Confidence Corridors for Multivariate Generalized Quantile Regression*

Shih-Kang Chao[†] Katharina Proksch[‡] Holger Dette[‡]
Wolfgang Härdle[†§]

May 12, 2014

Abstract

We focus on the construction of confidence corridors for multivariate nonparametric generalized quantile regression functions. This construction is based on asymptotic results for the maximal deviation between a suitable nonparametric estimator and the true function of interest which follow after a series of approximation steps including a Bahadur representation, a new strong approximation theorem and exponential tail inequalities for Gaussian random fields.

As a byproduct we also obtain confidence corridors for the regression function in the classical mean regression. In order to deal with the problem of slowly decreasing error in coverage probability of the asymptotic confidence corridors, which results in meager coverage for small sample sizes, a simple bootstrap procedure is designed based on the leading term of the Bahadur representation. The finite sample properties of both procedures are investigated by means of a simulation study and it is demonstrated that the bootstrap procedure considerably outperforms the asymptotic bands in terms of coverage accuracy. Finally, the bootstrap confidence corridors are used to study the efficacy of the National Supported Work Demonstration, which is a randomized employment enhancement program launched in the 1970s. This article has supplementary materials online.

Keywords: Bootstrap; Expectile regression; Goodness-of-fit tests; Quantile treatment effect; Smoothing and nonparametric regression.
JEL: C2, C12, C14

*Financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Economic Risk" (Teilprojekt B1), SFB 823 "Statistical modeling of nonlinear dynamic processes" (Teilprojekt C1, C4) and Einstein Foundation Berlin via the Berlin Doctoral Program in Economics and Management Science (BDPEMS) are gratefully acknowledged.

†Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. email: shih-kang.chao@cms.hu-berlin.de; haerdle@wiwi.hu-berlin.de.

‡Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany. email: katharina.proksch@rub.de; holger.dette@rub.de.

§Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore.

1. Introduction

Mean regression analysis is a widely used tool in statistical inference for curves. It focuses on the center of the conditional distribution, given d-dimensional covariates with $d \geq 1$. In a variety of applications though the interest is more in tail events, or even tail event curves such as the conditional quantile function. Applications with a specific demand in tail event curve analysis include finance, climate analysis, labor economics and systemic risk management.

Tail event curves have one thing in common: they describe the likeliness of extreme events conditional on the covariate $\boldsymbol{X}$. A traditional way of defining such a tail event curve is by translating "likeliness" with "probability" leading to conditional quantile curves. Extreme events may alternatively be defined through conditional moment behaviour leading to more general tail descriptions as studied by Newey and Powell (1987) and Jones (1994). We employ this more general definition of generalized quantile regression (GQR), which includes, for instance, expectile curves and study statistical inference of GQR curves through confidence corridors.

In applications parametric forms are frequently used because of practical numerical reasons. Efficient algorithms are available for estimating the corresponding curves. However, the "monocular view" of parametric inference has turned out to be too restrictive. This observation prompts the necessity of checking the functional form of GQR curves. Such a check may be based on testing different kinds of variation between a hypothesized (parametric) model and a smooth alternative GQR. Such an approach though involves either an explicit estimate of the bias or a pre-smoothing of the "null model". In this paper we pursue the Kolmogorov-Smirnov type of approach, that is, employing the maximal deviation between the null and the smooth GQR curve as a test statistic. Such a model check has the advantage that it may be displayed graphically as a confidence corridor (CC; also called "simultaneous confidence band" or "uniform confidence band/region") but has been considered so far only for univariate covariates. The basic technique for constructing CC of this type is extreme value theory for the sup-norm of an appropriately centered nonparametric estimate of the quantile curve.

For a one-dimensional predictor confidence corridors were developed under various settings. Classical one-dimensional results are confidence bands constructed for histogram estimators by Smirnov (1950) or more general one-dimensional kernel density estimators by Bickel and Rosenblatt (1973). The results were extended to a univariate nonparametric mean regression setting by Johnston (1982), followed by Härdle (1989) who derived CCs for one-dimensional kernel M-estimators. Claeskens and Van Keilegom (2003) proposed uniform confidence bands and a bootstrap procedure for regression curves and their derivatives.

In recent years, the growth of the literature body shows no sign of decelerating. In the same spirit of Härdle (1989), Härdle and Song (2010) and Guo and Härdle (2012) constructed uniform confidence bands for local constant quantile and expectile curves. Fan and Liu (2013) proposed an integrated approach for building simultaneous confidence band that covers semiparametric models. Giné and Nickl (2010) investigated adaptive density estimation based on linear wavelet and kernel density estimators and Lounici and Nickl (2011) extended the framework of Bissantz et al. (2007) to adaptive deconvolution density estimation. Bootstrap procedures are proposed as a remedy for the poor coverage performance of asymptotic confidence corridors. For example, the bootstrap for the density estimator is proposed in Hall (1991) and Mojirsheibani (2012), and for local constant quantile estimators in Song et al. (2012).

However, only recently progress has been achieved in the construction of confidence bands for regression estimates with a multivariate predictor. Hall and Horowitz (2013) derived an expansion for the bootstrap bias and established a somewhat different way to construct confidence bands without the use of extreme value theory. Their bands are uniform with respect to a fixed but unspecified portion (smaller than one) of points in a possibly multidimensional set in contrast to the classical approach where uniformity is achieved on the complete set considered. Proksch et al. (2014) proposed multivariate confidence bands for convolution type inverse regression models with fixed design.

To the best of our knowledge results of the classical Smirnov-Bickel-Rosenblatt type are not available for multivariate GQR or even mean regression with random design.

In this work we go beyond the earlier studies in three aspects. First, we extend the applicability of the CC to d-dimensional covariates with $d > 1$. Second, we present a more general approach covering not only quantile or mean curves but also GQR curves that are defined via a minimum contrast principle. Third, we propose a bootstrap procedure and we show numerically its improvement in the coverage accuracy as compared to the asymptotic approach.

Our asymptotic results, which describe the maximal absolute deviation of generalized quantile estimators, can not only be used to derive a goodness-of-fit test in quantile and expectile regression, but they are also applicable in testing the quantile treatment effect and stochastic dominance. We apply the new method to test the quantile treatment effect of the National Supported Work Demonstration program, which is a randomized employment enhancement program launched in the 1970s. The data associated with the participants of the program have been widely applied for treatment effect research since the pioneering study of LaLonde (1986). More recently, Delgado and Escanciano (2013) found that the program is beneficial for individuals of over 21 years of age. In our study, we find that the treatment tends to do better at raising the upper bounds of the earnings growth than raising the lower bounds. In other words, the program tends to increase the potential for high earnings growth but does not reduce the risk of negative earnings growth. The finding is particularly evident for those individuals who are older and spent more years at school. We should note that the tests based on the unconditional distribution cannot unveil the heterogeneity in the earnings growth quantiles in treatment effects.

The remaining part of this paper is organized as follows. In Section 2 we present our model, describe the estimators and state our asymptotic results. Section 3 is devoted to the bootstrap and we discuss its theoretical and practical aspects. The finite sample properties of both methods are investigated by means of a simulation study in Section 4 and the application of the new method is illustrated in a data example in Section 5. The assumptions for our asymptotic theory are listed and discussed after the references. All detailed proofs are available in the supplement material.

2. Asymptotic confidence corridors

In Section 2.1 we present the prerequisites such as the precise definition of the model and a suitable estimate. The result on constructing confidence corridors (CCs) based on the distribution of the maximal absolute deviation are given in Section 2.2. In Section 2.3 we describe how to estimate the scaling factors, which appear in the limit theorems, using residual based estimators. Section 3.1 introduce a new bootstrap method for constructing CCs, while Section 3.2 is devoted to specific issues related to bootstrap CCs for quantile regression. Assumptions are listed and discussed after the references.

2.1. Prerequisites

Let $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ be a sequence of independent identically distributed random vectors in $\mathbb{R}^{d+1}$ and consider the nonparametric regression model

$$Y_i = \theta(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{1}$$

where θ is an aspect of Y conditional on $\boldsymbol{X}$ such as the τ-quantile, the τ-expectile or the mean regression curve. The function $\theta(\boldsymbol{x})$ can be estimated by:

$$\hat{\theta}(\boldsymbol{x}) = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\rho(Y_i - \theta), \tag{2}$$

where $K_h(\boldsymbol{u}) = h^{-d} K(\boldsymbol{u}/h)$ for some kernel function $K : \mathbb{R}^d \to \mathbb{R}$, and a loss-function $\rho_\tau : \mathbb{R} \to \mathbb{R}$. In this paper we are concerned with the construction of uniform confidence corridors for quantile as well as expectile regression curves when the predictor is multivariate, that is, we focus on the loss functions

$$\rho_\tau(u) = \big|\mathbf{1}(u < 0) - \tau\big||u|^k,$$

for $k = 1$ and 2 associated with quantile and expectile regression. We derive the asymptotic distribution of the properly scaled maximal deviation $\sup_{\boldsymbol{x} \in \mathcal{D}} |\hat{\theta}_n(\boldsymbol{x}) - \theta(\boldsymbol{x})|$ for both cases, where $\mathcal{D} \subset \mathbb{R}^d$ is a compact subset. We use strong approximations of the empirical process, concentration inequalities for general Gaussian random fields and results from extreme value theory. To be precise, we show that

$$\mathrm{P}\left[(2\delta \log n)^{1/2}\Big\{\sup_{\boldsymbol{x}\in\mathcal{D}} \big|r_n(\boldsymbol{x})\big[\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\big]\big|/\|K\|_2 - d_n\Big\} < a\right] \to \exp\big\{-2\exp(-a)\big\}, \quad (3)$$

as $n \to \infty$, where $r(\boldsymbol{x})$ is a scaling factor which depends on $\boldsymbol{x}$, n and the loss function under consideration.

2.2. Asymptotic results

In this section we present our main theoretical results on the distribution of the uniform maximal deviation of the quantile and expectile estimator. The proofs of the theorems at their full lengths are deferred the appendix. Here we only give a brief sketch of proof of Theorem 2.1 which is the limit theorem for the case of quantile regression.

THEOREM 2.1. *Let $\hat{\theta}_n(\boldsymbol{x})$ and $\theta_0(\boldsymbol{x})$ be the local constant quantile estimator and the true quantile function, respectively and suppose that assumptions (A1)-(A6) in Section 5 hold. Let further vol$(\mathcal{D}) = 1$ and*

$$d_n = (2d\kappa \log n)^{1/2} + \big\{2d\kappa(\log n)\big\}^{-1/2}\left[\frac{1}{2}(d-1)\log\log n^{\kappa} + \log\big\{(2\pi)^{-1/2}H_2(2d)^{(d-1)/2}\big\}\right],$$

where $H_2 = \big(2\pi\|K\|_2^2\big)^{-d/2}\det(\Sigma)^{1/2}$, $\Sigma = \big(\Sigma_{ij}\big)_{i,j=1}^d = \Big(\int \frac{\partial K(\boldsymbol{u})}{\partial u_i}\frac{\partial K(\boldsymbol{u})}{\partial u_j}d\boldsymbol{u}\Big)_{i,j=1}^d$,

$$r(\boldsymbol{x}) = \sqrt{\frac{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}{\tau(1-\tau)}}\, f_{Y|\boldsymbol{X}}\big\{\theta_0(\boldsymbol{x})|\boldsymbol{x}\big\},$$

Then the limit theorem (3) holds.

Sketch of proof. A major technical difficulty is imposed by the fact that the loss-function ρ_τ is not smooth which means that standard arguments such as those based on Taylor's theorem do not apply. As a consequence the use of a different, extended methodology becomes necessary. In this context Kong et al. (2010) derived a uniform Bahadur representation for an M-regression function in a multivariate setting (see appendix). It holds uniformly for $x \in \mathcal{D}$, where $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$:

$$\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) = \frac{1}{nS_{n,0,0}(\boldsymbol{x})} \sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi_\tau\{Y_i - \theta_0(\boldsymbol{x})\} + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}, \quad a.s. \quad (4)$$

Here $S_{n,0,0}(\boldsymbol{x}) = \int K(\boldsymbol{u})g(\boldsymbol{x} + h\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{x} + h\boldsymbol{u})d\boldsymbol{u}$, $\psi_\tau(u) = \mathbf{1}(u < 0) - \tau$ is the piecewise derivative of the loss-function ρ_τ and

$$g(\boldsymbol{x}) = \frac{\partial}{\partial t}\mathsf{E}[\psi_\tau(Y - t)|\boldsymbol{X} = \boldsymbol{x}]\Big|_{t=\theta_0(\boldsymbol{x})}.$$

Notice that the error term of the Bahadur expansion does not depend on the design $\boldsymbol{X}$ and it converges to 0 with rate $\left(\log n/nh^d\right)^{\frac{3}{4}}$ which is much faster than the convergence rate $(nh^d)^{-\frac{1}{2}}$ of the stochastic term.

Rearranging (4), we obtain

$$S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\} = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi_\tau\{Y_i - \theta_0(\boldsymbol{x})\} + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}. \quad (5)$$

Now we express the leading term on the right hand side of (5) by means of the centered empirical process

$$Z_n(y, \boldsymbol{u}) = n^{1/2}\{F_n(y, \boldsymbol{u}) - F(y, \boldsymbol{u})\}, \quad (6)$$

where $F_n(y, \boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(Y_i \le y, X_{i1} \le x_1, ..., X_{id} \le x_d)$. This yields, by Fubini's theorem,

$$S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\} - b(\boldsymbol{x}) = n^{-1/2}\int\int K_h(\boldsymbol{x} - \boldsymbol{u})\psi_\tau\{y - \theta_0(\boldsymbol{x})\}dZ_n(y, \boldsymbol{u}) + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}, \quad (7)$$

where

$$b(\boldsymbol{x}) = -\mathsf{E}_{\boldsymbol{x}}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi\{Y_i - \theta_0(\boldsymbol{x})\}\right]$$

denotes the bias which is of order $\mathcal{O}(h^s)$ by Assumption (A3) in the Appendix. The variance of the first term of the right hand side of (7) can be estimated via a change of variables and Assumption (A5), which gives

$$
\begin{aligned}
&(nh^d)^{-2} n \mathsf{E}\big[K^2\{(\boldsymbol{x}-\boldsymbol{X}_i)/h\}\psi^2\{Y_i-\theta_0(\boldsymbol{x})\}\big] \\
&= (nh^d)^{-2} nh^d \int\int K^2(\boldsymbol{v})\psi^2\{y-\theta_0(\boldsymbol{x})\} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}-h\boldsymbol{v}) f_{\boldsymbol{X}}(\boldsymbol{x}-h\boldsymbol{v}) dy d\boldsymbol{v} \\
&= (nh^d)^{-1} \int\int K^2(\boldsymbol{v})\psi^2\{y-\theta_0(\boldsymbol{x})\} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) dy d\boldsymbol{v} + \mathcal{O}\big((nh^{d-1})^{-1}\big) \\
&= (nh^d)^{-1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})\|K\|_2^2 + \mathcal{O}\{(nh^d)^{-1}h\},
\end{aligned}
$$

where $\sigma^2(\boldsymbol{x}) = \mathsf{E}[\psi^2\{Y-\theta_0(\boldsymbol{x})\}|\boldsymbol{X}=\boldsymbol{x}]$. The standardized version of (5) can therefore be approximated by

$$
\begin{aligned}
&\frac{\sqrt{nh^d}}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})\|K\|_2} S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\} \\
&= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})\|K\|_2} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi\{Y_i-\theta_0(\boldsymbol{x})\} dZ_n(y,\boldsymbol{u}) + \mathcal{O}\big(\sqrt{nh^d}h^s\big) + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}.
\end{aligned}
\tag{8}
$$

The dominating term is defined by

$$
Y_n(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi\{y-\theta_0(\boldsymbol{x})\} dZ_n(y,\boldsymbol{u}). \tag{9}
$$

Involving strong Gaussian approximation and Bernstein-type concentration inequalities, this process can be approximated by a stationary Gaussian field:

$$
Y_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW(\boldsymbol{u}), \tag{10}
$$

where W denotes a Brownian sheet. The supremum of this process is asymptotically Gumbel distributed, which follows, e.g., by Theorem 2 of Rosenblatt (1976). Since the kernel is symmetric and of order s, we can estimate the term

$$
S_{n,0,0} = f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s).
$$

if (A5) holds. On the other hand, $\sigma^2(\boldsymbol{x}) = \tau(1-\tau)$ in quantile regression. Therefore, the statements of the theorem hold.

$\square$

Corollary 2.2 (CC for multivariate quantile regression)**.** Under the assumptions of Theorem 2.1, an approximate $(1-\alpha) \times 100\%$ confidence corridor is given by

$$\hat{\theta}_n(\boldsymbol{t}) \pm (nh^d)^{-1/2}\{\tau(1-\tau)\|K\|_2/\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})\}^{1/2}\hat{f}_{\varepsilon|\boldsymbol{X}}\{0|\boldsymbol{t}\}^{-1}\Big\{d_n + c(\alpha)(2\kappa d\log n)^{-1/2}\Big\},$$

where $\alpha \in (0,1)$ and $c(\alpha) = \log 2 - \log\big|\log(1-\alpha)\big|$ and $\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})$, $\hat{f}_{\varepsilon|\boldsymbol{X}}\{0|\boldsymbol{t}\}$ are consistent estimates for $f_{\boldsymbol{X}}(\boldsymbol{t})$, $f_{\varepsilon|\boldsymbol{X}}\{0|\boldsymbol{t}\}$ with convergence rate faster than $\mathcal{O}_p\big((\log n)^{-1/2}\big)$.

The expectile confidence corridor can be constructed in an analogous manner as the quantile confidence corridor. The two cases differ in the form and hence the properties of the loss function. Therefore we find for expectile regression:

$$S_{n,0,0}(\boldsymbol{x}) = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big)(2\tau-1) - \tau\big]f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s).$$

Through similar approximation steps as the quantile regression, we derive the following theorem.

THEOREM 2.3. *Let $\hat{\theta}_n(\boldsymbol{x})$ be the the local constant expectile estimator and $\theta_0(\boldsymbol{x})$ the true expectile function. If Assumptions (A1), (A3)-(A6) and (EA2) of Section 5 hold with a constant b_1 satisfying*

$$n^{-1/6}h^{-d/2-3d/(b_1-2)} = \mathcal{O}(n^{-\nu}), \quad \nu > 0.$$

Then the limit theorem (3) holds with a scaling factor

$$r(\boldsymbol{x}) = \sqrt{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma^{-1}(\boldsymbol{x})\left\{2\big[\tau - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})(2\tau-1)\big]\right\},$$

the same constants H_2 and d_n as defined in Theorem 2.1, where $\sigma^2(\boldsymbol{x}) = \mathsf{E}[\psi_\tau^2(Y-\theta_0(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}]$ and $\psi_\tau(u) = 2(\mathbf{1}(u \le 0) - \tau)|u|$ is the derivative of the expectile loss-function $\rho_\tau(u) = \big|\tau - \mathbf{1}(u<0)\big||u|^2$.

The proof of this result is deferred to the appendix. The next corollary shows the CC for expectiles.

Corollary 2.4 (CC for multivariate expectile regression)**.** Under the same assumptions of Theorem 2.3, an approximate $(1-\alpha)\times 100\%$ confidence corridor is given by

$$\hat{\theta}_n(\boldsymbol{t}) \pm (nh^d)^{-1/2}\big\{\hat{\sigma}^2(\boldsymbol{t})\|K\|_2/\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})\big\}^{1/2}\Big\{-2\big[\hat{F}_{\varepsilon|\boldsymbol{X}}\{0|\boldsymbol{t}\}(2\tau-1)-\tau\big]\Big\}^{-1}\Big\{d_n + c(\alpha)(2\kappa d\log n)^{-1/2}\Big\},$$

where $\alpha \in (0,1)$ $c(\alpha) = \log 2 - \log\big|\log(1-\alpha)\big|$ and $\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})$, $\hat{\sigma}^2(\boldsymbol{t})$ and $\hat{F}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ are consistent estimates for $f_{\boldsymbol{X}}(\boldsymbol{t})$, $\sigma^2(\boldsymbol{t})$ and $F_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ with convergence rate faster than $\mathcal{O}_p\big((\log n)^{-1/2}\big)$.

A further immediate consequence of Theorem 2.3 is a similar limit theorem in the context of local least squares estimation of the regression curve in classical mean regression.

Corollary 2.5 (CC for multivariate mean regression)**.** Consider the loss function $\rho(u) = u^2$ corresponding to $\psi(u) = 2u$. Under the assumptions of Theorem 2.3, with the same constants H_2 and d_n, (3) holds for the local constant estimator $\hat{\theta}$ and the regression function $\theta(x) = E[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ with scaling factor $r(\boldsymbol{x}) = \sqrt{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma^{-1}(\boldsymbol{x})$ and $\sigma^2(\boldsymbol{x}) =Var[Y \mid \boldsymbol{X} = \boldsymbol{x}]$.

For the appropriate bandwidth choice, it is enough to take $h = \mathcal{O}(n^{-1/(2s+d)-\delta})$, given $s > d$ and $\delta > 0$ to make our asymptotic theories hold, where s is the order of Hölder continuity of the function θ_0. In the simulation study we use the rule-of-thumb bandwidth with adjustments proposed by Yu and Jones (1998) for nonparametric quantile regression, and for expectile regression we use the rule-of-thumb bandwidth for the conditional distribution smoother of Y given $\boldsymbol{X}$, chosen with the `np` package in `R`. In the application, we use the cross-validated bandwidth for conditional distribution smoother of Y given $\boldsymbol{X}$, chosen with the `np` package in `R`. This package is based on the paper of Li et al. (2013).

2.3. Estimating the scaling factors

The performance of the confidence bands is greatly influenced by the scaling factors $\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$, $F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ and $\hat{\sigma}(\boldsymbol{x})^2$. The purpose of this subsection is thus to propose a way to estimate these factors and investigate their asymptotic properties.

