UNIVERSITÄT HAMBURG

Estimating the conditional error distribution in nonparametric regression

Sebastian Kiwitt and Natalie Neumeyer

Preprint–No. 2010-07 **Dezember 2010**

DEPARTMENT MATHEMATIK

SCHWERPUNKT MATHEMATISCHE STATISTIK
UND STOCHASTISCHE PROZESSE

BUNDESSTR. 55, D – 20146 HAMBURG

Estimating the conditional error distribution in nonparametric regression

Sebastian Kiwitt and Natalie Neumeyer*
University of Hamburg, Department of Mathematics
Bundesstrasse 55, 20146 Hamburg, Germany

December 10, 2010

Abstract

We consider the general nonparametric regression model $Y = m(X) + \varepsilon$, where the distribution of the error ε, given the covariate $X = x$, is modelled by a conditional distribution function $P(\varepsilon \leq y \mid X = x) = F_{\varepsilon|X}(y|x)$. For estimation of $F_{\varepsilon|X}$ a kernel approach as well as the (kernel based) empirical likelihood method are discussed. The latter method allows for incorporation of the centeredness assumption $E[\varepsilon \mid X = x] = 0$ and other assumptions on the error distribution into the estimation. We show weak convergence of the corresponding empirical processes to Gaussian processes. Further, we discuss possible application of the results to modified residual bootstrap in the general regression model and to hypotheses tests, e. g. tests for parametric structure of the conditional variance function $\sigma^2(x) = E[\varepsilon^2 \mid X = x]$.

AMS 2010 Classification: Primary 62G08, Secondary 62G30, 62G05, 62G09, 62G10

Keywords and Phrases: bootstrap, empirical distribution function, empirical likelihood, heteroscedasticity, hypothesis testing, kernel estimation

1 Introduction

Assume a sample of independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a bivariate random variable (X, Y) is collected, where one is interested in modelling the functional dependence of the observation Y on the covariable X by the regression function $m(x) = E[Y \mid X = x]$. Having estimated the regression function, one is interested in the remaining dependence of Y and X, namely the distribution of Y around its conditional mean, which is given by the conditional

[*] corresponding author, e-mail: neumeyer@math.uni-hamburg.de

distribution of the centered observation $Y - m(X)$, usually modelled as measurement error ε [cf. Efromovich (2005, 2007), for instance]. To this end consider the nonparametric regression model

$$Y \;=\; m(X) + \varepsilon \tag{1.1}$$

and let $\varepsilon_i = Y_i - m(X_i)$, $i = 1, \dots, n$. Here no structural assumptions on the regression function m are made except for smoothness, and the error ε is not observable.

Sometimes for statistical inference homoscedasticity is assumed in model (1.1), i. e. independence of the conditional variance $\mathrm{Var}(Y \mid X)$ of the covariable X. Often homoscedasticity is even modelled more restrictively by assuming independence of covariable and error, and this is what we will mean by "homoscedastic model" throughout the paper:

$$Y \;=\; m(X) + \varepsilon, \quad \text{where } X \perp \varepsilon, \quad E[\varepsilon] = 0. \tag{1.2}$$

Techniques of proofs for instance for asymptotic distributions of procedures in mathematical statistics are made easier by the independence assumption. The conditional error distribution, given the covariate, as mentioned before is then the (unconditional) distribution of ε, and can be estimated by the empirical distribution of residuals $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$, $i = 1, \dots, n$, where $\hat{m}$ denotes the estimation of the regression function. See Cheng (2002) for strong consistency and Müller, Schick & Wefelmeyer (2007, 2009) for central limit theorems. However, in many applications such a model is too restrictive, and a number of tests for constant variance (i. e. $\mathrm{Var}(Y \mid X)$ independent of X) are available in the literature, see for instance Dette & Munk (1998), Dette (2002), and Liero (2003). Tests for independence of ε and X [i. e. for validity of the homoscedastic model (1.2)] are given by Einmahl & Van Keilegom (2008a) and Neumeyer (2009).

