigr]^{1/2} \\ &\quad\le C(\mathrm{A})p^{2r}\sum_{\ell=2}^{k}\mathfrak{u}_{0}^{-2r(k-\ell)}\Biggl(\sum_{m=1}^{\ell}\boldsymbol{P}'_{\boldsymbol{\theta}^{*}}\Bigl\{L\bigl(W^{(m)},\widetilde{\boldsymbol{\theta}}_{m}(x),\widetilde{\boldsymbol{\theta}}_{\ell+1}(x)\bigr) > \mathfrak{z}_{m}\Bigr\}\Biggr)^{1/2}.\end{aligned}$$

Fix $c_{0} > \log(\mathfrak{u}_{0}^{-1})$ and consider $\mathfrak{z}_{m} = C_{1}p + C_{2}\mathtt{x}_{m}$ with $\mathtt{x}_{m} = 2c_{0}r(K-m) + 2\mathtt{x}$ for some

$\mathtt{x}$. Then (29) implies

$$
\begin{aligned}
&\boldsymbol{E}'_{\boldsymbol{\theta}^*}\bigl[L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)\bigr] \\
&\qquad \le C(\mathrm{A}) p^{2r} \sum_{\ell=2}^{K} \mathfrak{u}_0^{-2r(K-\ell)} \Bigl(\sum_{m=1}^{\ell} 2\mathrm{e}^{-\mathtt{x}_m} \Bigr)^{1/2} \\
&\qquad \le C(\mathrm{A}) p^{2r} \mathrm{e}^{-\mathtt{x}} \sum_{\ell=2}^{K} \exp\bigl[-2r(K-\ell)\bigl\{c_0 - \log(1/\mathfrak{u}_0)\bigr\}\bigr] \\
&\qquad \le C(\mathrm{A}) p^{2r} \mathrm{e}^{-\mathtt{x}}
\end{aligned}
$$

and the bound (11) follows with $\mathtt{x} = \log(1/\alpha) + r\log(p) + a_0$ for a proper a_0. □

6.4 Propagation Property and Stability

The oracle result is a consequence of two properties of the procedure: propagation under homogeneity and stability. The first one means that the algorithm does not terminate for $k < k^*$ (no false alarm) with a high probability. The stability property ensures that the estimation quality will not essentially deteriorate in the steps "after propagation" for $k > k^*$.

By construction, the procedure described in Section 2 provides the prescribed performance if the true quantile function $f(\cdot)$ follows the parametric model: at any intermediate step $k < K$ the non-adaptive estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ and the adaptive estimator $\widehat{\boldsymbol{\theta}}_k(x)$ coincide with high probability yielding that $\boldsymbol{E}_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)$ is small. The next theorem claims a similar performance of the k step estimator $\widehat{\boldsymbol{\theta}}_k(x)$ under the true nonparametric model $f(\cdot)$, however, the propagation property is only guaranteed for $k \le k^*$, that is, while the SMB assumption is fulfilled.

Theorem 6.5. *Assume $\Delta_{k^*}(\boldsymbol{\theta}) \le \Delta$ for some k^*. Then for any $k \le k^*$*

$$
\boldsymbol{E} \log\bigl\{1 + L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)/\mathcal{R}_r\bigr\} \le \Delta + \alpha, \tag{30}
$$

The result of the theorem follows from the next general inequality; see e.g. Spokoiny (2009).

Lemma 2. *Let $\boldsymbol{P}$, $\boldsymbol{P}_0$, be two measures s.t. $\boldsymbol{E}\log(d\boldsymbol{P}/d\boldsymbol{P}_0) \le \Delta < \infty$. For any random variable Z with $\boldsymbol{E}_0 Z < \infty$, it holds $\boldsymbol{E}\log(1+Z) \le \Delta + \boldsymbol{E}_0 Z$.*

The propagation result (30) explains well the behavior of the procedure for the first k^* steps. In addition, we also need a *stability property* which makes sure that at the further steps of the algorithm for $k > k^*$, the quality of the obtained adaptive estimator $\widehat{\boldsymbol{\theta}}_k(x)$ will not significantly deteriorate. The stability property can be stated as follows.