Since we consider the additive error model (1), the conditional distribution function $F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ and the conditional density $f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ can be replaced by $F_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$, respectively, where $F_{\varepsilon|\boldsymbol{X}}$ and $f_{\varepsilon|\boldsymbol{X}}$ are the conditional distribution and density functions of ε. Similarly, we have

$$\sigma^2(\boldsymbol{x}) = \mathsf{E}\left[\psi_\tau\big(Y - \theta_0(\boldsymbol{x})\big)^2 \big| \boldsymbol{X} = \boldsymbol{x}\right] = \mathsf{E}\left[\psi_\tau(\varepsilon)^2 \big| \boldsymbol{X} = \boldsymbol{x}\right]$$

where ε may depend on $\boldsymbol{X}$ due to heterogeneity. It should be noted that the kernel estimators for $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and $f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ are asymptotically equivalent, but show different finite sample behavior. We explore this issue further in the following section.

Introducing the residuals $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(\boldsymbol{X}_i)$ we propose to estimate $F_{\varepsilon|\boldsymbol{X}}$, $f_{\varepsilon|\boldsymbol{X}}$ and $\sigma^2(\boldsymbol{x})$ by

$$\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} G\left(\frac{v - \hat{\varepsilon}_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) / \hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{11}$$

$$\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} g_{h_0}\left(v - \hat{\varepsilon}_i\right) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) / \hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{12}$$

$$\hat{\sigma}^2(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \psi^2(\hat{\varepsilon}_i) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) / \hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{13}$$

where $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)$, G is a continuously differentiable cumulative distribution function and g is its derivative. The same bandwidth $\bar{h}$ is applied to the three estimators, but the choice of $\bar{h}$ will make the convergence rate of (13) sub-optimal. More details on the choice of $\bar{h}$ will be given later. Nevertheless, the rate of convergence of (13) is of polynomial order in n. The theory developed in this subsection can be generalized to the case of different bandwidth for different direction without much difficulty.

The estimators (11) and (12) belong to the family of residual-based estimators. The consistency of residual-based density estimators for errors in a regression model are explored in the literature in various settings. It is possible to obtain an expression for the residual based kernel density estimator as the sum of the estimator with the true residuals, the partial sum of the true residuals and a term for the bias of the nonparametrically estimated function, as shown in Muhsal and Neumeyer (2010), among others. The residual based conditional kernel density case is less considered in the literature. Kiwitt and Neumeyer (2012) consider the residual based kernel estimator for conditional distribution function conditioning on a one-dimensional variable.

Below we give consistency results for the estimators defined in (11), (12) and (13). The proof can be found in the appendix.

Lemma 2.6. Under conditions (A1), (A3)-(A5), (B1)-(B3) in Section 5, we have

1) $\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})-F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right|=\mathcal{O}_p\big(a_n\big)$,

2) $\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})-f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right|=\mathcal{O}_p\big(a_n\big)$,

3) $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{\sigma}^2(\boldsymbol{x})-\sigma^2(\boldsymbol{x})\right|=\mathcal{O}_p\big(b_n\big)$,

where $a_n=\mathcal{O}\big\{h_0^{s'}+h^s+\bar{h}^{s'}+(n\bar{h}^d)^{-1/2}\log n+(nh^d)^{-1/2}\log n\big\}=\mathcal{O}(n^{-\lambda})$, and $b_n=\mathcal{O}\big\{h^s+\bar{h}^{s'}+(n\bar{h}^d)^{-1/2}\log n+(nh^d)^{-1/2}\log n\big\}=\mathcal{O}(n^{-\lambda_1})$ for some constants $\lambda,\lambda_1>0$.

The factor of $\log n$ shown in the convergence rate is the price which we pay for the supnorm deviation. Since these estimators uniformly converge in a polynomial rate in n, the asymptotic distributions in Theorem 2.1 and 2.3 do not change if we plug these estimators into the formulae.

The choice of h_0 and $\bar{h}$ should minimize the convergence rate of the residual based estimators. Hence, observing that the terms related to h_0 and $\bar{h}$ are similar to those in usual $(d+1)$-dimensional density estimators, it is reasonable to choose $h_0\sim\bar{h}\sim n^{-1/(5+d)}$, given that L, g are second order kernels. We choose the rule-of-thumb bandwidths for conditional densities with the `R` package `np` in our simulation and application studies.

3. Bootstrap confidence corridors

3.1. Asymptotic theory

In the case of the suitably normed maximum of independent standard normal variables, it is shown in Hall (1979) that the speed of convergence in limit theorems of the form (3) is of order $1/\log n$, that is, the coverage error of the asymptotic CC decays only logarithmically.

This leads to unsatisfactory finite sample performance of the asymptotic methods, especially for small sample sizes. However, Hall (1991) suggests that the use of a bootstrap method, based on a proper way of resampling, can increase the speed of shrinking of coverage error to a polynomial rate of n. In this section we therefore propose a specific bootstrap technique and construct a confidence corridor for the objects to analyse.

Given the residuals $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(\boldsymbol{X}_i)$, the bootstrap observations $(\boldsymbol{X}_i^*, \varepsilon_i^*)$ are sampled from

$$\hat{f}_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{L}_{h_0}\left(\hat{\varepsilon}_i - v\right) L_h(\boldsymbol{x} - \boldsymbol{X}_i), \tag{14}$$

where L and $\tilde{L}$ are kernel functions with bandwidths h and h_0. In particular, in our simulation study, we choose L to be a product Gaussian kernel. In the following discussion P^* and E^* stand for the probability and expectation conditional on the data $(\boldsymbol{X}_i, Y_i)$, $i = 1, ..., n$.

We introduce the notation

$$A_n^*(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i^*)\psi_\tau(\varepsilon_i^*),$$

and define the so-called "one-step estimator" $\theta^*(\boldsymbol{x})$ from the bootstrap sample by

$$\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x}) = \hat{S}_{n,0,0}^{-1}(\boldsymbol{x})\left\{A_n^*(\boldsymbol{x}) - \mathsf{E}^*[A_n^*(\boldsymbol{x})]\right\}, \tag{15}$$

where

$$\hat{S}_{n,0,0}(\boldsymbol{x}) = \begin{cases} \hat{f}_{\varepsilon|\boldsymbol{X}}\big(0|\boldsymbol{x}\big)\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), & \text{quantile case;} \\ 2\big\{\tau - \hat{F}_{\varepsilon|\boldsymbol{X}}\big(0|\boldsymbol{x}\big)(2\tau - 1)\big\}\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), & \text{expectile case.} \end{cases} \tag{16}$$

note that $\mathsf{E}^*[\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x})] = 0$, so $\hat{\theta}^*(\boldsymbol{x})$ is unbiased for $\hat{\theta}_n(\boldsymbol{x})$ under E^*. As a remark, we note that undersmoothing is applied in our procedure for two reasons: first, the theory we developed so far is based on undersmoothing; secondly, it is suggested in Hall (1992) that undersmoothing is more effective than oversmoothing given that the goal is to achieve coverage accuracy.

Note that the bootstrap estimate (15) is motivated by the smoothed bootstrap procedure proposed in Claeskens and Van Keilegom (2003). In constrast to these authors we make use of the leading term of the Bahadur representation. Mammen et al. (2013) also use the leading term of a Bahadur representation proposed in Guerre and Sabbah (2012) to construct bootstrap samples. Song et al. (2012) propose a bootstrap for quantile regression based on oversmoothing, which has the drawback that it requires iterative estimation, and oversmoothing is in general less effective in terms of coverage accuracy.

For the following discussion define

$$Y_n^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sigma_*(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dZ_n^*(v,\boldsymbol{u}) \tag{17}$$

as the bootstrap analogue of the process (9), where

$$Z_n^*(y,\boldsymbol{u}) = n^{1/2}\left\{F_n^*(v,\boldsymbol{u}) - \hat{F}(v,\boldsymbol{u})\right\}, \quad \sigma_*(\boldsymbol{x}) = \sqrt{\mathsf{E}^*\left[\psi_\tau(\varepsilon_i^*)^2|\boldsymbol{x}\right]} \tag{18}$$

and

$$F_n^*(v,\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\varepsilon_i^* \le v, X_1^* \le u_1, ..., X_d^* \le u_d\right\}.$$

The process Y_n^* serves as an approximation of a standardized version of $\hat{\theta}_n^* - \hat{\theta}_n$, and similar to the previous sections the process Y_n^* is approximated by a stationary Gaussian field $Y_{n,5}^*$ under P^* with probability one, that is,

$$Y_{5,n}^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}}\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)dW^*(\boldsymbol{u}).$$

Finally, $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|Y_{5,n}^*(\boldsymbol{x})\right|$ is asymptotically Gumbel distributed conditional on samples.

THEOREM 3.1. *Suppose that assumptions (A1)-(A6), (C1) in Section 5 hold, and vol($\mathcal{D}$) = 1, let*

$$r^*(\boldsymbol{x}) = \sqrt{\frac{nh^d}{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma_*^2(\boldsymbol{x})}}\hat{S}_{n,0,0}(\boldsymbol{x}),$$

where $\hat{S}_{n,0,0}(\boldsymbol{x})$ is defined in (16) and $\sigma_*^2(\boldsymbol{x})$ is defined in (18). Then

$$P^*\left\{(2d\kappa\log n)^{1/2}\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left[r^*(\boldsymbol{x})|\hat{\theta}^*(\boldsymbol{x})-\hat{\theta}_n(\boldsymbol{x})|\right]/\|K\|_2-d_n\right)<a\right\}\to\exp\left\{-2\exp(-a)\right\},\quad a.s. \tag{19}$$

as $n\to\infty$ for the local constant quantile regression estimate. If (A1)-(A6) and (EC1) hold with a constant $b\geq 4$ satisfying

$$n^{-\frac{1}{6}+\frac{4}{b^2}-\frac{1}{b}}h^{-\frac{d}{2}-\frac{6d}{b}}=\mathcal{O}(n^{-\nu}),\quad \nu>0,$$

then (19) also holds for expectile regression with corresponding $\sigma_*^2(\boldsymbol{x})$.

The proof can be found in the appendix. The following lemma suggests that we can replace $\sigma_*(\boldsymbol{x})$ in the limiting theorem by $\hat{\sigma}(\boldsymbol{x})$.

Lemma 3.2. If assumptions (B1)-(B3), and (EC1) in Section 5 are satisfied with $b>2(2s'+d+1)/(2s'+3)$, then

$$\|\sigma_*^2(\boldsymbol{x})-\hat{\sigma}^2(\boldsymbol{x})\|=\scriptstyle{\mathcal{O}}_p^*\big((\log n)^{-1/2}\big),\quad a.s.$$

The following corollary is a consequence of Theorem 3.1.

Corollary 3.3. Under the same conditions as stated in Theorem 3.1, the (asymptotic) bootstrap confidence set of level $1-\alpha$ is given by

$$\left\{\theta:\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{\hat{S}_{n,0,0}(\boldsymbol{x})}{\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\hat{\sigma}^2(\boldsymbol{x})}}\left[\hat{\theta}_n(\boldsymbol{x})-\theta(\boldsymbol{x})\right]\right|\leq\xi_\alpha^*\right\}, \tag{20}$$

where ξ_α^* satisfies

$$\lim_{n\to\infty}P^*\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{\hat{S}_{n,0,0}(\boldsymbol{x})}{\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\hat{\sigma}^2(\boldsymbol{x})}}\left[\hat{\theta}^*(\boldsymbol{x})-\hat{\theta}_n(\boldsymbol{x})\right]\right|\leq\xi_\alpha^*\right)=1-\alpha,\quad a.s. \tag{21}$$

where $\hat{S}_{n,0,0}$ is defined in (16).

Note that it does not create much difference to standardize the $\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})$ in (19) with $\hat{f}_{\boldsymbol{X}}$ and $\hat{\sigma}^2(\boldsymbol{x})$ constructed from original samples or $\hat{f}_{\boldsymbol{X}}$ and $\hat{\sigma}^2(\boldsymbol{x})$ from the bootstrap samples. The simulation results of Claeskens and Van Keilegom (2003) show that the two ways of standardization give similar coverage probabilities for confidence corridors of kernel ML estimators.

3.2. Implementation

In this section, we discuss issues related to the implementation of the bootstrap for quantile regression.

The one-step estimator for quantile regression defined in (15) depends sensitively on the estimator of $\hat{S}_{n,0,0}(\boldsymbol{x})$. Unlike the expectile case, the function $\psi(\cdot)$ in quantile case is bounded, and as the result the bootstrapped density based on (20) is very easily influenced by the factor $\hat{S}_{n,0,0}(\boldsymbol{x})$; in particular, $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$. As pointed out by Feng et al. (2011), the residual of quantile regression tends to be less dispersed than the model error; thus $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ tends to over-estimate the true $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ for each $\boldsymbol{x}$.

The way of getting around this problem is based on the following observation: An additive error model implies the equality $f_{Y|\boldsymbol{X}}\{v + \theta_0(\boldsymbol{x})|\boldsymbol{x}\} = f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ but this property does not hold for the kernel estimators

$$\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}(\hat{\varepsilon}_i)\, L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) \tag{22}$$

$$\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_1}\left(Y_i - \hat{\theta}_n(\boldsymbol{x})\right) L_{\tilde{h}}(\boldsymbol{x} - \boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{23}$$

of the conditional density functions. In general$\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x}) \neq \hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$ in $\boldsymbol{x}$ although both estimates are asymptotically equivalent. In applications the two estimators can differ substantially due to the bandwidth selection because for data-driven bandwidths we usually have $h_0 \neq h_1$. For example, if acommon method for bandwidth selection such as a rule-of-thumb is used, h_1 will tend to be larger than h_0 since the sample variance of Y_i tends to be larger than that of $\hat{\varepsilon}_i$. Given that the same kernels are applied, it happens often that $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}) > f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$, even if $\hat{\theta}_n(\boldsymbol{x})$ is usually very close to $\theta_0(\boldsymbol{x})$. To correct such abnormality, we are motivated to set $h_1 = h_0$ which is the rule-of-thumb bandwidth of $\hat{f}_{\varepsilon|\boldsymbol{x}}(v|\boldsymbol{x})$ in (23). As the result, it leads to a more rough estimate for $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$.

In order to exploit the roughness of $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$ while making the CC as narrow as possible, we develop a trick depending on

$$\frac{\hat{f}_{Y|\boldsymbol{X}}\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\}}{\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})} = \frac{h_0}{h_1}\frac{\sum_{i=1}^n g_{h_1}\left(\{Y_i - \hat{\theta}_n(\boldsymbol{x})\}/h_1\right) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)}{\sum_{i=1}^n g_{h_0}\left(\hat{\varepsilon}_i/h_0\right) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)}. \tag{24}$$

As $n \to \infty$, (24) converges to 1. If we impose $h_0 = h_1$, as the multiple h_0/h_1 vanishes, (24) captures the deviation of the two estimators without the difference of the bandwidth in the way. In particular, the bandwidth $h_0 = h_1$ is selected with the rule-of-thumb bandwidth for $\hat{f}_{\varepsilon|\boldsymbol{X}}(y|\boldsymbol{x})$. This makes $\hat{f}_{\varepsilon|\boldsymbol{X}}(y|\boldsymbol{x})$ larger and thus leads to a narrower CC, as will be more clear below.

We propose the alternative bootstrap confidence corridor for quantile estimator:

$$\left\{\theta : \sup_{\boldsymbol{x}\in\mathcal{D}} \left|\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\hat{f}_{Y|\boldsymbol{X}}\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\}\left[\hat{\theta}_n(\boldsymbol{x}) - \theta(\boldsymbol{x})\right]\right| \le \xi_\alpha^\dagger\right\},$$

where $\xi_\alpha^\dagger$ satisfies

$$\mathrm{P}^*\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})^{-1/2}\frac{\hat{f}_{Y|\boldsymbol{X}}\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\}}{\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})}\left[A_n^*(\boldsymbol{x}) - \mathsf{E}^*A_n^*(\boldsymbol{x})\right]\right| \le \xi_\alpha^\dagger\right) = 1 - \alpha. \tag{25}$$

Note that the probability on the left-hand side of (25) can again be approximated by a Gumbel distribution function asymptotically, which follows by Theorem 3.1.

4. A simulation study

In this section we investigate the methods described in the previous sections by means of a simulation study. We construct confidence corridors for quantiles and expectiles for different levels τ and use the quartic (product) kernel. For the confidence based on asymptotic distribution theory, we use the rule of thumb bandwidth chosen from the R package `np`, and then rescale it as described in Yu and Jones (1998), finally multiply it by $n^{-0.05}$ for undersmoothing. The sample sizes are given by $n = 100, 300$ and 500, so the undersmoothing multiples are 0.794, 0.752 and 0.733 respectively. In the quantile regression bootstrap CC, the bandwidth h_1 used for estimating $\hat{f}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ is chosen to be the rule-of-thumb bandwidth of $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and multiplied by a multiple 1.5. This would give slightly wider CCs.

		Homogeneous			**Heterogeneous**		
Method	n	$\boldsymbol{\tau = 0.5}$	$\boldsymbol{\tau = 0.2}$	$\boldsymbol{\tau = 0.8}$	$\boldsymbol{\tau = 0.5}$	$\boldsymbol{\tau = 0.2}$	$\boldsymbol{\tau = 0.8}$
				$\boldsymbol{\sigma_0 = 0.2}$			
	100	.000(0.366)	.109(0.720)	.104(0.718)	.000(0.403)	.120(0.739)	.122(0.744)
	300	.000(0.304)	.130(0.518)	.133(0.519)	.002(0.349)	.136(0.535)	.153(0.537)
	500	.000(0.262)	.117(0.437)	.142(0.437)	.008(0.296)	.156(0.450)	.138(0.450)
				$\boldsymbol{\sigma_0 = 0.5}$			
	100	.070(0.890)	.269(1.155)	.281(1.155)	.078(0.932)	.300(1.193)	.302(1.192)
Asympt.	300	.276(0.735)	.369(0.837)	.361(0.835)	.325(0.782)	.380(0.876)	.394(0.877)
	500	.364(0.636)	.392(0.711)	.412(0.712)	.381(0.669)	.418(0.743)	.417(0.742)
				$\boldsymbol{\sigma_0 = 0.7}$			
	100	.160(1.260)	.381(1.522)	.373(1.519)	.155(1.295)	.364(1.561)	.373(1.566)
	300	.438(1.026)	.450(1.109)	.448(1.110)	.481(1.073)	.457(1.155)	.472(1.152)
	500	.533(0.888)	.470(0.950)	.480(0.949)	.564(0.924)	.490(0.984)	.502(0.986)
				$\boldsymbol{\sigma_0 = 0.2}$			
	100	.325(0.676)	.784(0.954)	.783(0.954)	.409(0.717)	.779(0.983)	.778(0.985)
	300	.442(0.457)	.896(0.609)	.894(0.610)	.580(0.504)	.929(0.650)	.922(0.649)
	500	.743(0.411)	.922(0.502)	.921(0.502)	.839(0.451)	.950(0.535)	.952(0.536)
				$\boldsymbol{\sigma_0 = 0.5}$			
	100	.929(1.341)	.804(1.591)	.818(1.589)	.938(1.387)	.799(1.645)	.773(1.640)
Bootst.	300	.950(0.920)	.918(1.093)	.923(1.091)	.958(0.973)	.919(1.155)	.923(1.153)
	500	.988(0.861)	.968(0.943)	.962(0.942)	.990(0.902)	.962(0.986)	.969(0.987)
				$\boldsymbol{\sigma_0 = 0.7}$			
	100	.976(1.811)	.817(2.112)	.808(2.116)	.981(1.866)	.826(2.178)	.809(2.176)
	300	.986(1.253)	.919(1.478)	.934(1.474)	.983(1.308)	.930(1.537)	.920(1.535)
	500	.996(1.181)	.973(1.280)	.968(1.278)	.997(1.225)	.969(1.325)	.962(1.325)

Table 1: *Nonparametric quantile model coverage probabilities. The nominal coverage is* 95%. *The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic quantile regression CC and bootstrap method corresponds to quantile regression bootstrap CC.*

The data are generated from the normal regression model

$$Y_i = f(X_{1,i}, X_{2,i}) + \sigma(X_{1,i}, X_{2,i})\varepsilon_i, \quad i = 1, \dots, n$$

where the independent variables (X_1, X_2) follow a joint uniform distribution taking values on $[0,1]^2$, $\mathrm{Cov}(X_1, X_2) = 0.2876$, $f(X_1, X_2) = \sin(2\pi X_1) + X_2$, and ε_i are independent standard Gaussian random variables. For both quantile and expectile, we look at three quantiles of

Method	n	Homogeneous $\tau = 0.5$	Homogeneous $\tau = 0.2$	Homogeneous $\tau = 0.8$	Heterogeneous $\tau = 0.5$	Heterogeneous $\tau = 0.2$	Heterogeneous $\tau = 0.8$
				$\sigma_0 = 0.2$			
	100	.000(0.428)	.000(0.333)	.000(0.333)	.000(0.463)	.000(0.362)	.000(0.361)
	300	.049(0.341)	.000(0.273)	.000(0.273)	.079(0.389)	.001(0.316)	.002(0.316)
	500	.168(0.297)	.000(0.243)	.000(0.243)	.238(0.336)	.003(0.278)	.002(0.278)
				$\sigma_0 = 0.5$			
	100	.007(0.953)	.000(0.776)	.000(0.781)	.007(0.997)	.000(0.818)	.000(0.818)
Asympt.	300	.341(0.814)	.019(0.708)	.017(0.709)	.355(0.862)	.017(0.755)	.018(0.754)
	500	.647(0.721)	.067(0.645)	.065(0.647)	.654(0.759)	.061(0.684)	.068(0.684)
				$\sigma_0 = 0.7$			
	100	.012(1.324)	.000(1.107)	.000(1.107)	.010(1.367)	.000(1.145)	.000(1.145)
	300	.445(1.134)	.021(1.013)	.013(1.016)	.445(1.182)	.017(1.062)	.016(1.060)
	500	.730(1.006)	.062(0.928)	.078(0.929)	.728(1.045)	.068(0.966)	.066(0.968)
				$\sigma_0 = 0.2$			
	100	.686(2.191)	.781(2.608)	.787(2.546)	.706(2.513)	.810(2.986)	.801(2.943)
	300	.762(0.584)	.860(0.716)	.876(0.722)	.788(0.654)	.877(0.807)	.887(0.805)
	500	.771(0.430)	.870(0.533)	.875(0.531)	.825(0.516)	.907(0.609)	.904(0.615)
				$\sigma_0 = 0.2$			
	100	.886(5.666)	.906(6.425)	.915(6.722)	.899(5.882)	.927(6.667)	.913(6.571)
Bootst.	300	.956(1.508)	.958(1.847)	.967(1.913)	.965(1.512)	.962(1.866)	.969(1.877)
	500	.968(1.063)	.972(1.322)	.972(1.332)	.972(1.115)	.971(1.397)	.974(1.391)
				$\sigma_0 = 0.2$			
	100	.913(7.629)	.922(8.846)	.935(8.643)	.929(8.039)	.935(9.057)	.932(9.152)
	300	.969(2.095)	.969(2.589)	.971(2.612)	.974(2.061)	.972(2.566)	.979(2.604)
	500	.978(1.525)	.976(1.881)	.967(1.937)	.981(1.654)	.978(1.979)	.974(2.089)

Table 2: *Nonparametric expectile model coverage probability. The nominal coverage is* 95%*. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic expectile regression CC and bootstrap method corresponds to expectile regression bootstrap CC.*

the distribution, namely $\tau = 0.2, 0.5, 0.8$.