If homoscedasticity is rejected, heteroscedastic regression is often modelled in the form

$$Y \;=\; m(X) + \varepsilon, \quad \varepsilon = \sigma(X)\eta, \quad \text{where } X \perp \eta, \quad E[\eta] = 0, \quad E[\eta^2] = 1. \tag{1.3}$$

Here again, techniques of proofs usually make extensive use of the independence of X and η. Akritas & Van Keilegom (2001) consider the estimation of the (unconditional) distribution of η by the empirical distribution function of standardized residuals $\hat{\eta}_i = (Y_i - \hat{m}(X_i))/\hat{\sigma}(X_i)$, where $\hat{\sigma}^2$ denotes some nonparametric estimator of the conditional variance function $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$. In some applications it might be questionable whether the independence assumption is met and Einmahl & Van Keilegom (2008b) and Neumeyer (2009) have proposed tests for the hypothesis $X \perp \eta$.

If the hypothesis is rejected, a more general model should be used, where the distribution of the error ε, given the covariate $X = x$, is modelled by a conditional distribution function $F_{\varepsilon|X}(\cdot|x)$, which is known to be centered (by definition of m). This is the model we will consider in the paper at hand, namely

$$Y \;=\; m(X) + \varepsilon, \quad \text{where } E[\varepsilon \mid X = x] \;=\; \int y F_{\varepsilon|X}(dy|x) \;=\; 0. \tag{1.4}$$

For example, Zhu, Fujikoshi & Naito (2001), Dette & Hetzler (2009a, 2009b), and Koul & Song (2010) consider such a model in the context of lack-of-fit tests for the variance function. To the authors' knowlegde the estimation of the conditional error distribution in model (1.4) has not yet been considered in the literature. It will be of major importance in the derivation of bootstrap procedures in model (1.4), and further can be applied for model tests in the context of model (1.4).

In section 2 we suggest a simple residual-based kernel estimator $\hat{F}_n(\cdot|x)$ for $F_{\varepsilon|X}(\cdot|x)$, we give an asymptotic expansion, and show weak convergence of a bias-corrected version of the process $(nh)^{1/2}(\hat{F}_n(\cdot|x) - F_{\varepsilon|X}(\cdot|x))$ to a centered Gaussian process, where h denotes some smoothing parameter.

In section 3 we suggest a modified kernel estimator $\hat{F}_n^*(\cdot|x)$ for $F_{\varepsilon|X}(\cdot|x)$ based on the empirical likelihood method as introduced by Owen (1988, 2001) [see also DiCiccio, Hall & Romano (1991), Kitamura (1997), Molanes Lopez, Van Keilegom & Veraverbeke (2009), among many others]. This method enables us to incorporate the centeredness assumption as in (1.4) into the estimation, as well as other information given by a condition

$$\int g(y,x)F_{\varepsilon|X}(dy|x) \;=\; 0 \tag{1.5}$$

for some known function g. We will show an asymptotic expansion of $\hat{F}_n^*(\cdot|x)$ and discuss weak convergence of a bias-corrected version of the process $(nh)^{1/2}(\hat{F}_n^*(\cdot|x) - F_{\varepsilon|X}(\cdot|x))$ to a centered Gaussian process. This generalizes results by Kiwitt, Nagel & Neumeyer (2008) who considered estimation of the distributions of ε in model (1.2) and η in model (1.3), respectively, by the empirical likelihood method.

We investigate whether the estimation of the conditional error distribution $F_{\varepsilon|X}(\cdot|x)$ can be improved by incorporating the additional information (1.5) [in comparison to the simple kernel estimator $F_n(\cdot|x)$]. To this end, in section 4 we consider examples in asymptotic theory as well as by means of a simulation study. We demonstrate the good performance of both estimators and show examples where bias as well as variance of the estimation can be reduced considerably by the empirical likelihood method.