Theorem 6.6. *The adaptive estimator* $\widehat{\boldsymbol{\theta}}(x)$ *fulfills*

$$L\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}(x)\big)\, \mathrm{I\!I}\{\widehat{k}(x) > k^*\} \le \mathfrak{z}_{k^*}.$$

In other words, if $\widehat{k}(x) \ge k^*$, then $\widehat{\boldsymbol{\theta}}(x)$ belongs to the confidence set $\mathcal{E}_{k^*}(\mathfrak{z}_{k^*})$ of the oracle estimator $\widetilde{\boldsymbol{\theta}}_{k^*}(x)$. This assertion follows from the setup of our procedure because the estimate $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}(x)}(x)$ is accepted. If $\widehat{k}(x) > k^*$, it should be consistent with $\widetilde{\boldsymbol{\theta}}_{k^*}(x)$, and thus it belongs to the confidence set of $\widetilde{\boldsymbol{\theta}}_{k^*(x)}(x)$.

6.5 Proof of the "oracle" property

The *propagation* and *stability* results yield

$$\begin{aligned}
&\boldsymbol{E} \log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \\
&= \boldsymbol{E}\left[\log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \mathrm{I\!I}(\widehat{k} \le k^*)\right] \\
&\qquad + \boldsymbol{E}\left[\log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \mathrm{I\!I}(\widehat{k} > k^*)\right] \\
&\le \varDelta + \boldsymbol{E}_{\boldsymbol{\theta}^*}\left[\frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right] \\
&\qquad + \boldsymbol{E} \log\left[1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\, \mathrm{I\!I}(\widehat{k} > k^*)\right] \\
&\le \varDelta + \rho + \log(1 + \mathfrak{z}_{k^*}/\mathcal{R}_r)
\end{aligned}$$

References

Belloni, A. and Chernozhukov, V. (2010). l1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39(1):82–130.

Cai, Z. and Wang, X. (2008). Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics*, 147:120–130.

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, 103(484):1595–1608.

Embrechts, P., M. A. and Straumann, D. (1999). Correlation and dependency in risk management: Properties and pitfalls. *Risk Management: Value at Risk and Beyond, Cambridge University Press*, pages 176–223.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles. *Journal of Business and Economic Statistics*, 22:367–381.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall.

Fan, J., Hu, T.-C., and Truong, Y. K. (1994). Robust non-parametric function estimation. *Scandinavian Journal of Statistics*, 21:433–446.

Fitzenberger, B. and Wilke, R. A. (2006). Using quantile regression for duration analysis. *Modern Econometric Analysis*, 90(1):105–120.

Härdle, W. K., Ritov, Y., and Song, S. (2012). Partial linear quantile regression and bootstrap confidence bands. *Journal of Multivariate Analysis, forthcoming.*

Härdle, W. K. and Song, S. (2010). Confidence bands in quantile regression. *Econometric Theory*, 26:1180–1200.

Hauksson, A. H., Michel, D., Thomas, D., Ulrich, M., and Gennady, S. (2001). Multivariate extremes, aggregation and risk estimation. *Quantitative Finance*, 1:79–75.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

James, G. M., Hastie, T. J., and Sugar, C. A. (2010). Principal component models for sparse functional data. *Biometrika*, 87:587–602.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. (2010). Additive models for quantile regression: Model selection and confidence banddaids. *Manuscript.*

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrika*, 46(1):33–50.

Koenker, R. and Ferreira, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statisitical Association*, 94:1296–1310.

Kong, E., Linton, O., and Xia, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory*, 26:159–166.

Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Annals of Statistics*, 25(3):929–947.

Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probability Theory and Realated Fields*, 135:335–362.

Portnoy, S. and Koenker, R. (1989). Adaptive l estimation of linear models. *Annals of Statistics*, 17:362–81.

Portnoy, S. and Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error vs. absolute-error estimators, with discussion. *Statistical Science*, 12:279–300.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression.* Cambridge University Press.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value at risk. *The Annals of Statistics*, 37(3):1405–1436.

Spokoiny, V. (2011). Parametric estimation. Finite sample theory. *Sumitted for publication.* Available at http://arxiv.org/abs/1111.3029.

Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807.

Wu, T. Z., Yu, K., and Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, 101:1607–1621.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93:228–237.