In the homogeneous model, we take $\sigma(X_1, X_2) = \sigma_0$, for $\sigma_0 = 0.2, 0.5, 0.7$. In the heterogeneous model, we take $\sigma(X_1, X_2) = \sigma_0 + 0.8X_1(1 - X_1)X_2(1 - X_2)$. 2000 simulation runs are carried out to estimate the coverage probability.

The upper part of Table 1 shows the coverage probability of the asymptotic CC for nonparametric quantile regression functions. It can be immediately seen that the asymptotic CC performs very poorly, especially when n is small. A comparison of the results with those of one-dimensional asymptotic simultaneous confidence bands derived in Claeskens and Van Keilegom (2003) or Fan and Liu (2013), shows that the accuracy in the two-dimensional case is much worse. Much to our surprise, the asymptotic CC performs better in the case of $\tau = 0.2, 0.8$ than in the case of $\tau = 0.5$. On the other hand, it is perhaps not so amazing to see that asymptotic CCs behave similarly under both homogeneous and heterogeneous models. As a final remark about the asymptotic CC we mention that it is highly sensitive with respect to σ_0. Increasing values of σ_0 yields larger CC, and this may lead to greater coverage probability.

The lower part of Table 1 shows that the bootstrap CCs for nonparametric quantile regression functions yield a remarkable improvement in comparison to the asymptotic CC. For the bootstrap CC the coverage probabilities are in general close to the nominal coverage of 95%. The bootstrap CCs are usually wider, and getting narrower when n increases. Such phenomenon can also be found in the simulation study of Claeskens and Van Keilegom (2003). Bootstrap CCs are less sensitive than asymptotic CCs with respect to the choice σ_0, which is also considered as an advantage. Finally, we note that the performance of bootstrap CCs does not depend on which variance specification is used too.

The upper part of Table 2 shows the coverage probabiltiy of the CC for nonparametric expectile regression functions. The results are similar to the case of quantile regression. The asymptotic CCs do not give accurate coverage probabilities, and in some cases like $\tau = 0.2$ and $\sigma_0 = 0.2$, not a single simulation in the 2000 iterations yields a case where surface is completely covered by the asymptotic CC.

The lower part of Table 2 shows that bootstrap CCs for expectile regression give more accurate approximates to the nominal coverage than the asymptotic CCs. One can see in the parenthesis that the volumes of the bootstrap CCs are significantly larger than those of the asymptotic CCs, especially for small n.

5. Application: a treatment effect study

The classical application of the proposed method consists in testing the hypothetical functional form of the regression function. Nevertheless, the proposed method can also be applied to test for a quantile treatment effect (see Koenker; 2005) or to test for conditional stochastic dominance (CSD) as investigated in Delgado and Escanciano (2013). In this section we shall apply the new method to test these hypotheses for data collected from a real government intervention.

The estimation of the quantile treatment effect (QTE) recovers the heterogeneous impact of intervention on various points of the response distribution. To define QTE, given vector-valued exogenous variables $\boldsymbol{X} \in \boldsymbol{\mathcal{X}}$ where $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^d$, suppose Y_0 and Y_1 are response variables associated with the control group and treatment group, and let $F_{0|\boldsymbol{X}}$ and $F_{1|\boldsymbol{X}}$ be the conditional distribution for Y_0 and Y_1, the QTE at level τ is defined by

$$\Delta_\tau(\boldsymbol{x}) \stackrel{\text{def}}{=} Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x}) - Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x}), \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \tag{26}$$

where $Q_{0|\boldsymbol{X}}(y|\boldsymbol{x})$ and $Q_{1|\boldsymbol{X}}(y|\boldsymbol{x})$ are the conditional quantile of Y_0 given $\boldsymbol{X}$ and Y_1 given $\boldsymbol{X}$ respectively. This definition corresponds to the idea of horizontal distance between the treatment and control distribution functions appearing in Doksum (1974) and Lehmann (1975).

A related concept in measuring the efficiency of a treatment is the so called "conditional stochastic dominance". Y_1 conditionally stochastically dominates Y_0 if

$$F_{1|\boldsymbol{X}}(y|\boldsymbol{x}) \leq F_{0|\boldsymbol{X}}(y|\boldsymbol{x}) \quad \text{a.s. for all } (y, \boldsymbol{x}) \in (\mathcal{Y}, \boldsymbol{\mathcal{X}}), \tag{27}$$

where $\mathcal{Y}$, $\boldsymbol{\mathcal{X}}$ are domains of Y and $\boldsymbol{X}$. For example, if Y_0 and Y_1 stand for the income of two groups of people G_0 and G_1, (27) means that the distribution of Y_1 lies on the right of that of Y_0, which is equivalent to saying that at a given $0 < \tau < 1$, the τ-quantile of Y_1 is greater than that of Y_0. Hence, we could replace the testing problem (27) by

$$Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x}) \geq Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x}) \quad \text{for all } 0 < \tau < 1 \text{ and } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}. \tag{28}$$

Comparing (28) and (26), one would find that (28) is just a uniform version of the test $\Delta_\tau(\boldsymbol{x}) \geq 0$ over $0 < \tau < 1$.

The method that we introduced in this paper is suitable for testing a hypothesis like $\Delta_\tau(\boldsymbol{x}) = 0$ where $\Delta_\tau(\boldsymbol{x})$ is defined in (26). One can construct CCs for $Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x})$ and $Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x})$ respectively, and then check if there is overlap between the two confidence regions. One can also extend this idea to test (28) by building CCs for several selected levels τ.

We use our method to test the effectiveness of the National Supported Work (NSW) demonstration program, which was a randomized, temporary employment program initiated in 1975 with the goal to provide work experience for individuals who face economic and social problems prior to entering the program. The data have been widely applied to examine techniques which estimate the treatment effect in a nonexperimental setting. In a pioneer study, LaLonde (1986) compares the treatment effect estimated from the experimental NSW data with that implied by nonexperimental techniques. Dehejia and Wahba (1999) analyse a subset of Lalonde's data and propose a new estimation procedure for nonexperimental treatment effect giving more accurate estimates than Lalonde's estimates. The paper that is most related to our study is Delgado and Escanciano (2013). These authors propose a test for hypothesis (27) and apply it to Lalonde's data, in which they choose "age" as the only conditional covariate and the response variable being the increment of earnings from 1975 to 1978. They cannot reject the null hypothesis of nonnegative treatment effect on the earnings growth.

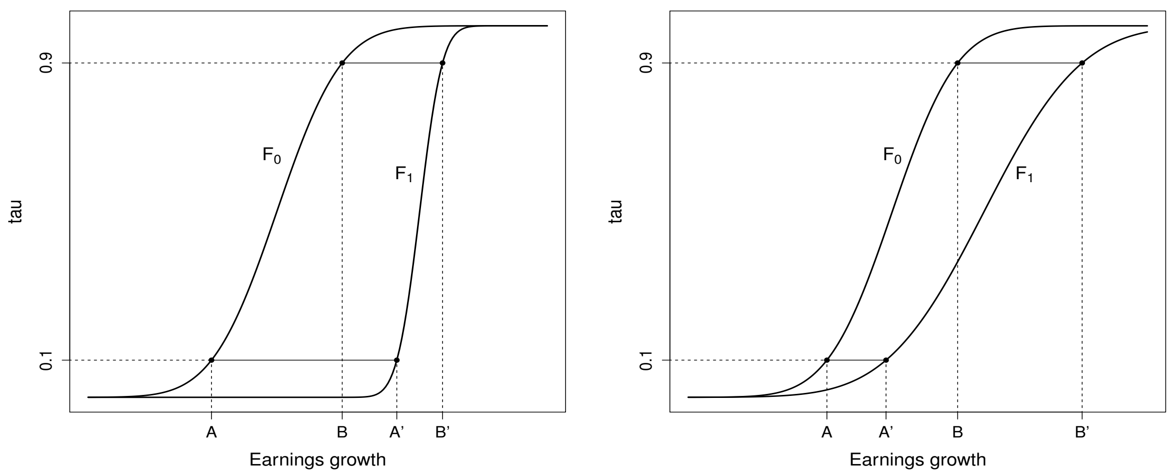


Figure 1: The illustrations for the two possible types of stochastic dominance.

The previous literature, however, has not addressed an important question. We shall depict this question by two pictures. In Figure 1, it is obvious that Y_1 stochastically dominates Y_0 in both pictures, but significant differences can be seen between them. For the left one, the 0.1 quantile improves more dramatically than the 0.9 quantile, as the distance between A and A' is greater than that between B and B'. In usual words, the gain of the 90% lower bound of the earnings growth is more than that of the 90% upper bound of the earnings growth after the treatment. "90% lower bound of the earnings growth" means the probability that the earnings growth is above the bound is 90%. This suggests that the treatment induces greater reduction in downside risk but less increase in the upside potential in the earnings growth. For the right picture the interpretation is just the opposite.

To see which type of stochastic dominance the NSW demonstration program belongs to, we apply the same data as Delgado and Escanciano (2013) for testing the hypothesis of positive quantile treatment effect for several quantile levels τ. The data consist of 297 treatment group observations and 423 control group observations. The response variable Y_0 (Y_1) denotes the difference in earnings of control (treatment) group between 1978 (year of postintervention) and 1975 (year of preintervention). We first apply common statistical procedures to describe the distribution of these two variables. Figure 2 shows the unconditional densities and distribution function. The cross-validated bandwidth for $\hat{f}_0(y)$ is 2.273 and 2.935 for $\hat{f}_1(y)$. The left figure of Figure 2 shows the unconditional densities of the income difference for treatment group and control group. The density of the treatment group has heavier tails while the density of the control group is more concentrated around zero. The right figure shows that the two unconditional distribution functions are very close on the left of the 50% percentile, and slight deviation appears when the two distributions are getting closer to 1. Table 3 shows that, though the differences are small, but the quantiles of the unconditional cdf of treatment group are mildly greater than that of the control group for each chosen τ. The two-sample Kolmogorov-Smirnov and Cramér-von Mises tests, however, yield results shown in the Table 4 which cannot reject the null hypothesis that the empirical cdfs for the two groups are the same with confidence levels 1% or 5%.

τ(%)	10	20	30	50	70	80	90
Treatment	-4.38	-1.55	0.00	1.40	5.48	8.50	11.15
Control	-4.91	-1.73	-0.17	0.74	4.44	7.16	10.56

Table 3: The unconditional sample quantiles of treatment and control groups.

Next we apply our test on quantile regression to evaluate the treatment effect. In order to compare with Delgado and Escanciano (2013), we first focus on the case of a one-dimensional covariate. The first covariate X_{1i} is the age. The second covariate X_{2i} is the number of years

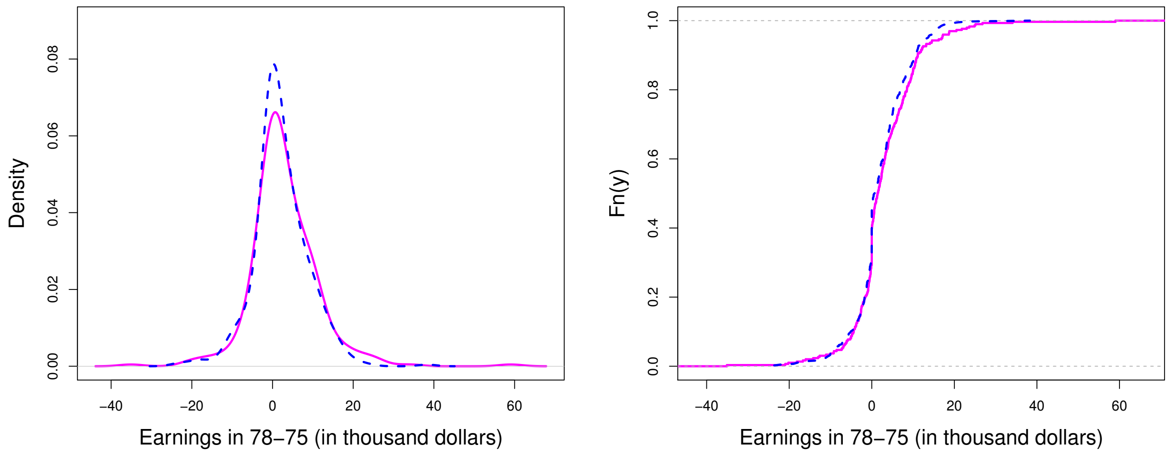


Figure 2: Unconditional empirical density function (left) and distribution function (right) of the difference of earnings from 1975 to 1978. The dashed line is associated with the control group and the solid line is associated with the treatment group.

Type of test	Statistics	p-value
Kolmogorov-Smirnov	0.0686	0.3835
Cramér-von Mises	0.2236	0.7739

Table 4: The two sample empirical cdf tests results for treatment and control groups.

of schooling. The sample values of schooling years lie in the range of $[3, 16]$ and age lies between $[17, 55]$. In order to avoid boundary effect and sparsity of the samples, we look at the ranges [7,13] for schooling years and [19,31] for age. We apply the bootstrap CC method for quantiles $\tau = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8$ and 0.9. We apply the quartic kernel. The cross-validated bandwidths are chosen in the same way as for conditional densities with the `R` package `np`. The resulting bandwidths are (2.2691,2.5016) for the treatment group and (2.7204, 5.9408) for the control group. In particular, for smoothing the data of the treatment group, for $\tau = 0.1$ and 0.9, we enlarge the cross-validated bandwidths by a constant of 1.7; for $\tau = 0.2, 0.3, 0.7, 0.8$, the cross-validated bandwidths are enlarged by constant factor 1.3. These inflated bandwidths are used to handle violent roughness in extreme quantile levels. The bootstrap CCs are computed with 10,000 repetitions. The level of the test is $\alpha = 5\%$.

The results of the two quantile regressions with one-dimensional covariate, and their CCs for various quantile levels are presented in Figure 3 and 4. We observe that for all chosen quantile levels the quantile estimates associated to the treatment group lie above that of the control group when age is over certain levels, and particularly for $\tau = 10\%, 50\%, 80\%$ and 90%, the quantile estimates for treatment group exceeds the upper CCs for the quantile estimates of the control group. On the other hand, at $\tau = 10\%$, the quantile estimates for the control group drop below the CC for treatment group for age greater than 27. Hence, the results here show a tendency that both the downside risk reduction and the upside potential enhancement of earnings growth are achieved, as the older individuals benefit the most from the treatment. Note that we observe a heterogeneous treatment effect in age and the weakly dominance of the conditional quantiles of the treatment group to that of the control group, i.e., (28) holds for the chosen quantile levels, which are in line with the findings of Delgado and Escanciano (2013). We now turn to Figure 4, where the covariate is the years of schooling. The treatment effect is not significant for conditional quantiles at levels $\tau = 10\%, 20\%$ and 30%. This suggests that the treatment does little to reduce the downside risk of the earnings growth for individuals with various degree of education. Nonetheless, we constantly observe that the regression curves of the treatment group rise above that of the control group after a certain level of the years of schooling for quantiles level $\tau = 50\%, 70\%, 80\%$ and 90%. Notice that for $\tau = 50\%$ and 80% the regression curves associated to the treatment group reach the upper boundary of the CC of the control group. This suggests that the treatment effect tends to raise the upside potential of the earnings growth, in particular for those individuals who spent more years in the school. It is worth noting that we also see a heterogeneous treatment effect in schooling years, although the heterogeneity in education is less strong than the heterogeneity in age.

The previous regression analyses separately conditioning on covariates age and schooling years only give a limited view on the performance of the program, we now proceed to the analysis conditioning on the two covariates jointly (X_{1i}, X_{2i}). The estimation settings are similar to the case of univariate covariate. Figure 5 shows the quantile regression CCs. From a first glance of the pictures, the τ-quantile CC of the treatment group and that of the control group overlap extensively for all τ. We could not find sufficient evidence to reject the null hypothesis that the conditional distribution of treatment group and control group are equivalent. The second observation obtained from comparing subfigures in Figure 6, we

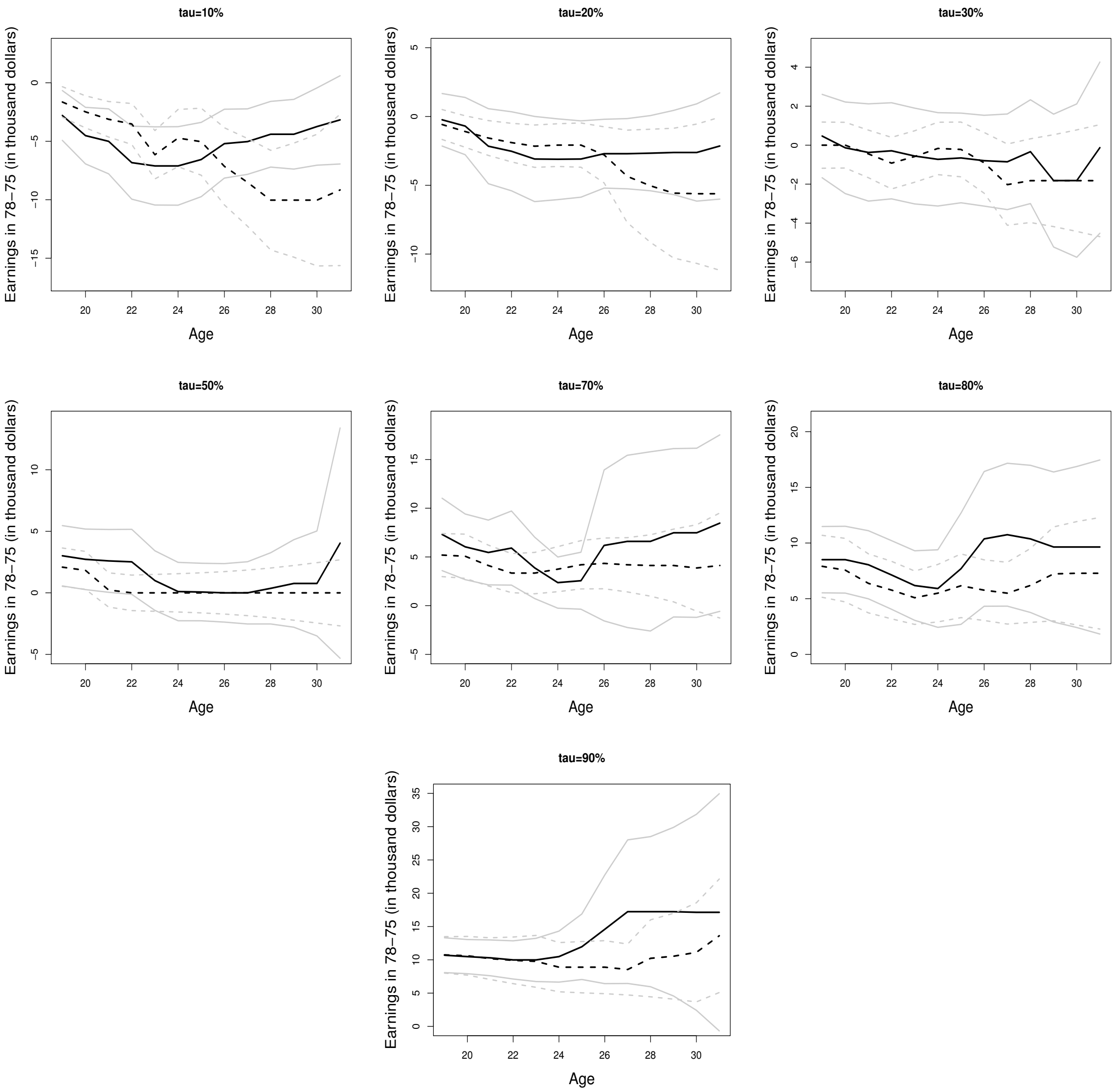


Figure 3: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of age. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.

find that the treatment has larger impact in raising the upper bound of the earnings growth than improving the lower bound. For lower quantile levels $\tau = 10\%, 20\%$ and 30% the solid surfaces uniformly lie inside the CC of the control group, while for $\tau = 50\%, 70\%, 80\%$ and 90%, we see several positive exceedances over the upper boundary of the CC of the control group. Hence, the program tends to do better at raising the upper bound of the earnings growth but does worse at improving the lower bound of the earnings growth. In other words, the program tends to increase the potential for high earnings growth but does