In section 5 possible applications are discussed. First, an important motivation for application of the empirical likelihood estimation is for constructing bootstrap versions of statistical procedures in the general model (1.4). In the homoscedastic model (1.2) usually residual bootstrap as introduced by Härdle & Bowman (1988) can be applied, whereas in the heteroscedastic model (1.3) to preserve the conditional variance often the wild bootstrap is used, see Härdle & Mammen (1993), Härdle & Marron (1991), and Stute, González Manteiga & Presedo Quindimil (1998), among others. However, in the general model considered here, both methods usually cannot be applied directly and need to be modified. We will discuss how this can be done and how the estimators $\hat{F}_n$ and $\hat{F}_n^*$ can be useful. We particularly consider the bootstrap versions of tests for the hypothesis of a parametric structure of the variance function, $H_0 : \exists\vartheta \in \Theta$ s. t. $\sigma^2(\cdot) = \sigma^2_\vartheta(\cdot)$, where the function $\sigma^2_\vartheta(\cdot)$ is known except

for some finite dimensional parameter ϑ. Dette, Neumeyer & Van Keilegom (2007) consider a test for such hypothesis in the heteroscedastic model (1.3) and Zhu, Fujikoshi & Naito (2001), Dette & Hetzler (2009a, 2009b), and Koul & Song (2010) in model (1.4). Tests for constant variance (independence of $\mathrm{Var}(Y \mid X)$ of X, often called homoscedasticity), as mentioned earlier, are a special case.
Secondly, hypotheses tests for validity of the information (1.5) are discussed briefly, which are based on a measure of distance between the simple kernel estimator and the empirical likelihood estimator for the conditional error distribution function. A special case are again goodness-of-fit tests for the variance function.
The last subsection of section 5 concludes the paper.

The technical assumptions and all proofs are given in an appendix.

2 Kernel based estimation

As the errors $\varepsilon_1, ..., \varepsilon_n$ in model (1.4) are not observable, they need to be estimated by residuals

$$\hat{\varepsilon}_i \;=\; Y_i - \hat{m}(X_i), \quad i = 1, \ldots, n.$$

To this end let $\tilde{K}$ be a kernel function and $b = b_n$ a sequence of bandwidths, and let $\hat{m}$ denote the Nadaraya-Watson estimator for the regression function m [Nadaraya (1964), Watson (1964)] defined as

$$\hat{m}(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{b}\tilde{K}\Big(\frac{X_i - x}{b}\Big) Y_i \frac{1}{\hat{f}_X(x)},$$

where $\hat{f}_X$ denotes the kernel estimator for the covariate density f_X, i.e.

$$\hat{f}_X(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{b}\tilde{K}\Big(\frac{X_i - x}{b}\Big).$$

The residuals are used to estimate the conditional distribution function $F_{\varepsilon|X}$ also by kernel approach. Let K denote a kernel function and $h = h_n$ a sequence of bandwidths and let the estimator be defined as

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} \frac{K(\frac{X_i - x}{h})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h})} I\{\hat{\varepsilon}_i \le y\}. \tag{2.1}$$

We further define the analogous estimator built from the errors $\varepsilon_1, ..., \varepsilon_n$ as

$$F_n(y|x) = \sum_{i=1}^{n} \frac{K(\frac{X_i - x}{h})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h})} I\{\varepsilon_i \le y\}. \tag{2.2}$$

This function is not known, but will be used for the proof of the asymptotic expansion given in the theorem below.

Kernel based estimators such as $F_n(\cdot|x)$ for conditional distributions $F_{\epsilon|X}(\cdot|x)$ based on iid samples $(X_i, \varepsilon_i), i = 1, ..., n$, have been considered by Stute (1986), Horwáth & Yandell (1988), and Hall, Wolff & Yao (1999), among others. Note that the approach (2.2) could be used in our case to estimate the conditional distribution of the observations Y, given the covariate $X = x$, i. e. $F_{Y|X}(\cdot|x)$, by $\hat{G}_n(\cdot|x)$ based on the iid-sample $(X_i, Y_i), i = 1, ..., n$. Due to the equality $F_{\varepsilon|X}(y|x) = F_{Y|X}(y + m(x)|x)$ an alternative to the estimator $\hat{F}_n(\cdot|x)$ is given by $\hat{G}_n(\cdot + \hat{m}(x)|x)$. Both estimators have very similar asymptotic properties and we will only consider $\hat{F}_n$ in the following.