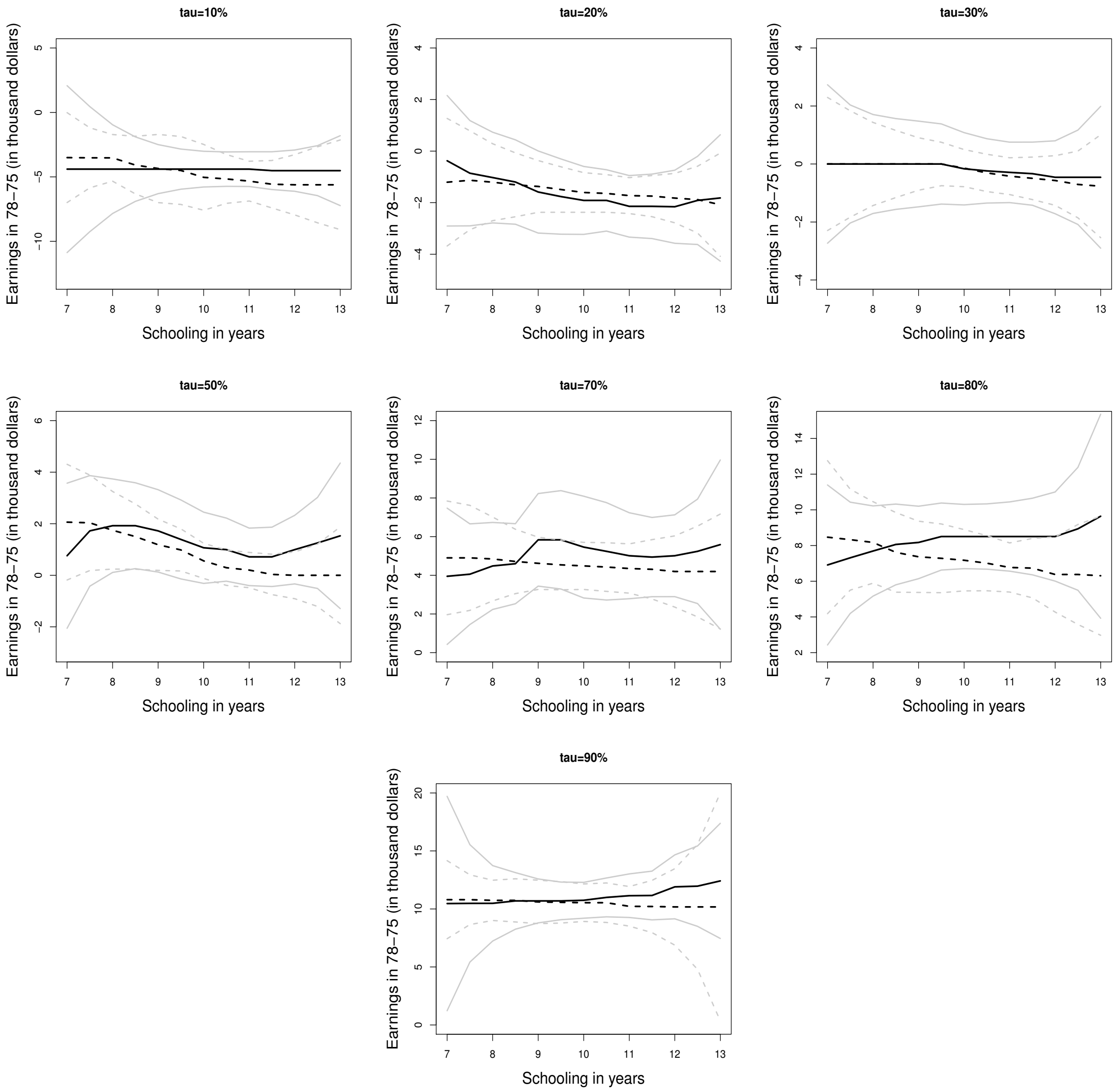
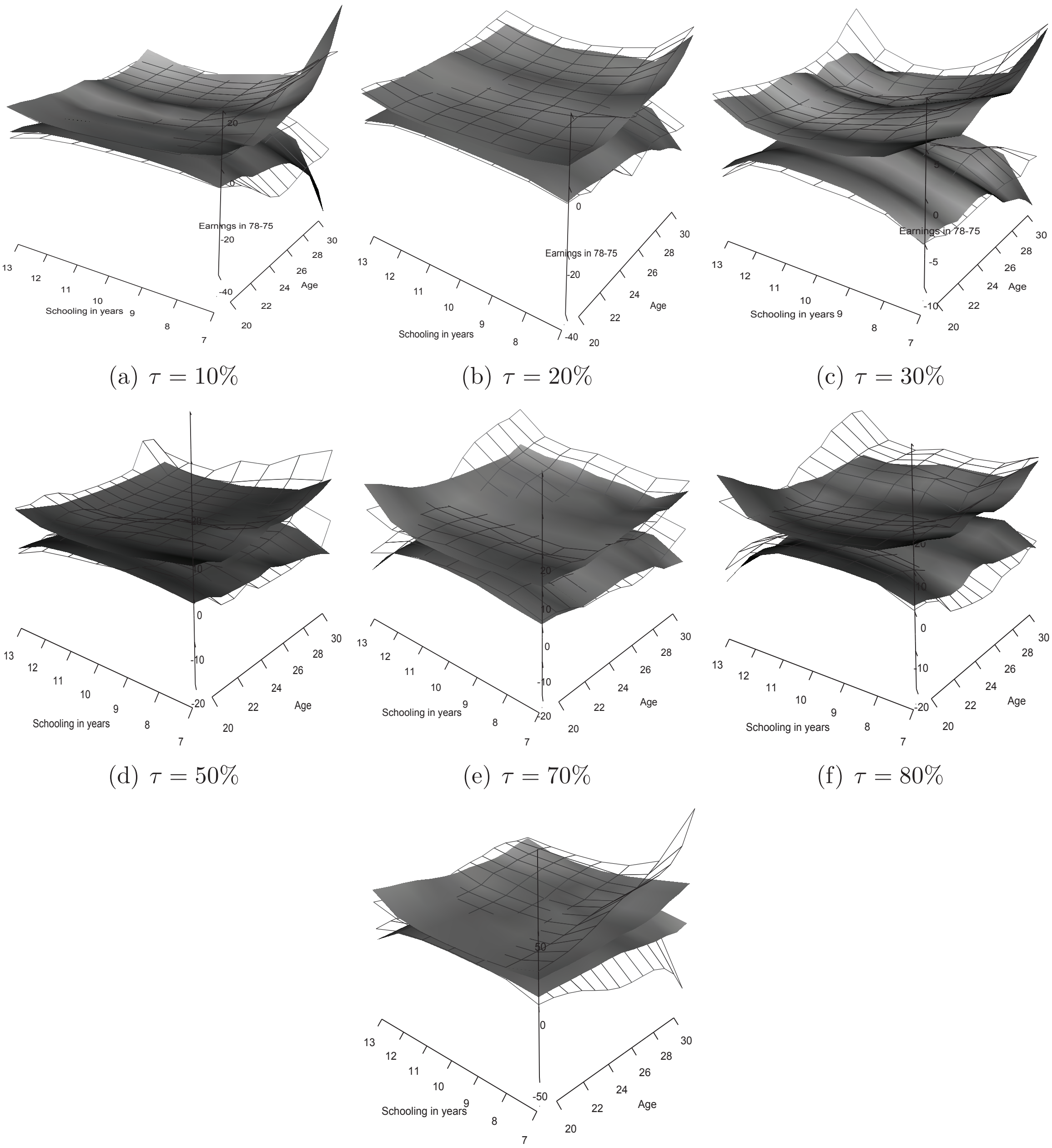


Figure 4: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of years of schooling. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.

little in reducing the risk of negative earnings growth. Our last conclusion comes from inspecting the shape of the surfaces: conditioning on different levels of years of schooling (age), the treatment effect is heterogeneous in age (years of schooling). The most interesting cases happens when conditioning on high age and high years of schooling. Indeed, when considering the cases of $\tau = 80\%$ and 90%, when conditioning on the years of schooling at 12 (corresponding to finishing the high school), the earnings increment of the treatment group rises above the upper boundary of the CC of the control group. This suggests that



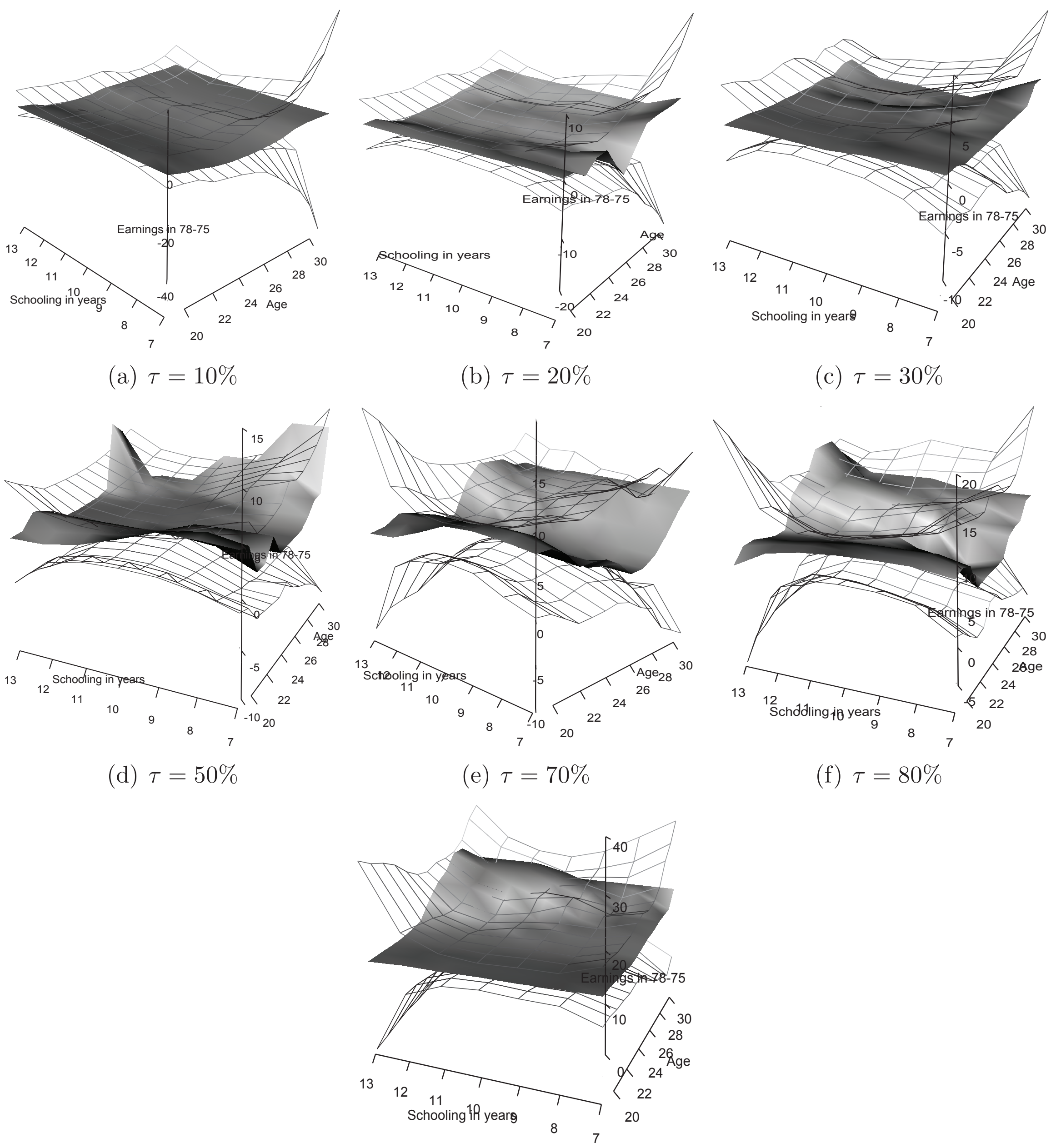
(a) $\tau = 10\%$ (b) $\tau = 20\%$ (c) $\tau = 30\%$

(d) $\tau = 50\%$ (e) $\tau = 70\%$ (f) $\tau = 80\%$

(g) $\tau = 90\%$

Figure 5: The CCs for the treatment group and the control group. The net surface corresponds to the control group quantile CC and the solid surface corresponds to the treatment group quantile CC.

the individuals who are older and have more years of schooling tend to benefit more from the treatment.



(a) $\tau = 10\%$ (b) $\tau = 20\%$ (c) $\tau = 30\%$

(d) $\tau = 50\%$ (e) $\tau = 70\%$ (f) $\tau = 80\%$

(g) $\tau = 90\%$

Figure 6: The conditional quantiles (solid surfaces) for the treatment group and the CCs (net surfaces) for the control group.

Supplementary Materials

Section A contains the detailed proofs of Theorems 2.1, 2.3, 3.1 and Lemmas 2.6 and 3.2, as well as other intermediate results. Section B contains some results obtained by other authors, which we use in our study. We incorporate them here for completeness.

References

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* **1**(6): 1071–1095.

Bissantz, N., Dümbgen, L., Holzmann, H. and Munk, A. (2007). Nonparametric confidence bands in deconvolution density estimation, *Journal of the Royal Statistical Society: Series B* **69**(3): 483–506.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives, *The Annals of Statistics* **31**(6): 1852–1884.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**(448): 1053–1062.

Delgado, M. A. and Escanciano, J. C. (2013). Conditional stochastic dominance testing, *Journal of Business & Economic Statistics* **31**(1): 16–28.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case, *The Annals of Statistics* **2**(2): 267–277.

Fan, Y. and Liu, R. (2013). A direct approach to inference in nonparametric and semiparametric quantile regression models, Preprint.

Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression, *Biometrika* **98**(4): 995–999.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation, *The Annals of Statistics* **38**(2): 1122–1170.

Guerre, E. and Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function, *Econometric Theory* **28**(1): 87–129.

Guo, M. and Härdle, W. (2012). Simultaneous confidence bands for expectile functions, *AStA Advances in Statistical Analysis* **96**(4): 517–541.

Hall, P. (1979). On the rate of convergence of normal extremes, *Journal of Applied Probability* **16**(2): 433–439.

Hall, P. (1991). On convergence rates of suprema, *Probability Theory and Related Fields* **89**(4): 447–455.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density, *The Annals of Statistics* **20**(2): 675–694.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions, *The Annals of Statistics* **41**(4): 1892–1921.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers, *Journal of Multivariate Analysis* **29**(2): 163–179.

Härdle, W. and Song, S. (2010). Confidence bands in quantile regression, *Econometric Theory* **26**(4): 1180–1200.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates, *Journal of Multivariate Analysis* **12**(3): 402–414.

Jones, M. C. (1994). Expectiles and M-quantiles are quantiles, *Statistics & Probability Letters* **20**(2): 149–153.

Kiwitt, S. and Neumeyer, N. (2012). Estimating the conditional error distribution in nonparametric regression, *Scandinavian Journal of Statistics* **39**(2): 259–281.

Koenker, R. (2005). *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, New York.

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**(5): 1529–1564.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *The American Economic Review* **76**(4): 604–620.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Models Based on Ranks*, Springer, San Francisco, CA.

Li, Q., Lin, J. and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions, *Journal of Business & Economic Statistics* **31**(1): 57–65.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton university press, New Jersey.

Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators, *The Annals of Statistics* **39**(1): 201–231.

Mammen, E., Van Keilegom, I. and Yu, K. (2013). Expansion for moments of regression quantiles with applications to nonparametric testing, *ArXiv e-prints* .

Mojirsheibani, M. (2012). A weighted bootstrap approximation of the maximal deviation of kernel density estimates over general compact sets, *Journal of Multivariate Analysis* **112**: 230–241.

Muhsal, B. and Neumeyer, N. (2010). A note on residual-based empirical likelihood kernel density estimator, *Electronic Journal of Statistics* **4**: 1386–1401.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**(4): 819–847.

Proksch, K., Bissantz, N. and Dette, H. (2014). Confidence bands for multivariate and time dependent inverse regression models, *Bernoulli* . To appear.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates, *The*

Annals of Probability **4**(6): 1009–1015.

Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables, *Doklady Akad. Nauk SSSR* **74**: 189–191.

Song, S., Ritov, Y. and Härdle, W. (2012). Partial linear quantile regression and bootstrap confidence bands, *Journal of Multivariate Analysis* **107**: 244–262.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression, *Journal of the American Statistical Association* **93**(441): 228–237.

Assumptions

(A1) K is of order $s-1$ (see (A3)) has bounded support $[-A, A]^d$, is continuously differentiable up to order d with bounded derivatives, i.e. $\partial^{\boldsymbol{\alpha}} K \in L^1(\mathbb{R}^d)$ exists and is continuous for all multi-indices $\boldsymbol{\alpha} \in \{0,1\}^d$

(A2) Let a_n be an increasing sequence, $a_n \to \infty$ as $n \to \infty$, and the marginal density f_Y be such that

$$(\log n) h^{-3d} \int_{|y|>a_n} f_Y(y) dy = \mathcal{O}(1) \tag{29}$$

and

$$(\log n) h^{-d} \int_{|y|>a_n} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy = \mathcal{O}(1), \text{ for all } \boldsymbol{x} \in \mathcal{D}$$

as $n \to \infty$ hold.

(A3) The function $\theta_0(\boldsymbol{x})$ is continuously differentiable and is in Hölder class with order $s > d$.

(A4) $f_{\boldsymbol{X}}(\boldsymbol{x})$ is bounded, continuously differentiable and its gradient is uniformly bounded. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.

(A5) The joint probability density function $f(y, \boldsymbol{u})$ is bounded, positive and continuously differentiable up to sth order (needed for Rosenblatt transform). The conditional density $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ exists and is boudned and continuouly differentiable with respect to $\boldsymbol{x}$. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big) > 0$.

(A6) h satisfies $\sqrt{nh^d}h^s\sqrt{\log n} \to 0$ (undersmoothing), and $nh^{3d}(\log n)^{-2} \to \infty$.

(EA2) $\sup_{\boldsymbol{x}\in\mathcal{D}} \left|\int v^{b_1} f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv\right| < \infty$, for some $b_1 > 0$.

(B1) L is a Lipschitz, bounded, symmetric kernel. G is Lipschitz continuous cdf, and g is the derivative of G and is also a density, which is Lipschitz continuous, bounded, symmetric and five times continuously differentiable kernel.

(B2) $F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ is in $s'+1$ order Hölder class with respect to v and continuous in $\boldsymbol{x}$, $s' > \max\{2, d\}$. $f_{\boldsymbol{X}}(\boldsymbol{x})$ is in second order Hölder class with respect to $\boldsymbol{x}$ and v. $\mathsf{E}[\psi^2(\varepsilon_i)|\boldsymbol{x}]$ is second order continuously differentiable with respect to $\boldsymbol{x} \in \mathcal{D}$.

(B3) $nh_0\bar{h}^d \to \infty$, $h_0, \bar{h} = \mathcal{O}(n^{-\nu})$, where $\nu > 0$.

(C1) There exist an increasing sequence c_n, $c_n \to \infty$ as $n \to \infty$ such that

$$(\log n)^3(nh^{6d})^{-1}\int_{|v|>c_n/2} f_{\varepsilon}(v)dv = \mathcal{O}(1), \tag{30}$$

as $n \to \infty$.

(EC1) $\sup_{\boldsymbol{x}\in\mathcal{D}} \left|\int v^{b} f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv\right| < \infty$, for some $b > 0$.

The assumptions (A1)-(A5) are assumptions frequently seen in the papers of confidence corridors, such as Härdle (1989), Härdle and Song (2010) and Guo and Härdle (2012). (EA2) and (EC1) essentially give the uniform bound on the 2nd order tail variation, which is crucial in the sequence of approximations for expectile regression. (B1)-(B3) are similar to the assumptions listed in chapter 6.1 of Li and Racine (2007). (A6) characterizes the two conflicting conditions: the undersmoothing of our estimator and the convergence of the strong approximation. To make the condition hold, sometimes we need large s for high dimension, the smoothness of the true function. (C1) and (EC1) are relevant to the theory of bootstrap, where we need bounds on the tail probability and 2nd order variation.

Supplement materials to "Confidence Corridors for Multivariate Generalized Quantile Regression"

A. Proof of Theorems

We list the assumptions here for the easy of reference.

(A1) K is of order $s-1$ (see (A3)) has bounded support $[-A, A]^d$, is continuously differentiable up to order d with bounded derivatives, i.e. $\partial^{\boldsymbol{\alpha}} K \in L^1(\mathbb{R}^d)$ exists and is continuous for all multi-indices $\boldsymbol{\alpha} \in \{0,1\}^d$

(A2) Let a_n be an increasing sequence, $a_n \to \infty$ as $n \to \infty$, and the marginal density f_Y be such that

$$(\log n) h^{-3d} \int_{|y|>a_n} f_Y(y) dy = \mathcal{O}(1) \tag{1}$$

and

$$(\log n) h^{-d} \int_{|y|>a_n} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy = \mathcal{O}(1), \text{ for all } \boldsymbol{x} \in \mathcal{D}$$

as $n \to \infty$ hold.

(A3) The function $\theta_0(\boldsymbol{x})$ is continuously differentiable and is in Hölder class with order $s > d$.

(A4) $f_{\boldsymbol{X}}(\boldsymbol{x})$ is bounded, continuously differentiable and its gradient is uniformly bounded. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.

(A5) The joint probability density function $f(y, \boldsymbol{u})$ is bounded, positive and continuously differentiable up to sth order (needed for Rosenblatt transform). The conditional density $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ exists and is boudned and continuouly differentiable with respect to $\boldsymbol{x}$. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big) > 0$.

(A6) h satisfies $\sqrt{nh^d} h^s \sqrt{\log n} \to 0$ (undersmoothing), and $nh^{3d}(\log n)^{-2} \to \infty$.

(EA2) $\sup_{\boldsymbol{x}\in\mathcal{D}} \left| \int v^{b_1} f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) dv \right| < \infty$, for some $b_1 > 0$.

(B1) $L \in L^1(\mathbb{R}^d)$ is a Lipschitz, bounded, symmetric kernel. G is Lipschitz continuous cdf with $G(x), 1 - G(x) \le Ce^{-x}$ for $C > 0$, and $g \in L^1(\mathbb{R})$ is the derivative of G and is also a density, which is Lipschitz continuous, bounded, symmetric and five times continuously differentiable kernel.

(B2) $F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ is in $s'+1$ order Hölder class with respect to v and continuous in $\boldsymbol{x}$, $s' > \max\{2, d\}$. $f_{\boldsymbol{X}}(\boldsymbol{x})$ is in second order Hölder class with respect to $\boldsymbol{x}$ and v. $\mathsf{E}[\psi^2(\varepsilon_i)|\boldsymbol{x}]$ is second order continuously differentiable with respect to $\boldsymbol{x} \in \mathcal{D}$.

(B3) $nh_0\bar{h}^d \to \infty$, $h_0, \bar{h} = \mathcal{O}(n^{-\nu})$, where $\nu > 0$.

(C1) There exist an increasing sequence c_n, $c_n \to \infty$ as $n \to \infty$ such that

$$(\log n)^3 (nh^{6d})^{-1} \int_{|v|>c_n/2} f_\varepsilon(v)dv = \mathcal{O}(1), \tag{2}$$

as $n \to \infty$.

(EC1) $\sup_{\boldsymbol{x}\in\mathcal{D}} \left|\int |v|^b f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv\right| < \infty$, for some $b > 0$.

Define the approximating processes

$$Y_n(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y - \theta_0(\boldsymbol{x}))dZ_n(y, \boldsymbol{u}). \tag{3}$$

$$Y_{0,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y - \theta_0(\boldsymbol{x}))dZ_n(y, \boldsymbol{u}), \tag{4}$$

where $\Gamma_n = \{y : |y| \le a_n\}$ and $\sigma_n^2(\boldsymbol{x}) = \mathsf{E}[\psi^2(Y - \theta_0(\boldsymbol{x}))\mathbf{1}(Y_i \le a_n)|\boldsymbol{X} = \boldsymbol{x}]$.

$$Y_{1,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y - \theta_0(\boldsymbol{x}))dB_n\big(T(y, \boldsymbol{u})\big) \tag{5}$$

where $B_n\big\{T(y, \boldsymbol{u})\big\} = W_n\big\{T(y, \boldsymbol{u})\big\} - F(y, \boldsymbol{u})W_n(1, ..., 1)$ and $T(y, \boldsymbol{u})$ is the Rosenblatt transformation

$$T(y, \boldsymbol{u}) = \big\{F_{X_1|Y}(u_1|y), F_{X_2|Y}(u_2|u_1, y), ..., F_{X_d|X_{d-1},...,X_1,Y}(u_d|u_{d-1}, ..., u_1, y), F_Y(y)\big\}.$$

$$Y_{2,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y - \theta_0(\boldsymbol{x}))dW_n\big(T(y, \boldsymbol{u})\big) \tag{6}$$

$$Y_{3,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y - \theta_0(\boldsymbol{u}))dW_n\big(T(y, \boldsymbol{u})\big) \tag{7}$$

$$Y_{4,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int \sqrt{\sigma_n^2(\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(\boldsymbol{u}\big). \tag{8}$$

$$Y_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(\boldsymbol{u}\big). \tag{9}$$

In these approximating processes, the function

$$\psi_\tau(u) = \begin{cases} \mathbf{1}(u \le 0) - \tau, & \text{Quantile;} \\ 2(\mathbf{1}(u \le 0) - \tau)|u|, & \text{Expectile.} \end{cases}$$

In the proofs, we suppress the subscript "τ".

Next we introduce some notations which are used repeatedly in the following proofs.

Definition 1 (Neighboring Block in $\mathcal{D} \subset \mathbb{R}^d$, Bickel and Wichura (1971) p.1658)**.** A *block* $B \subset \mathcal{D}$ is a subset of $\mathcal{D}$ of the form $B = \Pi_i(s_i, t_i]$ with s and t in $\mathcal{D}$; the *pth-face* of B is $\Pi_{i \neq p}(s_i, t_i]$. Disjoint blocks B and C are *p-neigbbors* if they abut and have the same pth face; they are *neighbors* if they are p-neighbors for some $p \geq 1$.

To illustrate the idea of neighboring block, take $d = 3$ for example, the blocks $(s,t] \times (a,b] \times (c,d]$ and $(t,u] \times (a,b] \times (c,d]$ are 1-neighbors for $s \leq t \leq u$.

Definition 2 (Bickel and Wichura (1971) p.1658)**.** Let $X : \mathbb{R}^d \to \mathbb{R}$. The *increment of* X *on the block* B, denoted $X(B)$, is defined by

$$X(B) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^d} (-1)^{d - |\boldsymbol{\alpha}|} X\big\{\boldsymbol{s} + \boldsymbol{\alpha} \odot (\boldsymbol{t} - \boldsymbol{s})\big\}, \tag{10}$$

where "$\odot$" denotes the *componentwise product*; that is, for any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{u} \odot \boldsymbol{v} = (u_1 v_1, u_2 v_2, ..., u_d v_d)$.

Below we give some examples of the increment of a multivariate function X on a block:

- $d = 1$: $B = (s,t]$, $X(B) = X(t) - X(s)$;
- $d = 2$: $B = (s_1, t_1] \times (s_2, t_2]$. $X(B) = X(t_1, t_2) - X(t_1, s_2) + X(s_1, s_2) - X(s_1, t_2)$.

A.1. Proof of Theorem 2.1

LEMMA A.1.

$$\|Y_n(\boldsymbol{x}) - Y_{0,n}(\boldsymbol{x})\| = \mathcal{O}_p\big\{(\log n)^{-1/2}\big\},$$

where $\|\cdot\|$ *denotes the sup norm with respect to* $\boldsymbol{x} \in \mathcal{D}$.

PROOF. By the triangle inequality we have

$$\|Y_n - Y_{n,0}\| \leq \|Y_n - \hat{Y}_{n,0}\| + \|\hat{Y}_{n,0} - Y_{n,0}\| \stackrel{\text{def}}{=} E_1 + E_2,$$

where $\hat{Y}_{n,0} = \sigma^2(\boldsymbol{x})/\sigma_n(\boldsymbol{x}) Y_{n,0}(\boldsymbol{x})$ and the terms E_1 and E_2 are defined in an obvious manner. We now show that $E_j = \mathcal{O}_p\big\{(\log n)^{-1/2}\big\}$, $j = 1, 2$. Note that

$$|\hat{Y}_{n,0}(\boldsymbol{x}) - Y_{n,0}(\boldsymbol{x})| = \Big|\big(\sigma(\boldsymbol{x})/\sigma_n(\boldsymbol{x}) - 1\big) Y_{n,0}(\boldsymbol{x})\Big|.$$

It is shown later that $\|Y_{n,0}\| = \mathcal{O}_p\big(\sqrt{\log n}\big)$, hence it remains to prove that

$$\sup_{\boldsymbol{x} \in \mathcal{D}} \big|\sigma(\boldsymbol{x})/\sigma_n(\boldsymbol{x}) - 1\big| = \mathcal{O}\big\{(\log n)^{-1}\big\}. \tag{11}$$

To this end let $\tilde{\sigma}_n^2 = \mathsf{E}[\psi^2\{Y_i - \theta_0(\boldsymbol{x})\}\mathbf{1}(|Y_i| > a_n)|\boldsymbol{X} = \boldsymbol{x}]$. Since $\sigma_n^2(\boldsymbol{x}) \to \tau(1-\tau) > 0$ for $n \to \infty$, by (1), and $\psi^2(\cdot) \leq \max\{\tau^2, (1-\tau)^2\}$, $\big|(\log n)^2 \tilde{\sigma}_n^2(\boldsymbol{x})/\sigma_n^2(\boldsymbol{x})\big| \leq \big|(\log n) h^d \mathcal{O}(1)\big| \to 0$. Therefore,

$$(\log n) \sup_{\boldsymbol{x} \in \mathcal{D}} \left|\sqrt{\frac{\sigma^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}} - 1\right| = (\log n) \sup_{\boldsymbol{x} \in \mathcal{D}} \left|\sqrt{\frac{\tilde{\sigma}_n^2(\boldsymbol{x}) + \sigma_n^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}} - 1\right| \leq \sup_{\boldsymbol{x} \in \mathcal{D}} \left|\sqrt{\frac{(\log n)^2 \tilde{\sigma}_n^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}}\right| \to 0,$$

as $n \to \infty$, hence $E_2 = o_p\big((\log n)^{-1/2}\big)$. We now use Lemma B.2 in order to show that E_1 too is negligible.

$$
\begin{aligned}
(\log n)^{1/2} E_1 &= (\log n)^{1/2} \sup_{\boldsymbol{x}\in\mathcal{D}} |Y_n(\boldsymbol{x}) - \hat{Y}_{n,0}(\boldsymbol{x})| \\
&= (\log n)^{1/2} \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})}} \int\int_{\{|y|>a_n\}} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi\{y-\theta_0(\boldsymbol{x})\} dZ_n(y,\boldsymbol{u}) \right| \\
&= \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})}} V_n(\boldsymbol{x}) \right|,
\end{aligned}
$$

where

$$
V_n(\boldsymbol{x}) = \sum_{i=1}^n W_{n,i}(\boldsymbol{x}),
$$

and

$$
\begin{aligned}
W_{n,i}(\boldsymbol{x}) = (\log n)^{1/2}(nh^d)^{-1/2} \bigg\{ &\psi(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n) K\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big) \\
&- \mathsf{E}\Big[\psi(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n) K\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big)\Big] \bigg\}.
\end{aligned}
$$

Note that $f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x})\tau(1-\tau) > 0$ for $\boldsymbol{x} \in \mathcal{D}$ by Assumption (A4).

$$
\begin{aligned}
\mathsf{E}[W_{n,i}(\boldsymbol{x})^2] &\le (\log n)(nh^d)^{-1} \mathsf{E}\left[\psi^2(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n) K^2\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big)\right] \\
&\le (\log n)(nh^d)^{-1} C_{\psi,K} \int_{\{|y|>a_n\}} f_Y(y) dy.
\end{aligned}
$$

Thus, from (1),

$$
\mathsf{E}\Big[\Big(\sum_{i=1}^n W_{n,i}(\boldsymbol{x})\Big)^2\Big] \le (\log n) h^{-d} C_{\psi,K} \int_{\{|y|>a_n\}} f_Y(y) dy = h^{2d} \mathcal{O}_p(1) \to 0,
$$

as $n \to \infty$. From Markov's inequality, $|V_n(\boldsymbol{x})| \xrightarrow{p} 0$ for each fixed $\boldsymbol{x} \in \mathcal{D}$.

We now show the tightness of $V_n(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{D}$ in order to obtain the uniform convergence. To simplify the expression, define

$$
g(\boldsymbol{x}) \stackrel{\text{def}}{=} \psi\{y - \theta_0(\boldsymbol{x})\} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right).
$$

Take arbitrary neighboring blocks $B, C \subset \mathcal{D}$ (see Definition 1) and suppose $B = \Pi_{i=1}^d (s_i, t_i]$,

$$
\begin{aligned}
\mathsf{E}[V_n(B)^2]^{1/2} \le (\log n)^{1/2} h^{-d/2} \bigg\{ &\mathsf{E}\Big[\mathbf{1}(Y_i > a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} g\big(\boldsymbol{s} + \boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})\big)\Big)^2\Big] \\
&+ \mathsf{E}\Big[\mathbf{1}(Y_i < -a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} g\big(\boldsymbol{s} + \boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})\big)\Big)^2\Big] \bigg\}^{1/2} \\
&\stackrel{\text{def}}{=} (\log n)^{1/2} h^{-d/2} (I_1 + I_2)^{1/2},
\end{aligned}
$$

where I_1 and I_2 are defined in an obvious manner. When n is large, a_n is large as well and the integral is restricted to the set $\{Y_i > a_n\}$. Taking into account that θ is uniformly bounded on the compact set $\mathcal{D}$ by Assumption (A4) we deduce that $\psi(Y_i - \theta_0(\boldsymbol{x})) = \tau$ for sufficiently large n on the event $\{Y_i > a_n : i = 1, ..., n\}$. Hence, I_1 can be estimated as

$$I_1 \leq \tau^2 \int\int \mathbf{1}(y > a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} K\Big[\big(\boldsymbol{s} + \boldsymbol{\alpha} \odot (\boldsymbol{t} - \boldsymbol{s}) - \boldsymbol{u}\big)/h\Big]\Big)^2 f(y, \boldsymbol{u}) dy d\boldsymbol{u}.$$

Note that

$$\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} K\Big[\big(\boldsymbol{s} + \boldsymbol{\alpha} \odot (\boldsymbol{t} - \boldsymbol{s}) - \boldsymbol{u}\big)/h\Big] = \int_B \partial^{(1,\dots,1)} K\Big(\frac{\boldsymbol{v} - \boldsymbol{u}}{h}\Big) d\boldsymbol{v} \leq h^{-d} C_{K'} \mu(B),$$

where the constant $C_{K'}$ satisfies $\sup_{\boldsymbol{u}\in\mathcal{D}} |\partial^{\boldsymbol{\alpha}} K(\boldsymbol{u})| \leq C_{K'}$ and $\mu(\cdot)$ is the Lebesgue measure. As consequence it follows that

$$I_1 \leq \tau^2 \int\int \mathbf{1}(y > a_n)\big(C_{K'}\mu(B)\big)^2 f(y, \boldsymbol{u}) dy d\boldsymbol{u} = \tau^2 \big(h^{-d} C_{K'} \mu(B)\big)^2 \int_{\{y > a_n\}} f_Y(y) dy.$$

Similarly, $I_2 \leq (1-\tau)^2 \big(C_{K'} h^{-d} \mu(B)\big)^2 \int_{\{y < -a_n\}} f_Y(y) dy$. Hence,

$$\begin{aligned}
\mathsf{E}[V_n(B)^2]^{1/2} &\leq (\log n)^{1/2} h^{-3d/2} C_{K'} \mu(B) \left(\tau^2 \int_{\{y > a_n\}} f_Y(y) dy + (1-\tau)^2 \int_{\{y < -a_n\}} f_Y(y) dy\right)^{1/2} \\
&\leq (\log n)^{1/2} h^{-3d/2} C_{K'} \max(\tau, 1-\tau) \left(\int_{\{|y| > a_n\}} f_Y(y) dy\right)^{1/2} \mu(B).
\end{aligned}$$

Analogously we obtain the estimate

$$\mathsf{E}[V_n(C)^2]^{1/2} \leq (\log n)^{1/2} h^{-3d/2} C_{K'} \max(\tau, 1-\tau) \left(\int_{\{|y| > a_n\}} f_Y(y) dy\right)^{1/2} \mu(C),$$

which finally yields

$$\begin{aligned}
\mathsf{E}[|V_n(B)||V_n(C)|] &\leq \mathsf{E}[|V_n(B)|^2]^{1/2} \mathsf{E}[|V_n(C)|^2]^{1/2} \\
&\leq (\log n) h^{-3d} C_{K'}^2 \max(\tau, 1-\tau)^2 \left(\int_{\{|y| > a_n\}} f_Y(y) dy\right) \mu(C) \mu(B).
\end{aligned}$$

By Assumption (A2) it follows $(\log n) h^{-3d} \int_{\{|y| > a_n\}} f_Y(y) dy$ is bounded. Thus, applying Lemma B.2 with $\gamma_1 = \gamma_2 = \lambda_1 = \lambda_2 = 1$ yields the desired result. □

LEMMA A.2. $\|Y_{0,n} - Y_{1,n}\| = \mathcal{O}_p\big(n^{-1/6} h^{-d/2} (\log n)^{\epsilon + (2d+4)/3}\big)$, *a.s. for any* $\epsilon > 0$.

PROOF. We adopt the notation that if $\boldsymbol{\alpha} \in \{0,1\}^{d+1}$, then we write $\boldsymbol{\alpha} = (\alpha_1, \boldsymbol{\alpha}_2)$ where $\alpha_1 \in \{0,1\}$ and $\boldsymbol{\alpha}_2 \in \{0,1\}^d$. In the computation below, we focus on $B_{\boldsymbol{x}} = \Pi_{j=1}^d \big[x_j - Ah, x_j + Ah\big]$ instead of $\mathbb{R}^d$ since K has compact support. Recall definition 1 of an increment of a function X over a block

B. Integration by parts yields

$$Y_{0,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \Bigg[\int_{B_{\boldsymbol{x}}} \int_{\Gamma_n} Z_n(y, \boldsymbol{u})\ d\big(\psi(y - \theta_0(\boldsymbol{x}))K((\boldsymbol{x} - \boldsymbol{u})/h)\big)$$
$$+ \Big\{ Z_n(\cdot_1, \cdot_2)\psi\big(\cdot_1 - \theta_0(\boldsymbol{x})\big) K\Big(\frac{\boldsymbol{x} - \cdot_2}{h}\Big) \Big\} \big(\Gamma_n \times B_{\boldsymbol{x}}\big) \tag{12}$$
$$+ \Big\{ \sum_{\boldsymbol{\alpha} \in \{0,1\}^{d+1} - \{\boldsymbol{0},\boldsymbol{1}\}} \int\int_{(\Gamma_n \times B_{\boldsymbol{x}})_{\boldsymbol{\alpha}}} Z_n(\cdot_1, \cdot_2)\ d^{\alpha_1}\psi\big(\cdot_1 - \theta_0(\boldsymbol{x})\big)\partial^{\boldsymbol{\alpha}_2} K\big((\boldsymbol{x} - \cdot_2)/h\big) \Big\} \big(\Gamma_n \times B_{\boldsymbol{x}}\big)_{\boldsymbol{1}-\boldsymbol{\alpha}} \Bigg]$$

where $\boldsymbol{1} = (1, ..., 1) \in \{0, 1\}^{d+1}$ and $\boldsymbol{0} = (0, ..., 0) \in \{0, 1\}^{d+1}$. $\big(\Gamma_n \times B_{\boldsymbol{x}}\big)$ is a $d + 1$ dimensional cube. $\cdot_1$ corresponds to the one-dimensional variable y and $\cdot_2$ corresponds to the two-dimensional variable u. The second term in (12) can be evaluated with the formula (10). $\big(\Gamma_n \times B_{\boldsymbol{x}}\big)_{\boldsymbol{1}-\boldsymbol{\alpha}}$ can be viewed as the projection of $\Gamma_n \times B_{\boldsymbol{x}}$ on to the space spanned by those axes whose numbers correspond to positions of ones of the multi-index $\boldsymbol{1} - \boldsymbol{\alpha}$. This leaves us with an $|\boldsymbol{\alpha}|$-fold integral.

Moreover, $d\{\psi(y - \theta_0(\boldsymbol{x}))K((\boldsymbol{x} - \boldsymbol{u})/h)\} = d\psi(y - \theta_0(\boldsymbol{x}))\partial^{\boldsymbol{1}_2} K\big((\boldsymbol{x} - \boldsymbol{u})/h\big)$, where $\boldsymbol{1}_2 = (1, ..., 1) \in \{0, 1\}^d$ and $d\psi(y - \theta_0(\boldsymbol{x})) = \delta_{\theta_0(\boldsymbol{x})}(y)$ denotes the Dirac measure at $\theta_0(\boldsymbol{x})$.

By integration by parts applied to $Y_{1,n}$ and an application of Theorem 3.2 in Dedecker et al. (2014) we obtain for every $\epsilon > 0$,

$$h^{d/2} n^{1/6} (\log n)^{-\epsilon-(2d+4)/3} |Y_{0,n} - Y_{1,n}|$$
$$\le \mathcal{O}(1) \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \right| \Bigg\{ \Big| \int_{B_{\boldsymbol{x}}} dK((\boldsymbol{x} - \boldsymbol{u})/h) \Big|$$
$$+ \Big| \Big\{ \psi\big(\cdot_1 - \theta_0(\boldsymbol{x})\big) K\Big(\frac{\boldsymbol{x} - \cdot_2}{h}\Big) \Big\} \big(\Gamma_n \times B_{\boldsymbol{x}}\big) \Big|$$
$$+ \Big| \sum_{\alpha_1 = 1, \boldsymbol{\alpha}_2 \in \{0,1\}^d - \{\boldsymbol{1}\}} \int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}} \partial^{\boldsymbol{\alpha}_2} K\big((\boldsymbol{x} - \cdot_2)/h\big) \Big| (B_{\boldsymbol{x}})_{\boldsymbol{1}_2 - \boldsymbol{\alpha}_2}$$
$$+ \Big| \sum_{\alpha_1 = 0, \boldsymbol{\alpha}_2 \in \{0,1\}^d - \{\boldsymbol{0}\}} \int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}} \partial^{\boldsymbol{\alpha}_2} K\big((\boldsymbol{x} - \cdot_2)/h\big) \Big| |\psi\big(\cdot_1 - \theta_0(\boldsymbol{x})\big)| \big(\Gamma_n \times (B_{\boldsymbol{x}})_{\boldsymbol{1}_2 - \boldsymbol{\alpha}_2}\big) \Bigg\} \quad a.s. \tag{13}$$

By (A1), K is of bounded variation in the sense of Hardy and Krause (Owen (2005) definition 2), and this leads to the desired result that (13) is bounded. □

LEMMA A.3. $\|Y_{1,n} - Y_{2,n}\| = \mathcal{O}_p\big(h^{d/2}\big)$.

PROOF. Since $B_n\big(T(y, \boldsymbol{u})\big) = W_n\big(T(y, \boldsymbol{u})\big) - F(y, \boldsymbol{u})W(1, ..., 1)$, where $T(y, \boldsymbol{u})$ is the Rosenblatt transformation and the Jacobian of $T(y, \boldsymbol{u})$ is $f(y, \boldsymbol{u})$,by a change of variables and the first order approximation to $f(y, \boldsymbol{x} - h\boldsymbol{v})$:

$$
\begin{aligned}
&|Y_{1,n}(\boldsymbol{x}) - Y_{2,n}(\boldsymbol{x})| \\
&\le \left| \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi(y - \theta_0(\boldsymbol{x})) f(y, \boldsymbol{u}) dy d\boldsymbol{u} \right| |W(1, ..., 1)| \\
&\le \left| \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K(\boldsymbol{v})\, \psi(y - \theta_0(\boldsymbol{x})) f(y, \boldsymbol{x} - h\boldsymbol{v}) h^d dy d\boldsymbol{v} \right| |W(1, ..., 1)| \\
&\le h^{d/2} \left| \int K(\boldsymbol{v})\, d\boldsymbol{v} \right| \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int_{\Gamma_n} |\psi(y - \theta_0(\boldsymbol{x}))| f(y, \boldsymbol{x}) dy + \mathcal{O}(h) \right| |W(1, ..., 1)| \\
&\le h^{d/2} \left| \int K(\boldsymbol{v})\, d\boldsymbol{v} \right| \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \max\{\tau, 1 - \tau\} + \mathcal{O}(h) \right| |W(1, ..., 1)|,
\end{aligned}
$$

note that $|W(1, ..., 1)| = \mathcal{O}_p(1)$. □

LEMMA A.4. $\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p\big(h^{1/2-\delta}\big)$ *for an arbitrarily small* $0 < \delta < 1/2$.

REMARK A.1. *We note that the rate of* $h^{1/2-\delta}$ *is not sharp rate but sufficiently fast for our purpose.*

PROOF. Define

$$
\begin{aligned}
V_n(\boldsymbol{x}) &\overset{\text{def}}{=} Y_{2,n}(\boldsymbol{x}) - Y_{3,n}(\boldsymbol{x}) \\
&= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} \{\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u}))\} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(T(y, \boldsymbol{u})\big). \quad (14)
\end{aligned}
$$

$\|V_n\| = \mathcal{O}_p\big(h^{1/2-\delta}\big)$ if

$$
\lim_{\eta\to\infty} \mathrm{P}\left\{ \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \frac{V(\boldsymbol{x})}{\sqrt{h}} \right| > \eta h^{-\delta} \right\} = 0, \text{ for all } n \in \mathbb{N}.
$$

Since $\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u})) = \operatorname{sign}(\theta_0(\boldsymbol{u}) - \theta_0(\boldsymbol{x}))\mathbf{1}\big\{[\theta_0(\boldsymbol{x}) \wedge \theta_0(\boldsymbol{u}), \theta_0(\boldsymbol{x}) \vee \theta_0(\boldsymbol{u})]\big\}$, thus

$$
\big\{\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u}))\big\}^2 = \mathbf{1}\big\{[\theta_0(\boldsymbol{x}) \wedge \theta_0(\boldsymbol{u}), \theta_0(\boldsymbol{x}) \vee \theta_0(\boldsymbol{u})]\big\}.
$$

By assumption the conditional distribution function $F_{Y|\boldsymbol{X}}$ and the function θ_0 are both continuously differentiable and change of variables and an application of the multivariate mean value theorem gives

$$
\begin{aligned}
\mathsf{E}\left[\left\{\frac{V_n(\boldsymbol{x})}{\sqrt{h}}\right\}^2\right] &= \frac{1}{h^{d+1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int\int_{\Gamma_n} \{\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u}))\}^2 K^2\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) f(y, \boldsymbol{u}) dy d\boldsymbol{u} \\
&\le \frac{1}{h^{d+1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int \big|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{u})|\boldsymbol{u})\big| K^2\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \frac{1}{h f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int K^2(\boldsymbol{z}) \left| \sum_{|\boldsymbol{\alpha}|=1} \partial^{\boldsymbol{\alpha}} \Big(F_{Y|\boldsymbol{X}} \circ \theta_0\Big)(\boldsymbol{\xi}) \right| |h\boldsymbol{z}| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{z} + \mathcal{O}(h) \\
&\le \frac{1}{\sigma_n^2(\boldsymbol{x})} \left\| \sum_{|\boldsymbol{\alpha}|=1} \partial^{\boldsymbol{\alpha}} \Big(F_{Y|\boldsymbol{X}} \circ \theta_0\Big) \right\| \left(\int |\boldsymbol{z}| K^2(\boldsymbol{z}) d\boldsymbol{z} \right) + \mathcal{O}(h),
\end{aligned}
$$

where $\boldsymbol{\xi}$ lies on the line connecting $\boldsymbol{x}$ and $\boldsymbol{u}$. Note that $\sigma_n^2(\boldsymbol{x}) \geq \min\{\tau^2, (1-\tau)^2\}$. It follows from the continuous differentiability of $F_{Y|\boldsymbol{X}}$ and θ_0 that $\left\|\partial^{\boldsymbol{\alpha}}(F_{Y|\boldsymbol{X}} \circ \theta_0)\right\|$ is bounded.

$$\sigma^2 \stackrel{\text{def}}{=} \sup_{\boldsymbol{x}} \mathsf{E}\left[\left(\frac{V_n(\boldsymbol{x})}{\sqrt{h}}\right)^2\right] \leq C + \mathcal{O}(h), \tag{15}$$

Now we compute $d(\boldsymbol{s}, \boldsymbol{t})$ defined in Lemma B.3. Again from $\sigma_n^2(\boldsymbol{x}) \geq \min\{\tau^2, (1-\tau)^2\}$ and (A4),

$$\begin{aligned}
&\mathsf{E}\left[\left(\frac{V_n(\boldsymbol{t}) - V_n(\boldsymbol{s})}{\sqrt{h}}\right)^2\right] \\
&\leq C\frac{1}{h^{d+1}} \int\int_{\Gamma_n} \Big\{\big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u}))\big] K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) \\
&\quad - \big[\psi(y-\theta_0(\boldsymbol{s})) - \psi(y-\theta_0(\boldsymbol{u}))\big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big\}^2 f(y,\boldsymbol{u}) dy d\boldsymbol{u} \\
&= C\frac{1}{h^{d+1}} \int\int_{\Gamma_n} \Big\{\big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u}))\big] \left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right] \\
&\quad - \big[(\psi(y-\theta_0(\boldsymbol{s})) - \psi(y-\theta_0(\boldsymbol{u}))) - (\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u})))\big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big\}^2 f(y,\boldsymbol{u}) dy d\boldsymbol{u},
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathsf{E}\left[\left(\frac{V_n(\boldsymbol{t}) - V_n(\boldsymbol{s})}{\sqrt{h}}\right)^2\right] \\
&\leq \frac{2C}{h^{d+1}} \int\int_{\Gamma_n} \big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{s}))\big]^2 K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) f(y,\boldsymbol{u}) dy d\boldsymbol{u} \\
&\quad + \frac{2C}{h^{d+1}} \int\int_{\Gamma_n} \big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u}))\big]^2 \left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right]^2 f(y,\boldsymbol{u}) dy d\boldsymbol{u} \stackrel{\text{def}}{=} I_1 + I_2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
I_1 &\leq \frac{2C}{h^{d+1}} \int \left|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{t})|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{s})|\boldsymbol{u})\right| K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&\leq \frac{2CD}{h^{d+1}} \|\boldsymbol{s}-\boldsymbol{t}\|_\infty \int K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \leq \frac{2C'D}{h} \|\boldsymbol{s}-\boldsymbol{t}\|_\infty,
\end{aligned}$$

where $\|\boldsymbol{s}-\boldsymbol{t}\|_\infty = \sup_j |s_j - t_j|$. A change of variables and the fact that K is bounded yield

$$\begin{aligned}
I_2 &\leq \frac{2C}{h^{d+1}} \int \left|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{t})|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{u})|\boldsymbol{u})\right| \left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right]^2 f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&\leq \frac{4C}{h} \frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h} \int \left|K(\boldsymbol{z}) - K\left(\boldsymbol{z} + \frac{\boldsymbol{s}-\boldsymbol{t}}{h}\right)\right| d\boldsymbol{z} \\
&\leq 4C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^2} \left[\int_{[-A,A]^d} |K(\boldsymbol{z})| d\boldsymbol{z} + \int_{[-A,A]^d - \frac{\boldsymbol{s}-\boldsymbol{t}}{h}} \left|K\left(\boldsymbol{z} + \frac{\boldsymbol{s}-\boldsymbol{t}}{h}\right)\right| d\boldsymbol{z}\right] = 4C' \frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^2}
\end{aligned}$$

Thus, for the function γ defined in Lemma B.3 we obtain the estimate $\gamma(\epsilon) \leq C(\sqrt{\epsilon}/h)$ and thus

$$Q(m) \leq (2+\sqrt{2})\frac{C}{h}\int_1^\infty \sqrt{m2^{-y^2}}dy \leq C'\frac{\sqrt{m}}{h},$$

where $C' > 0$. Observe that the graph of the inverse of a univariate, injective function $Q(m)$ is its reflection about the line $y = x$, so the inverse of an upper bound for Q would be a lower bound for Q^{-1}. Given the upper bound above, we can therefore bound Q^{-1} from below by

$$Q^{-1}(a) \geq (C')^{-2}h^2a^2.$$

We have $Q^{-1}\big(1/(\eta h^{-\delta})\big) \geq (C')^{-2}\eta^{-1}h^{2+2\delta}$. Applying Lemma B.3 yields

$$\mathrm{P}\left\{\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{V_n(\boldsymbol{x})}{\sqrt{h_n}}\right| > \eta h_n^{-\delta}\right\} \leq C''\eta^d h_n^{-2d(1+\delta)}\exp\left\{-C'''\eta^2h_n^{-2\delta}\right\} \to 0,$$

as $\eta \to \infty$ for all $n \in \mathbb{N}$. □

LEMMA A.5. $Y_{3,n} \stackrel{\mathcal{L}}{=} Y_{4,n}$.

PROOF. Since both processes are Gaussian with mean zero, we only need to check the equality of the covariance functions of the two processes at any given time points $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{D}$. Ignoring the normalizing factors in the front, the covariance of $Y_{3,n}$ function is:

$$\begin{aligned} r_3(\boldsymbol{s},\boldsymbol{t}) &= \int\int_{\Gamma_n}\psi^2\big(y-\theta_0(\boldsymbol{u})\big)K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)f(y,\boldsymbol{u})dyd\boldsymbol{u} \\ &= \int \mathsf{E}\Big[\psi^2\big(Y_i-\theta_0(\boldsymbol{u})\big)\mathbf{1}(|Y_i|\leq a_n)|\boldsymbol{u}\Big]K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\ &= \int \sigma_n^2(\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)d\boldsymbol{u} = r_4(\boldsymbol{s},\boldsymbol{t}) \end{aligned}$$

which is, up to a factor, the covariance function of $Y_{4,n}$. □

LEMMA A.6. $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p(h^{1-\delta})$, *for* $0 < \delta < 1$.

PROOF. We will proceed as in Lemma A.4 and apply Lemma B.3. Set

$$\tilde{Y}(\boldsymbol{x}) \stackrel{\text{def}}{=} Y_{4,n} - Y_{5,n} = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\int\left(\sqrt{\sigma_n^2(\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})}\right)K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)dW(\boldsymbol{u}).$$

Notice that

$$\sigma_n^2(\boldsymbol{u}) = \tau(1-\tau) - \int_{\{|y|>a_n\}}\psi^2\big(y-\theta_0(\boldsymbol{u})\big)f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy,$$

where

$$\int_{\{|y|>a_n\}}\psi^2\big(y-\theta_0(\boldsymbol{u})\big)f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy \leq \int_{\{|y|>a_n\}}f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy.$$

(1) suggests that

$$\int_{\{|y|>a_n\}}f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy = \mathcal{O}(h^d(\log n)^{-1}).$$

Hence, we have $\sigma_n^2(\boldsymbol{u}) \le C_\tau + E_n$, where $E_n = \mathcal{O}(h^d(\log n)^{-1})$, and $C_\tau = \tau(1-\tau)$.

$$\begin{aligned}
\mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t})}{h}\Big)^2\Big] &= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{t})\sigma_n^2(\boldsymbol{t})} \int \left(\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\right)^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{t})\sigma_n^2(\boldsymbol{t})} \int \Big\{ \sqrt{\sigma_n^2(\boldsymbol{u})}\Big[\sqrt{f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{f_{\boldsymbol{X}}(\boldsymbol{t})}\Big] \\
&\qquad + \sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]\Big\}^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&\le 2Ch^{-d-2}\Big\{ \max\{\tau^2, (1-\tau)^2\} \int \Big[\sqrt{f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{f_{\boldsymbol{X}}(\boldsymbol{t})}\Big]^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&\qquad + C\int \Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u}\Big\},
\end{aligned}$$

Since

$$\Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]^2 = \left[\frac{\sigma_n^2(\boldsymbol{u}) - \sigma_n^2(\boldsymbol{t})}{\sqrt{\sigma_n^2(\boldsymbol{u})} + \sqrt{\sigma_n^2(\boldsymbol{t})}}\right]^2 \le CE_n^2 = \mathcal{O}(h^{2d}(\log n)^{-2});$$

moreover, $\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}$ is continuously differentiable on $\mathcal{D}$ by assumption (A4). Along with $\int |z|^2 K(z) < \infty$, we may bound

$$\sup_{\boldsymbol{t}\in\mathcal{D}} \mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t})}{h}\Big)^2\Big] \le C + \mathcal{O}(h^{2d-2}(\log n)^{-2}).$$

On the other hand,

$$\begin{aligned}
&\mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t}) - \tilde{Y}(\boldsymbol{s})}{h}\Big)^2\Big] \\
&\le Ch^{-d-2}\int \Big\{ \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big] K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) \\
&\qquad - \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big\}^2 d\boldsymbol{u} \\
&= Ch^{-d-2}\int \Big\{ \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big] \Big[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big] \\
&\qquad + \Big[\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big\}^2 d\boldsymbol{u} \\
&\le 2Ch^{-d-2}\int \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big]^2 \Big[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big]^2 d\boldsymbol{u} \\
&\qquad + 2Ch^{-d-2}\int \Big[\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big]^2 K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \overset{\text{def}}{=} I_1 + I_2.
\end{aligned}$$

From

$$\Big[\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big]^2 = \left[\frac{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t}) - \sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}{\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} + \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}}\right]^2 \le C\|\boldsymbol{t}-\boldsymbol{s}\|_\infty^2,$$

we obtain

$$I_2 = C\frac{\|\boldsymbol{t}-\boldsymbol{s}\|_\infty^2}{h^2}.$$

By change of variables and a similar argument as to bound I_2 in the proof of Lemma A.4, it follows

$$I_1 \leq C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^3}.$$

Hence, under the condition that $\|\boldsymbol{s}-\boldsymbol{t}\|_\infty < 1$ and $h \to 0$, we conclude that

$$\mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t})-\tilde{Y}(\boldsymbol{s})}{h}\Big)^2\Big] \leq C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^3}. \tag{16}$$

With the same notations as in Lemma B.3, (16) implies $\gamma(\epsilon) \leq Ch^{-3/2}\sqrt{\epsilon}$, which gives $Q(m) \leq Ch^{-3/2}\sqrt{m}$. Therefore,

$$Q^{-1}(a) \geq Ch^3a^2, \tag{17}$$

and

$$Q^{-1}\big((\eta h^{-\delta})^{-1}\big) \geq Ch^3\eta^{-2}h^{2\delta}. \tag{18}$$

Lemma B.3 asserts that

$$\mathrm{P}\Big\{\sup_{\boldsymbol{x}\in\mathcal{D}}\Big|\frac{\tilde{Y}(\boldsymbol{x})}{h}\Big| > \eta h^{-\delta}\Big\} \leq Ch^{-(3+2\delta)d}\eta^{2d}\exp\Big\{-h^{-2\delta}\eta^2\Big\} \to 0,$$

as $\eta \to \infty$ and $h \to 0$. □

Finally, an application of Theorem 2 of Rosenblatt (1976) to $Y_{5,n}(\boldsymbol{x})$ concludes the proof of Theorem 2.1.

A.2. Proof of Theorem 2.3

Now let $\rho_\tau(u) = |\tau - \mathbf{1}(u<0)|u^2$, be the loss function associated to quantile regression. Then $\psi_\tau(u) = -2\{\tau - \mathbf{1}(u<0)\}|u|$ and

$$g(\boldsymbol{x}) = \frac{\partial}{\partial t}\mathsf{E}[\varphi(Y-t)|\boldsymbol{X}=\boldsymbol{x}]\Big|_{t=\theta_0(\boldsymbol{x})} = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big)(2\tau-1)-\tau\big].$$

It is obvious that $g(\boldsymbol{x}) > 0$ for $0<\tau<1$, and consequently

$$S_{n,0,0}(\boldsymbol{x}) = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big)(2\tau-1)-\tau\big]f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s).$$

LEMMA A.7. $\|Y_n - Y_{0,n}\| = o_p\big((\log n)^{-1/2}\big)$.

PROOF. We have $\|Y_n - Y_{0,n}\| \leq \|Y_n - \hat{Y}_{n,0}\| + \|\hat{Y}_{n,0} - Y_{0,n}\|$, where $\hat{Y}_{n,0}$ is defined as in Lemma A.1, with $a_n \asymp (h^{-3d}\log n)^{1/(b_1-2)}$. With such a choice we have

$$h^{-3d}\log n\sup_{\boldsymbol{x}\in\mathcal{D}}\Big|\int_{|y|>a_n} y^2 f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy\Big| = \mathcal{O}(1) \tag{19}$$

which implies $h^{-3d}\log n\int_{|y|>a_n} y^2 f_Y(y)dy = \mathcal{O}(1)$. It follows that $\|Y_n - \hat{Y}_{n,0}\| = o\big((\log n)^{-1/2}\big)$ via similar arguments as in Lemma A.1.

Since

$$\mathsf{E}\big[W_{n,i}^2(\boldsymbol{x})\big] \le (\log n)(nh^d)^{-1}C\int_{|y|>a_n} y^2 f_Y(y)dy,$$

we conclude by Markov's inequality that $|V_n(\boldsymbol{x})| \to 0$ for each $\boldsymbol{x}\in\mathcal{D}$.

As to the tightness, we have

$$\begin{aligned}
I_1 &\le 4\tau^2\int\int \mathbf{1}(y>a_n)\Big[\sum_{\boldsymbol{\alpha}\in\{0,1\}^d}(-1)^{d-|\boldsymbol{\alpha}|}\big(y-\theta_0(\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s}))\big)K\left(\frac{\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})-\boldsymbol{u}}{h}\right)\Big]f(y,\boldsymbol{u})dyd\boldsymbol{u}\\
&\le 8\tau^2\left\{\big(h^{-d}C\mu(B)\big)^2\int_{y>a_n} y^2 f_Y(y)dy + \big(h^{-d}C\mu(B)\big)^2\int_{y>a_n} f_Y(y)dy\right\}\\
&\le 8\tau^2\big(h^{-d}C\mu(B)\big)^2\int_{y>a_n} y^2 f_Y(y)dy.
\end{aligned}$$

Hence,

$$\mathsf{E}\big[V(B)^2\big]^{1/2} \le (\log n)^{1/2}h^{-3d/2}C\left(\int_{y>a_n} y^2 f_Y(y)dy\right)^{1/2}\mu(B).$$

The desired result follows by similar arguments as those used to prove Lemma A.1. □

LEMMA A.8. *If $n^{-1/6}h^{-d/2-3d/(b_1-2)} = \mathcal{O}(n^{-\nu})$, $\nu>0$, $\|Y_{0,n}-Y_{1,n}\| = \mathcal{O}_p\big(n^{-1/6}h^{-d/2}(\log n)^{\epsilon+(2d+4)/3}a_n\big)$ for any $\epsilon>0$.*

PROOF. With similar arguments as in Lemma A.2,

$$\begin{aligned}
&h^{d/2}n^{1/6}(\log n)^{-\epsilon-(2d+4)/3}a_n^{-1}|Y_{0,n}-Y_{1,n}|\\
&\le \mathcal{O}(1)\left|\frac{a_n^{-1}}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\right|\Bigg\{\big|(\tau-1)(\theta_0(\boldsymbol{x})+a_n)+\tau(a_n-\theta_0(\boldsymbol{x}))\big|\left|\int_{B_{\boldsymbol{x}}} dK((\boldsymbol{x}-\boldsymbol{u})/h)\right|\\
&\quad + \big|\tau(a_n-\theta_0(\boldsymbol{x}))+(\tau-1)(a_n-\theta_0(\boldsymbol{x}))\big|\left|K\left(\frac{\boldsymbol{x}-\cdot_2}{h}\right)\right|(B_{\boldsymbol{x}})\\
&\quad + \big|(\tau-1)(\theta_0(\boldsymbol{x})+a_n)+\tau(a_n-\theta_0(\boldsymbol{x}))\big|\left|\sum_{\alpha_1=1,\boldsymbol{\alpha}_2\in\{0,1\}^d-\{\mathbf{1}_2\}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\right|(B_{\boldsymbol{x}})_{\mathbf{1}_2-\boldsymbol{\alpha}_2}\\
&\quad + \big|\tau(a_n-\theta_0(\boldsymbol{x}))+(\tau-1)(a_n-\theta_0(\boldsymbol{x}))\big|\left|\sum_{\alpha_1=0,\boldsymbol{\alpha}_2\in\{0,1\}^d-\{\mathbf{0}_2\}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\right|(B_{\boldsymbol{x}})_{\mathbf{1}_2-\boldsymbol{\alpha}_2}\Bigg\},\ a.s.
\end{aligned}\tag{20}$$

by the assumption on the kernel K, (20) is almost surely bounded bounded. $h^{d/2}n^{1/6}(\log n)^{-\epsilon-(2d+4)/3} = o(1)$ by the choice of a_n given in Lemma A.7. □

LEMMA A.9. $\|Y_{1,n}-Y_{2,n}\| = \mathcal{O}_p\big(h^{d/2}\big)$.

PROOF. Since $B_n\big(T(y,\boldsymbol{u})\big) = W_n\big(T(y,\boldsymbol{u})\big) - F(y,\boldsymbol{u})W_n(1,\dots,1)$, we obtain by a change of variables and a first order approximation to $f(y, \boldsymbol{x} - h\boldsymbol{v})$:

$$
\begin{aligned}
&\|Y_{1,n} - Y_{2,n}\| \\
&\leq 2h^{d/2}\left|\int K(\boldsymbol{v})\,d\boldsymbol{v}\right| \left\|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\int_{\Gamma_n} |\varphi(y - \theta_0(\boldsymbol{x}))| f(y, \boldsymbol{x})dy + \mathcal{O}(h)\right\| |W(1, ..., 1)| \\
&\leq 2h^{d/2}\left|\int K(\boldsymbol{v})\,d\boldsymbol{v}\right| \left\|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \max\{\tau, 1-\tau\}\left|\mathsf{E}\big[|Y_i|\big|\boldsymbol{x}\big] + \theta_0(\boldsymbol{x})\right| + \mathcal{O}(h)\right\| |W(1, ..., 1)|.
\end{aligned}
$$

Note that $|W(1, ..., 1)| = \mathcal{O}_p(1)$, Y_i has a finite second moment by assumption and θ_0 is uniformly bounded on $\mathcal{D}$. □

LEMMA A.10. *$\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p\big(h^{1-\delta}\big)$, where $0 < \delta < 1$.*

PROOF. Note that the derivative of expectile loss function is $2\big[\mathbf{1}(u \leq 0) - \tau\big]|u|$, which is Lipschitz continuous with Lipschitz constant $2\max\{\tau, 1-\tau\}$. Define $V(\boldsymbol{x})$ as in Lemma A.4.

$$
\begin{aligned}
\mathsf{E}\left[\left(\frac{V(\boldsymbol{x})}{h}\right)^2\right] &= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}\int\int_{\Gamma_n} \{\varphi(y - \theta_0(\boldsymbol{x})) - \varphi(y - \theta_0(\boldsymbol{u}))\}^2 K^2\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right) f(y, \boldsymbol{u})dyd\boldsymbol{u} \\
&\leq \frac{C_{\theta_0}\max\{\tau, 1-\tau\}^2}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}\int \left(F_{Y|\boldsymbol{X}}(a_n|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(-a_n|\boldsymbol{u})\right)|\boldsymbol{x} - \boldsymbol{u}|^2 K^2\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\
&\leq \frac{C^2}{h^2 f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}\int K^2(\boldsymbol{z})|h\boldsymbol{z}|^2 f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{z} + O(h) \leq \frac{2C^2}{\sigma_n^2(\boldsymbol{x})}\|K\|_2^2 + \mathcal{O}(h),
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{E}\left[\left(\frac{V(\boldsymbol{t}) - V(\boldsymbol{s})}{h}\right)^2\right] \leq \frac{2C}{h^{d+2}}\int\int_{\Gamma_n} \big[\varphi(y - \theta_0(\boldsymbol{t})) - \varphi(y - \theta_0(\boldsymbol{s}))\big]^2 K^2\left(\frac{\boldsymbol{s} - \boldsymbol{u}}{h}\right) f(y, \boldsymbol{u})dyd\boldsymbol{u} + \\
&\frac{2C}{h^{d+2}}\int\int_{\Gamma_n} \big[\varphi(y - \theta_0(\boldsymbol{t})) - \varphi(y - \theta_0(\boldsymbol{u}))\big]^2 \left[K\left(\frac{\boldsymbol{t} - \boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s} - \boldsymbol{u}}{h}\right)\right]^2 f(y, \boldsymbol{u})dyd\boldsymbol{u} \overset{\text{def}}{=} I_1 + I_2,
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 &\leq \frac{C}{h^{d+2}}\int \|\boldsymbol{t} - \boldsymbol{s}\|_\infty^2 K^2\left(\frac{\boldsymbol{s} - \boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\
&\leq \frac{C}{h^{d+2}}\|\boldsymbol{s} - \boldsymbol{t}\|_\infty^2 \int K^2\left(\frac{\boldsymbol{s} - \boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \leq C\frac{\|\boldsymbol{s} - \boldsymbol{t}\|_\infty^2}{h^2} + \mathcal{O}(1).
\end{aligned}
$$

By a change of variables and a similar argument as used to bound I_2 in Lemma A.4, we obtain

$$
I_2 \leq C\frac{\|\boldsymbol{s} - \boldsymbol{t}\|_\infty}{h^3}.
$$

for $\|\boldsymbol{s} - \boldsymbol{t}\| < 1$. Following the lines of proof of Lemma A.4 or Lemma A.6 completes the proof of the lemma. □

LEMMA A.11. *$Y_{3,n} \overset{d}{=} Y_{4,n}$*

PROOF. The proof resembles the proof for Lemma A.5 and is omitted for brevity. □

LEMMA A.12. *$\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p\big(h^{1-\delta}\big)$, where $0 < \delta < 1$.*

PROOF. The proof resembles the proof for Lemma A.6 by using (19). The details are omitted for brevity. □

A.3. Proof of Lemma 2.6

We first show assertion 1.). Let $\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ be defined as

$$\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} G\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{21}$$

Since $\sup_{\boldsymbol{x}\in\mathcal{D}} |\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) - f_{\boldsymbol{X}}(\boldsymbol{x})| = \mathcal{O}_p(\bar{h}^s + (n\bar{h}^d)^{-1/2}\log n)$, linearisation yields

$$\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = \frac{\tilde{M}(v,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} + R_n,$$

where $R_n = \mathcal{O}_p(\bar{h}^2 + (n\bar{h}^d)^{-1/2}\log n)$ uniformly over $\boldsymbol{x}\in\mathcal{D}$ by assumption (B2), where $\tilde{M}(v,\boldsymbol{x}) = \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} G\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$. By Theorem 6.2. (i) of Li and Racine (2007), $\mathsf{E}\big[\tilde{M}(v,\boldsymbol{x}) - F_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x})\big]$ is of order $\mathcal{O}(h_0^2 + d\bar{h}^2)$. It remains to show that

$$\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}} \left|\tilde{M}(v,\boldsymbol{x}) - \mathsf{E}\big[\tilde{M}(v,\boldsymbol{x})\big]\right| = \mathcal{O}_p\left((n\bar{h}^d)^{-1/2}\log n\right). \tag{22}$$

By Theorem 6.2. (ii) of Li and Racine (2007), $\mathrm{Var}\left(\tilde{M}(v,\boldsymbol{x})\right) = \mathcal{O}\{(n\bar{h}^d)^{-1}\}$. By virtue of a standard δ_n-net discretization argument and the Bernstein inequality we obtain (22).

Next we show that $\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right| = \mathcal{O}_p(h^2 + (nh^d)^{-1/2}\log n)$. We have

$$\begin{aligned}
\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) &= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{G\left(\frac{v-\varepsilon_i}{h_0}\right) - G\left(\frac{v-\hat{\varepsilon}_i}{h_0}\right)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{h_0^{-1} g\left(\frac{v-\varepsilon_i}{h_0}\right)(\varepsilon_i - \hat{\varepsilon}_i)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) + R_{1,n},
\end{aligned}$$

where $R_{1,n}$ is of negligible order by (B1) under the claim in Section 3.3 of Muhsal and Neumeyer (2010). $\varepsilon_i - \hat{\varepsilon}_i = \hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)$, which is stochastically bounded with $h^s + (nh^d)^{-1/2}\log n$, for arbitrary $\delta > 0$. Moreover, observe that

$$\frac{1}{n}\sum_{i=1}^{n} h_0^{-1} g\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$$

is a kernel density estimator which has standard bias and variance and which is is stochastically bounded. Hence, in order to shrink $\mathrm{P}\left\{\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right| > \eta n^{-\lambda}\right\}$, splitting the probability of under the event $\left\{\left|\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right| > h^s + (nh^d)^{-1/2}\log n\right\}$ and its complement, where $n^{-\lambda} = h_0^2 + h^s + \bar{h}^2 + (nh_0\bar{h}^d)^{-1/2}\log n + (nh^d)^{-1/2}\log n$, we get the desired result.

Next we show assertion 2.). Let $\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ be defined as

$$\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}(v-\varepsilon_i)\, L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{23}$$

By standard theory for kernel density estimation, we have

$$\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = \frac{\tilde{m}(v,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} + R_n,$$

where $R_n = \mathcal{O}_p(\bar{h}^s + (n\bar{h}^d)^{-1/2}\log n)$ uniformly over $\boldsymbol{x} \in \mathcal{D}$ by assumption (B2), where $\tilde{m}(v,\boldsymbol{x}) = \tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n g_{h_0}(\varepsilon_i - v) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$. It follows from the standard theory of density estimation that

$$\|\tilde{m}(v,\boldsymbol{x}) - f_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x})\| = \mathcal{O}_p(h_0^2 + \bar{h}^2 + (nh_0\bar{h}^d)^{-1/2}\log n). \tag{24}$$

A Taylor expansion yields

$$\begin{aligned}\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) &= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^n \{g_{h_0}(v-\varepsilon_i) - g_{h_0}(v-\hat{\varepsilon}_i)\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\\ &= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^n \left\{h_0^{-2}g'\left(\frac{v-\varepsilon_i}{h_0}\right)\left(\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) + R_{2,n}\end{aligned}$$

it follows from Muhsal and Neumeyer (2010) that $R_{2,n}$ is negligible under condition (B1). Again

$$\frac{1}{n}\sum_{i=1}^n h_0^{-2}g'\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$$

is a kernel estimator for the derivative of the conditional density function and is thus stochastically bounded. Applying the stochastic bound for $\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)$ and similar probability separating argument for proving 1.), assertion 2.) follows.

For the third estimator 3.), define

$$\tilde{\sigma}^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \psi^2(\varepsilon_i)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}).$$

Using a weak uniform consistency result for kernel regression, see, for instance, Hansen (2008), $\left\|\tilde{\sigma}^2(\boldsymbol{x}) - \sigma^2(\boldsymbol{x})\right\| = \mathcal{O}_p\left(\bar{h}^2 + (n\bar{h}^d)^{-1/2}\log n\right)$. Below we separately discuss the quantile and expectile case.

In the quantile case, $\psi(u) = \mathbf{1}(u<0) - \tau$, then

$$\begin{aligned}\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x}) &= n^{-1}\sum_{i=1}^n \left[\psi^2(\hat{\varepsilon}_i) - \psi^2(\varepsilon_i)\right] L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\\ &= (1-2\tau)n^{-1}\sum_{i=1}^n \left[\mathbf{1}(\hat{\varepsilon}_i<0) - \mathbf{1}(\varepsilon_i<0)\right] L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i).\end{aligned}$$

Note that $\mathbf{1}(\hat{\varepsilon}_i<0) - \mathbf{1}(\varepsilon_i<0) = \mathbf{1}(\theta_0(\boldsymbol{X}_i) < Y_i < \hat{\theta}_n(\boldsymbol{X}_i)) - \mathbf{1}(\hat{\theta}_n(\boldsymbol{X}_i) < Y_i < \theta_0(\boldsymbol{X}_i))$. Applying the fact that $\sup_{\boldsymbol{x}\in\mathcal{D}}|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})|$ stochastically bounded, we first restrict our focus on the event $\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i) < h^s + (nh^d)^{-1/2}\log n$. If $\tau = 1/2$, then $\psi^2(\hat{\varepsilon}_i) - \psi^2(\varepsilon_i) = 0$ and we are done.

Given $\tau \neq 1/2$,

$$
\begin{aligned}
(1-2\tau)^{-1}\mathsf{E}\big[\hat{\sigma}^2(\boldsymbol{x})-\tilde{\sigma}^2(\boldsymbol{x})\big] &= \mathsf{E}\left[\big(\mathbf{1}(\hat{\varepsilon}_i<0)-\mathbf{1}(\varepsilon_i<0)\big)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right] \\
&= 2\int \Big\{F(\hat{\theta}_n(\boldsymbol{u})|\boldsymbol{u})-F(\theta_0(\boldsymbol{u})|\boldsymbol{u})\Big\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\
&= 2\int f(\theta^\dagger(\boldsymbol{u})|\boldsymbol{u})(\hat{\theta}_n(\boldsymbol{u})-\theta_0(\boldsymbol{u}))L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u},
\end{aligned}
$$

where $\theta^\dagger(\boldsymbol{u})$ lies between $\hat{\theta}_n(\boldsymbol{u})$ and $\theta_0(\boldsymbol{u})$. By condition (B2), $f(v|\boldsymbol{x})$ is uniformly bounded, we deduce that $\mathsf{E}\big[\hat{\sigma}^2(\boldsymbol{x})-\tilde{\sigma}^2(\boldsymbol{x})\big]=\mathcal{O}(h^s+(nh^d)^{-1/2}\log n)$. Observe that $(\mathbf{1}(\hat{\varepsilon}_i<0)-\mathbf{1}(\varepsilon_i<0))^2 = \mathbf{1}\big\{[\hat{\theta}_n(\boldsymbol{X}_i)\wedge\theta_0(\boldsymbol{X}_i),\hat{\theta}_n(\boldsymbol{X}_i)\vee\theta_0(\boldsymbol{X}_i)]\big\}$. It follows from similar computations

$$
\mathsf{E}\big[\big\{\hat{\sigma}^2(\boldsymbol{x})-\tilde{\sigma}^2(\boldsymbol{x})\big\}^2\big]=\mathcal{O}\big(h^s+(nh^d)^{-1/2}\log n\big).
$$

Again observe that $\mathbf{1}\big\{[\hat{\theta}_n(\boldsymbol{X}_i)\wedge\theta_0(\boldsymbol{X}_i),\hat{\theta}_n(\boldsymbol{X}_i)\vee\theta_0(\boldsymbol{X}_i)]\big\}$ is independent of the variable $\boldsymbol{x}$, a discretization argument and the Bernstein inequality yield the result that $n^{\lambda_1}\cdot\|\hat{\sigma}^2(\boldsymbol{x})-\tilde{\sigma}^2(\boldsymbol{x})\|$ is stochastically bounded.

For the expectile case, $\psi(u)=2\big[\mathbf{1}(u<0)-\tau\big]|u|$. Since

$$
\begin{aligned}
\psi^2(\varepsilon_i)-\psi^2(\hat{\varepsilon}_i) &= 4\big\{\mathbf{1}(\varepsilon_i<0)-\tau\big\}^2|\varepsilon_i|^2-4\big\{\mathbf{1}(\hat{\varepsilon}_i<0)-\tau\big\}^2|\hat{\varepsilon}_i|^2 \\
&= 4\big\{\mathbf{1}(\hat{\varepsilon}_i<0)-\tau\big\}^2\big(|\varepsilon_i|^2-|\hat{\varepsilon}_i|^2\big)+4\big\{\mathbf{1}(\varepsilon_i<0)-\mathbf{1}(\hat{\varepsilon}_i<0)\big\}^2|\varepsilon_i|^2,
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\hat{\sigma}^2(\boldsymbol{x})-\tilde{\sigma}^2(\boldsymbol{x}) &= 4n^{-1}\sum_{i=1}^n\big\{\mathbf{1}(\hat{\varepsilon}_i<0)-\tau\big\}\big(|\varepsilon_i|^2-|\hat{\varepsilon}_i|^2\big)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&\quad + 4(1-2\tau)n^{-1}\sum_{i=1}^n\big\{\mathbf{1}(\varepsilon_i<0)-\mathbf{1}(\hat{\varepsilon}_i<0)\big\}|\varepsilon_i|^2L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&\stackrel{\text{def}}{=} 4R_{3,n}(\boldsymbol{x})+4(1-2\tau)R_{4,n}(\boldsymbol{x}).
\end{aligned}
$$

Again, it is sufficient to focus on the set $\{|\hat{\theta}_n(\boldsymbol{X}_i)-\theta_0(\boldsymbol{X}_i)|<n^{-\lambda_0}\}$, where $n^{-\lambda_0}\sim h^s+(nh^d)^{-1/2}\log n$. For $R_{3,n}(\boldsymbol{x})$, notice that

$$
|\varepsilon_i|^2-|\hat{\varepsilon}_i|^2=\big(\theta_0(\boldsymbol{X}_i)-\hat{\theta}_n(\boldsymbol{X}_i)\big)\big(\theta_0(\boldsymbol{X}_i)+\hat{\theta}_n(\boldsymbol{X}_i)-2Y_i\big)=R_{5,n}(\boldsymbol{u})\big(2\theta_0(\boldsymbol{u})+R_{5,n}(\boldsymbol{u})-2Y_i\big),
$$

where $\sup_{\boldsymbol{x}\in\mathcal{D}}|R_{5,n}(\boldsymbol{x})|=\mathcal{O}(n^{-\lambda_0})$, so

$$
\begin{aligned}
\mathsf{E}R_{3,n}(\boldsymbol{x}) &= \mathsf{E}\left[\big\{\mathbf{1}(\hat{\varepsilon}_i<0)-\tau\big\}\big(\theta_0(\boldsymbol{X}_i)-\hat{\theta}_n(\boldsymbol{X}_i)\big)\big(\theta_0(\boldsymbol{X}_i)+\hat{\theta}_n(\boldsymbol{X}_i)-2Y_i\big)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right] \\
&= (1-\tau)^2\int\int_{y<\hat{\theta}_n(\boldsymbol{u})}R_{5,n}(\boldsymbol{u})(2\theta_0(\boldsymbol{u})+R_{5,n}(\boldsymbol{u})-2y)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u} \\
&\quad -\tau^2\int\int_{y>\hat{\theta}_n(\boldsymbol{u})}R_{5,n}(\boldsymbol{u})(2\theta_0(\boldsymbol{u})+R_{5,n}(\boldsymbol{u})-2y)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u}.
\end{aligned}
$$

Hence, $|\mathsf{E} R_{3,n}(\boldsymbol{x})| < Cn^{-\lambda_0}$ for some constant C.

$$\begin{aligned}\operatorname{Var}\left\{R_{3,n}(\boldsymbol{x})\right\} &\le n^{-1}\max\{(1-\tau)^2,\tau^2\}^2 \int\int R_{5,n}^2(\boldsymbol{u})(2\theta_0(\boldsymbol{u})+R_{5,n}(\boldsymbol{u})-2y)^2 L_{\bar{h}}^2(\boldsymbol{x}-\boldsymbol{u}) f_{Y|\boldsymbol{X}}(y|\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u}\\ &\le C(n\bar{h}^d)^{-1}n^{-2\lambda_0}.\end{aligned}$$

One can apply discretization and the Bernstein inequality to show that $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|R_{3,n}(\boldsymbol{x})\right| = \mathcal{O}_p(n^{-\lambda_0}\log n)$.

For $R_{4,n}(\boldsymbol{x})$, again suppose without loss of generality that $\{|\hat{\theta}_n(\boldsymbol{X}_i)-\theta_0(\boldsymbol{X}_i)| < n^{-\lambda_0}\}$, where $n^{-\lambda_0} \sim h^s + (nh^d)^{-1/2}\log n$,

$$|\mathsf{E}[R_{4,n}(\boldsymbol{x})]| \le 2\int\left[\int_{|y-\theta_0(\boldsymbol{u})|<R_{6,n}(\boldsymbol{u})} |y-\theta_0(\boldsymbol{u})|^2 f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy\right] |L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})| f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} = \mathcal{O}(n^{-2\lambda_0}).$$

An application of Markov's inequality yields the desired result.

A.4. Proof of Theorem 3.1

Proof of Lemma 3.2. We will discuss the case of quantile and expectile regression separately.

Consider first $\psi(u) = \mathbf{1}(u<0)-\tau$.

$$\sigma_*^2(\boldsymbol{x}) - \hat{\sigma}^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n\left\{\int \psi^2(v)g_{h_0}(v-\hat{\varepsilon}_i) - \psi^2(\hat{\varepsilon}_i)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{25}$$

By a change of variables,

$$\begin{aligned}\left|\int \psi^2(v)g_{h_0}(v-\hat{\varepsilon}_i)dv - \psi^2(\hat{\varepsilon}_i)\right| &\le \int\left|\psi^2(\hat{\varepsilon}_i + wh_0) - \psi^2(\hat{\varepsilon}_i)\right| g(w)dw\\ &\le 2\max\{\tau,1-\tau\}\int|\psi(\hat{\varepsilon}_i+wh_0)-\psi(\hat{\varepsilon}_i)|\,g(w)dw\\ &= C_\tau\Bigg\{\mathbf{1}(\hat{\varepsilon}_i > \log(n)\cdot h_0)\int_{-\infty}^{-\hat{\varepsilon}_i/h_0} g(w)dw + \mathbf{1}(\hat{\varepsilon}_i < -\log(n)\cdot h_0)\int_{-\hat{\varepsilon}_i/h_0}^{\infty} g(w)dw\\ &+ \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0)\int_{\mathbb{R}} g(w)dw\Bigg\}\\ &\le C_\tau\Bigg\{\mathbf{1}(\hat{\varepsilon}_i > \log(n)\cdot h_0)\int_{-\infty}^{-\log(n)} g(w)dw + \mathbf{1}(\hat{\varepsilon}_i < -\log(n)\cdot h_0)\int_{\log(n)}^{\infty} g(w)dw\\ &+ \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0)\Bigg\}\\ &\le C_\tau\left\{\int_{-\infty}^{-\log(n)} g(w)dw + \int_{\log(n)}^{\infty} g(w)dw + \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0)\right\}.\end{aligned}$$

Hence, the sup norm of (25) is bounded by $I_1 + I_2 + \sup_{\boldsymbol{x}}|I_3(\boldsymbol{x})|$, where $I_1 \overset{\text{def}}{=} C_\tau G(-\log n)$, $I_2 \overset{\text{def}}{=} C_\tau\big(1-G(\log n)\big)$ and

$$I_3(\boldsymbol{x}) \overset{\text{def}}{=} n^{-1}\sum_{i=1}^n \mathbf{1}(|\hat{\varepsilon}_i| \le h_0\log n)\big|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\big|/|\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})|,$$

since $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$. I_1 and I_2 decay polynomially in n by assumption (A1). Note that for any $\kappa > 0$,

$$\mathrm{P}\left\{\sup_{\boldsymbol{x}}\left|(1-\mathsf{E})\sum_{i=1}^{n}\mathbf{1}(|\hat{\varepsilon}_i| \le h_0\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|\right| > n(\log n)^{-1}\kappa\right\} \le$$
$$\mathrm{P}\left\{\sup_{\boldsymbol{x}}\left|(1-\mathsf{E})\sum_{i=1}^{n}\mathbf{1}(|\hat{\varepsilon}_i| \le h_0\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|\right| > n(\log n)^{-1}\kappa, \left\|\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\right\| \le E_n\log n\right\}$$
$$+\,\mathrm{P}\left\{\left\|\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\right\| > E_n\log n\right\}, \tag{26}$$

where $E_n = h^s + (nh^d)^{-1/2}\log n$. The uniform convergence of $\hat{\theta}_n(\boldsymbol{x})$ to $\theta_0(\boldsymbol{x})$ yields that

$$\sum_{n=1}^{\infty}\mathrm{P}\left\{\left\|\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\right\| > E_n\log n\right\} < \infty. \tag{27}$$

For the first probability, it is easy to see that it is bounded by the sum

$$A_n + B_n \stackrel{\text{def}}{=} \mathrm{P}\left\{\sup_{\boldsymbol{x}}\left|(1-\mathsf{E})\sum_{i=1}^{n}\mathbf{1}(|\varepsilon_i| \le h_0\log n + E_n\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right| > \frac{1}{2}n(\log n)^{-1}\kappa\right\}$$
$$+\,\mathbf{1}\left(\sup_{\boldsymbol{x}}\left|\mathsf{E}\left[\sum_{i=1}^{n}\mathbf{1}(h_0\log n - E_n\log n < |\varepsilon_i| \le h_0\log n + E_n\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right]\right| > \frac{1}{2}n(\log n)^{-1}\kappa\right).$$

After an explicit computation of the expectation, one concludes that B_n is equal to zero for any $\kappa > 0$ if n is sufficiently large. Now we need to bound A_n. Note that for any fixed $\boldsymbol{x}$, we can estimate the variance by

$$\mathrm{Var}\left(\sum_{i=1}^{n}\mathbf{1}(|\varepsilon_i| \le h_0\log n + E_n\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|\right) \le C_L n h_0 \bar{h}^{-d}\log n,$$

applying a concentration inequality, one gets for any $\kappa > 0$,

$$\mathrm{P}\left\{(1-\mathsf{E})\sum_{i=1}^{n}\mathbf{1}(|\varepsilon_i| \le h_0\log n + E_n\log n)|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| > n(\log n)^{-1}\kappa\right\}$$
$$\le 2\exp\left\{-\frac{1}{4}\frac{n^2(\log n)^{-4}\kappa^2}{C_L n h_0\bar{h}^{-d}\log n + C_L n\bar{h}^{-d}(\log n)^{-2}\kappa}\right\},$$

which decreases exponentially in n since $n\bar{h}^d \to \infty$ polynomially in n by assumption (B3). By a discretization argument, one can show that A_n is also summable (the grid size grows polynomially in n). Hence, we conclude that the probability (26) is summable. The stochastic part of the numerator of $I_3(\boldsymbol{x})$ is therefore of $\mathcal{O}_p((\log n)^{-1})$ a.s. by an application of the Borel-Cantelli lemma.

The mean of the numerator of $I_3(\boldsymbol{x})$ can be estimated by the law of iterative expectation:

$$\mathsf{E}\left[\mathsf{E}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}(|\hat{\varepsilon}_i| \le h_0\log n)\big|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\big|\,\middle|\, X, \hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\right]\right]$$
$$= \mathsf{E}\left[\int_{\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})-h_0\log n}^{\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})+h_0\log n} f\left\{e|X,\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\right\}de|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|\right] \le 2h_0\log nC = \mathcal{O}((\log n)^{-1}),$$

since the density $f\{e|X, \hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\}$ is bounded and $L \in L^1(\mathbb{R}^d)$. Finally, applying a linearization argument we obtain that $\|I_3(\boldsymbol{x})\| = \mathcal{O}_p((\log n)^{-1}) = \scriptstyle\mathcal{O}((\log n)^{-1/2})$ a.s.

In the case of expectile regression, we need to consider $\psi(u) = 2(\mathbf{1}(u<0)-\tau)|u|$, which is Lipschitz continuous (see Lemma A.10). Note that $|\hat{\varepsilon}_i| \le |\varepsilon_i| + E_n$, where $E_n = \mathcal{O}(h^s + (nh^d)^{-1/2}\log n)$ a.s. by the Bahadur representation of θ_n, a discretization argument and an application of the Bernstein inequality. Hence,

$$
\begin{aligned}
&\left| n^{-1}\sum_{i=1}^{n}\left\{\int \psi^2(v)g_{h_0}(v-\hat{\varepsilon}_i) - \psi^2(\hat{\varepsilon}_i)dv\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right| \\
&\le n^{-1}C_\tau \sum_{i=1}^{n}\int h_0(2|\hat{\varepsilon}_i| + h_0|w|)|w|g(w)dw|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| \\
&= C_\tau h_0^2 \int |w|^2 g(w)dw + C_{\tau,g}2h_0 n^{-1}\sum_{i=1}^{n}|\hat{\varepsilon}_i||L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| \\
&\le C_\tau h_0^2 \int |w|^2 g(w)dw + 2C_{\tau,g}h_0 E_n n^{-1}\sum_{i=1}^{n}|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| + 2C_{\tau,g}h_0 n^{-1}\sum_{i=1}^{n}|\varepsilon_i||L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|.
\end{aligned}
$$

The first term converges almost surely to 0, faster than $(\log n)^{-1}$, based on assumption (B3). The second term and the third term can be handled by similar argument for showing the uniform almost sure convergence of the Nadaraya-Watson estimator, see Hansen (2008) for more details.

□

Our strategy is to follow the sequence of approximation steps that are similar to Section A.1 and A.2. Define

$$
Y^*_{0,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_{n,*}(\boldsymbol{x})}} \int\int_{\Gamma^*_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dZ^*_n(v,\boldsymbol{u}), \tag{28}
$$

where $\sigma^2_{n,*}(\boldsymbol{x}) = \mathsf{E}^*\left[\psi_\tau(\varepsilon^*_i)^2\mathbf{1}(|\varepsilon^*_i| < b_n)|\boldsymbol{x}\right]$, and $\Gamma^*_n = \{v : |v| \le b_n\}$.

$$
Y^*_{1,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_{n,*}(\boldsymbol{x})}} \int\int_{\Gamma^*_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dB^*_n\{\hat{T}(v,\boldsymbol{u})\}, \tag{29}
$$

where $B^*_n\{\hat{T}(v,\boldsymbol{u})\} = W^*_n\{\hat{T}(v,\boldsymbol{u})\} - \hat{F}(v,\boldsymbol{u})W^*_n(1,...,1)$, W^* is a Brownian motion defined conditional on the sample, and $\hat{T}(v,\boldsymbol{u})$ is the Rosenblatt transformation:

$$
\hat{T}(v,\boldsymbol{u}) = \left\{\hat{F}_{X_1|\varepsilon}(u_1|v), \hat{F}_{X_2|\varepsilon}(u_2|u_1,v), ..., \hat{F}_{X_d|X_{d-1},...,X_1,\varepsilon}(u_d|u_{d-1},...,u_1,v), \hat{F}_\varepsilon(v)\right\},
$$

given $\hat{F}_{X_1|\varepsilon}(u_1|v), \hat{F}_{X_2|\varepsilon}(u_2|u_1,v), ..., \hat{F}_{X_d|X_{d-1},...,X_1,\varepsilon}(u_d|u_{d-1},...,u_1,v), \hat{F}_\varepsilon(v)$ are associated cdfs obtained from integrating $\hat{f}_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{u})$.

$$
Y^*_{2,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_{n,*}(\boldsymbol{x})}} \int\int_{\Gamma^*_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dW^*_n\{\hat{T}(v,\boldsymbol{u})\}, \tag{30}
$$

$$Y^*_{4,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_{n,*}(\boldsymbol{x})}} \int \sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{u})\sigma^2_{n,*}(\boldsymbol{u})} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW^*_n(\boldsymbol{u}), \tag{31}$$

$$Y^*_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW^*_n(\boldsymbol{u}). \tag{32}$$

From (28) to (29) the proof resembles Lemma A.1 for quantile regression and A.7 for expectile regression. For the bootstrap version of these proofs to hold, it is sufficient to verify the conditions

$$(\log n)h^{-3d} \int_{|v|>c_n} \hat{f}_\varepsilon(v)dv = \mathcal{O}(1), \text{ a.s.} \tag{33}$$

for quantile regression and

$$(\log n)h^{-3d} \int_{|v|>c_n} v^2 \hat{f}_\varepsilon(v)dv = \mathcal{O}(1), \text{ a.s.} \tag{34}$$

for expectile regression, where $\hat{f}_\varepsilon(v) = (nh_0)^{-1}\sum_{i=1}^n g((v-\hat{\varepsilon}_i)/h_0)$. The rest follows from similar arguments in Lemma A.1 and A.7.

We will only consider the kernel g with compact support; in particular, with support $[-1,1]$. Via standard arguments one could generalize the result here immediately to, e.g., the Gaussian kernel.

Let $\delta_n = (\log n)^{-1}h^{3d}$. Let $E_n = h^s + (nh^d)^{-1/2}\log n$.

$$\begin{aligned}
\int_{|v|>c_n} \hat{f}_\varepsilon(v)dv &= \frac{1}{nh_0}\sum_{i=1}^n \int_{|v|>c_n} g\left(\frac{\hat{\varepsilon}_i - v}{h_0}\right) dv \le \frac{1}{nh_0} C_g \sum_{i=1}^n \int_{|v|>c_n} \mathbf{1}(|\hat{\varepsilon}_i - v| \le h_0)dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \int_{|v|>c_n} \mathbf{1}(|v| - |\hat{\varepsilon}_i| \le h_0)\mathbf{1}(\hat{\varepsilon}_i - h_0 \le v \le \hat{\varepsilon}_i + h_0)dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \int_{|v|>c_n} \mathbf{1}(\hat{\varepsilon}_i - h_0 \le v \le \hat{\varepsilon}_i + h_0)dv \\
&\le \frac{2}{n} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)
\end{aligned} \tag{35}$$

where C_g is a constant depending on g. For any $\kappa > 0$ and a constant $\lambda > 0$ small such that $E_n n^\lambda \to 0$ as $n \to \infty$, consider

$$\begin{aligned}
&\mathsf{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right| > 2\delta_n\kappa\right\} \le \\
&\mathsf{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right| > 2\delta_n\kappa, \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| \le E_n n^\lambda\right\} + \mathsf{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda\right\} \\
&\stackrel{\text{def}}{=} P_{1,n} + P_{2,n}.
\end{aligned}$$

$P_{2,n}$ is summable by similar argument in the proof of Lemma 3.2. Without loss of generality, we assume c_n is large enough so that $h_0 + E_n n^\lambda < c_n/2$ since $h_0, E_n n^\lambda \to 0$. Thus,

$$P_{1,n} \le \mathrm{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n/2 \le |\varepsilon_i|)\right| > 2\delta_n\kappa\right\}.$$

Let $S_n = \sum_{i=1}^{n}\mathbf{1}(c_n/2 \le |\varepsilon_i|)$. From (2) in assumption (C2),

$$\mathrm{Var}(S_n) = n\int_{|v|>c_n/2} f_\varepsilon(v)dv = \mathcal{O}(n^2(\log n)^{-3}h^{6d}) = \mathcal{O}(n^2(\log n)^{-1}\delta_n^2).$$

This yields

$$\sum_{n=1}^{\infty}\mathrm{P}\left(|S_n| > 2\kappa n\delta_n\right) \le 2\sum_{n=1}^{\infty}\exp\left\{-\frac{4n^2\kappa^2\delta_n^2}{4\,\mathrm{Var}(S_n)+8n\kappa\delta_n}\right\} = 2\sum_{n=1}^{\infty}\exp\left\{-\frac{\kappa^2\log n}{1+2\kappa\log^2 n/(nh^{3d})}\right\} < \infty, \tag{36}$$

given that $\kappa > 1$, since $nh^{3d}(\log n)^{-2} \to \infty$ by assumption (A7). It follows by the Borel-Cantelli lemma that the stochastic part of (35) is of $\mathcal{O}_p(\delta_n)$. For the expectation, we note that

$$\begin{aligned}
&\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \le \mathbf{1}(c_n - h_0 \le |\varepsilon_i| + \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\|)\\
&\le \mathbf{1}(c_n - h_0 - E_n n^\lambda \le |\varepsilon_i|)\mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| \le E_n n^\lambda)\\
&\qquad + \mathbf{1}(c_n - h_0 \le |\varepsilon_i| + \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\|)\mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda)\\
&\le \mathbf{1}(c_n/2 \le |\varepsilon_i|) + \mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda).
\end{aligned} \tag{37}$$

Therefore,

$$\begin{aligned}
&\mathsf{E}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right] \le \mathsf{E}\left[\mathbf{1}(c_n/2 \le |\varepsilon_i|)\right] + \mathrm{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda\right\}\\
&= \int_{|v|>c_n/2} f_\varepsilon(v)dv + \mathcal{O}(e^{-n^{\mu_1}}) = \mathcal{O}((\log n)^{-3}nh^{6d}),
\end{aligned}$$

for some $\mu_1 > 0$.

Next we show (34). The sequence c_n will be chosen appropriately later,

$$\begin{aligned}
\int_{v>c_n} v^2 \hat{f}_\varepsilon(v)dv &\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \int_{|v|>c_n} v^2 \mathbf{1}(|v| \le h_0 + |\hat{\varepsilon}_i|)dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)(2h_0\hat{\varepsilon}_i^2 + 2h_0^3) \\
&\le \frac{2}{n} C_g \sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) + \underbrace{\frac{2h_0^2}{n} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)}_{T_{1,n}} \\
&\le \frac{4}{n} C_g \sum_{i=1}^n \varepsilon_i^2 \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) + \underbrace{\frac{4}{n} C_g \sum_{i=1}^n \left[\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right]^2 \mathbf{1}(c_n - h_0 \le \hat{\varepsilon}_i)}_{T_{2,n}} + T_{1,n} \\
&= T_{3,n} + T_{2,n} + T_{1,n}. \qquad (38)
\end{aligned}$$

Choosing $c_n \asymp (n^{4/b-1}(\log n)^{1+8/b}\delta_n^{-2})^{1/(b-2)}$. Note $c_n > ((\log n)^3(nh^{6d})^{-1})^{1/b}$, and therefore (2) holds naturally in this case, by assumption (EC1),

$$\int_{|v|>c_n} f_\varepsilon(v)dv \le \int_{|v|>c_n} \frac{|v|^b}{|c_n|^b} f_\varepsilon(v)dv = \mathcal{O}(c_n^b) = \mathcal{O}(n^{4/b-1}(\log n)^{1+8/b}\delta_n^{-2}).$$

It can be shown via similar arguments for showing (33) that $T_{i,n} = \mathcal{O}_p^*((\log n)^{-1}h^{3d})$ a.s. for $i = 1, 2$.

To bound $T_{3,n}$, given b from (EC1), we choose $M_n = n^{1/b}(\log n)^{2/b}$ and obtain

$$\begin{aligned}
&\mathrm{P}\left\{|(1-\mathsf{E})T_{3,n}| > 2\delta_n\kappa\right\} \\
&\le \mathrm{P}(|(1-\mathsf{E})S_n'| > 2n\kappa\delta_n, \varepsilon_i < M_n, \forall i) + n\mathrm{P}(|\varepsilon_i| \ge M_n) + \mathrm{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda\right\} \\
&\overset{\text{def}}{=} U_{1,n} + U_{2,n} + U_{3,n},
\end{aligned}$$

where $S_n' = C_g \sum_{i=1}^n \varepsilon_i^2 \mathbf{1}(c_n/2 \le |\varepsilon_i|)$, the term $U_{2,n}$ is of order $o(M_n^{-b})$ by (EC1) and hence summable. $U_{3,n}$ is summable by a similar argument as used in the proof of (33). Restricting S_n' to the set $\cap_{i=1}^n \{|\varepsilon_i| < M_n\}$, we find

$$\operatorname{Var}(S_n') \le M_n^4 n C_g^2 \int_{c_n/2}^\infty f_\varepsilon(v)dv \le C_{g,b} M_n^4 n c_n^{-b} = \mathcal{O}(n^2(\log n)^{-1}\delta_n^2).$$

This yields

$$\sum_{n=1}^\infty U_{1,n} \le 2\sum_{n=1}^\infty \exp\left\{-\frac{4n^2\kappa^2\delta_n^2}{4\operatorname{Var}(S_n') + 8n\kappa\delta_n}\right\} = 2\sum_{n=1}^\infty \exp\left\{-\frac{\kappa^2\log n}{1 + 2\kappa\log^2 n/(nh^{3d})}\right\} < \infty, \qquad (39)$$

given that $\kappa > 1$ and assumption (EA2). It follows by the Borel-Cantelli lemma that $(1-\mathsf{E})T_{3,n} = \mathcal{O}(\delta_n)$ a.s. It left to control the expectation. By computation in (37),

$$\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \le \mathbf{1}(c_n - h_0 \le \mathbf{1}(c_n/2 \le |\varepsilon_i|) + \mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda).$$

Thus, by law of iterative expectation,

$$\begin{aligned}\mathsf{E}[T_{3,n}] &\le \mathsf{E}\big[\varepsilon_i \mathbf{1}(c_n/2 \le |\varepsilon_i|)\big] + \mathsf{E}\left[n^{-1}\sum_{i=1}^{n} \varepsilon_i \mathrm{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^{\lambda}\right\}\right] \\ &= \mathcal{O}(c_n^{2-b}) + \mathcal{O}(e^{-n^{\mu_2}}).\end{aligned}$$

It follows immediately by the order of c_n that $\mathsf{E}[T_{3,n}] = \mathcal{O}(\delta_n)$.

In order to show the almost sure uniform convergence of $Y^*_{4,n}(\boldsymbol{x})$ to $Y^*_{5,n}(\boldsymbol{x})$ we need to verify that for quantile regression

$$h^{-d} \log n \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \int_{|v|>c_n} \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv \right| = \mathcal{O}(1), \quad \text{a.s.} \tag{40}$$

and for expectile regression

$$h^{-d} \log n \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \int_{|v|>c_n} v^2 \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv \right| = \mathcal{O}(1). \quad \text{a.s.} \tag{41}$$

The first condition can be shown in the same way as showing (33), and the second one is similar to (34) given $b \ge 4$. A discretization argument is needed in both cases, but the grid size only grows in polynomial rate in n. The proofs are omitted for brevity.

Using analogous arguments as in Lemma A.1 for quantile regression and A.7 for expectile regression with (33) and (34), it can be shown that $Y^*_n(\boldsymbol{x})$ converges uniformly in probability to $Y^*_{0,n}(\boldsymbol{x})$. The almost sure uniform convergence in probability of $Y^*_{0,n}(\boldsymbol{x})$ to $Y^*_{5,n}(\boldsymbol{x})$ follows by similar arguments in Lemma A.2, A.3, A.5 and A.6 for quantile regression and Lemma A.8, A.9, A.11 and A.12 for expectile regression, except that $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\sigma_n^2(\boldsymbol{x})$, $F(y,\boldsymbol{x})$ are replaced by $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})$, $\sigma^2_{*,n}(\boldsymbol{x})$, $\hat{F}(v,\boldsymbol{x})$ respectively, and that the approximation shown in Lemma A.4 and A.10 is not needed here. Finally, the proof of Theorem 3.1 is completed by an application of the extreme value theorem of Rosenblatt (1976) to $Y^*_{5,n}(\boldsymbol{x})$.

B. Supporting Lemmas

LEMMA B.1 (Kong et al. (2010))**.** *Under (A1),(A3)-(A5), for some $s \ge 0$, and $\mathcal{D}$ is an compact subset of $\mathbb{R}^d$. Then*

$$\sup_{x\in\mathcal{D}} \left| H_n\left\{\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})\right\} - \beta^*_n(\mathbf{x})\right| = \mathcal{O}\left(\left\{\frac{\log n}{nh^d}\right\}^{\lambda(s)}\right). \tag{42}$$

where

$$\beta^*_n(\mathbf{x}) = -\frac{1}{nh^d} S^{-1}_{K,g,f} H_n^{-1} \left(\sum_{i=1}^{n} K_h(\mathbf{X}_i - \mathbf{x}_i)\varphi(\varepsilon_i)\right)(1, \mathbf{X}_i - \mathbf{x})^\top; \tag{43}$$

$$\tag{44}$$

φ is the piecewise derivative of ρ, and

$$\lambda(s) = \min\left\{\frac{2}{2+s}, \frac{6+2s}{8+4s}\right\}. \tag{45}$$

Note that under the i.i.d. case, the constant s, which controls the weak dependence, is 0.

LEMMA B.2 (Bickel and Wichura (1971): Tightness of processes on a multidimensional cube)**.** *If $\{X_n\}_{n=1}^{\infty}$ is a sequence in $D[0,1]^d$, $\mathrm{P}(X \in [0,1]^d) = 1$. For neighboring blocks B, C in $[0,1]^d$ (see Definition 1) constants $\lambda_1 + \lambda_2 > 1$, $\gamma_1 + \gamma_2 > 0$, $\{X_n\}_{n=1}^{\infty}$ is tight if*

$$\mathsf{E}[|X_n(B)|^{\gamma_1}|X_n(C)|^{\gamma_2}] \le \mu(B)^{\lambda_1}\mu(C)^{\lambda_2}, \tag{46}$$

where $\mu(\cdot)$ is a finite nonnegative measure on $[0,1]^d$ (for example, Lebesgue measure), where the increment of X_n on the block B is defined by

$$X_n(B) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} X_n\big(\boldsymbol{s} + \boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})\big).$$

LEMMA B.3 (Meerschaert, M. M., Wang, W. and Xiao, Y. (2013))**.** *Suppose that $Y = \{Y(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^d\}$ is a centered Gaussian random field with values in $\mathbb{R}$, and denote*

$$d(\boldsymbol{s},\boldsymbol{t}) \stackrel{\text{def}}{=} d_Y(\boldsymbol{s},\boldsymbol{t}) = \big(\mathsf{E}|Y(\boldsymbol{t}) - Y(\boldsymbol{s})|^2\big)^{1/2}, \quad \boldsymbol{s},\boldsymbol{t} \in \mathbb{R}^d.$$

Let $\mathcal{D}$ be a compact set contained in a cube with length r in $\mathbb{R}^d$ and let $\sigma^2 = \sup_{\boldsymbol{t}\in\mathcal{D}} \mathsf{E}[Y(\boldsymbol{t})^2]$. For any $m > 0$, $\epsilon > 0$, define

$$\gamma(\epsilon) = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}, \|\boldsymbol{s}-\boldsymbol{t}\|\le\epsilon} d(\boldsymbol{s},\boldsymbol{t})$$

and

$$Q(m) = (2+\sqrt{2})\int_1^{\infty} \gamma(m2^{-y^2})dy.$$

Then for all $a > 0$ which satisfy $a \ge (1 + 4d\log 2)^{1/2}(\sigma + a^{-1})$,

$$\mathrm{P}\left\{\sup_{\boldsymbol{t}\in S}|Y(\boldsymbol{t})| > a\right\} \le 2^{2d+2}\left(\frac{r}{Q^{-1}(1/a)} + 1\right)^d \frac{\sigma + a^{-1}}{a} \exp\left\{-\frac{a^2}{2(\sigma + a^{-1})^2}\right\}, \tag{47}$$

where $Q^{-1}(a) = \sup\{m : Q(m) \le a\}$.

References

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multivariate stochastic processes and some applications, *The Annals of Mathematical Statistics* **42**(5): 1656–1670.

Dedecker, J., Merlevéde, F. and Rio, E. (2014). Strong approximation of the empirical distribution function for absolutely regular sequences in $\mathbb{R}^d$, *Electronic Journal of Probability* **19**(9): 1–56.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory* **24**(3): 726–748.

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**(5): 1529–1564.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton university press, New Jersey.

Meerschaert, M. M., Wang, W. and Xiao, Y. (2013). Fernique-type inequalities and moduli of continuity for anisotropic Gaussian random fields, *Transations of the American Mathematical Society* **365**(2): 1081–1107.

Muhsal, B. and Neumeyer, N. (2010). A note on residual-based empirical likelihood kernel density estimator, *Electronic Journal of Statistics* **4**: 1386–1401.

Owen, A. B. (2005). *Multidimensional variation for quasi-Monte Carlo*, Vol. 2 of *Ser. Biostat.*, World Sci. Publi., Hackensack, NJ., pp. 49–74.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates, *The Annals of Probability* **4**(6): 1009–1015.