Technical assumptions are listed in the appendix for reasons of clarity and better readability. Note that we assume $\lim_{n\to\infty} \frac{b_n}{h_n} = \lambda \in [0, 1)$ and smoothness of m and f_X.

Theorem 2.1 *(a) Under model (1.4) and the assumptions listed in section A we have the asymptotic expansion*

$$\begin{aligned}\hat{F}_n(y|x) &= F_{\varepsilon|X}(y|x) + \frac{1}{f_X(x)} \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)(I\{\varepsilon_i \le y\} - F_{\varepsilon|X}(y|x)) \\ &\quad + \frac{1}{f_X(x)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \int \frac{1}{h} K\Big(\frac{t - x}{h}\Big) f_{\varepsilon|X}(y|t) \frac{1}{b} \tilde{K}\Big(\frac{X_i - t}{b}\Big)\, dt + b^2 B(y|x) \\ &\quad + o_p\Big(\frac{1}{\sqrt{nh}}\Big)\end{aligned}$$

uniformly with respect to $y \in \mathbb{R}$ for each $x \in (0, 1)$, where

$$B(y|x) = \frac{1}{2} \frac{f_{\varepsilon|X}(y|x)}{f_X(x)} ((mf_X)''(x) - (mf_X'')(x)) \int \tilde{K}(u) u^2\, du.$$

(b) Under model (1.4) and the assumptions listed in section A the process

$$G_n(y|x) = \sqrt{nh}\Big(\hat{F}_n(y|x) - F_{\varepsilon|X}(y|x) - h^2 H(y|x) - b^2 B(y|x)\Big), \quad y \in \mathbb{R},$$

converges weakly to a centered Gaussian process $G(y|x)$, $y \in \mathbb{R}$, with covariance structure

$$\begin{aligned}&\operatorname{Cov}(G(y|x), G(z|x)) \\ &= \frac{1}{f_X(x)} \Big(F_{\varepsilon|X}(y \wedge z|x) - F_{\varepsilon|X}(y|x) F_{\varepsilon|X}(z|x)\Big) \int K^2(u)\, du \\ &\quad + \frac{1}{f_X(x)} E[\varepsilon^2 \mid X = x] f_{\varepsilon|X}(y|x) f_{\varepsilon|X}(z|x) k_2(\lambda) \\ &\quad + \frac{1}{f_X(x)} \Big(E[\varepsilon I\{\varepsilon \le y\} \mid X = x] f_{\varepsilon|X}(z|x) + E[\varepsilon I\{\varepsilon \le z\} \mid X = x] f_{\varepsilon|X}(y|x)\Big) k_1(\lambda).\end{aligned}$$

Here, $k_1(0) = k_2(0) = \int K^2(u)\, du$ and for $\lambda > 0$,

$$k_1(\lambda) = \int (K * \tilde{K}_\lambda)(u) K(u)\, du, \quad k_2(\lambda) = \int (K * \tilde{K}_\lambda)^2(u)\, du,$$

where $*$ denotes convolution and $\tilde{K}_\lambda(\cdot) = \frac{1}{\lambda}\tilde{K}(\frac{\cdot}{\lambda})$. Further, $B(y|x)$ is defined in (a), and

$$H(y|x) = \frac{1}{2f_X(x)}\Big(\frac{\partial^2(F_{\varepsilon|X}(y|t)f_X(t))}{\partial t^2}\Big|_{t=x} - F_{\varepsilon|X}(y|x)f_X''(x)\Big)\int K(u)u^2\,du.$$

The proof of Theorem 2.1 is given in the appendix.

Remark 2.2 Note that for $\lambda = 0$ (i. e. $b = o(h)$) the bias term of order b^2 is $o((nh)^{-1/2})$ and hence negligible. Further in this case due to $\int K^2(u)\,du = k_2(0) = k_1(0)$ the covariance does no longer depend on the kernel $\tilde{K}$ used for estimating the regression function.

Remark 2.3 From the proof of Theorem 2.1 it follows for the estimator based on the true errors [see (2.2)] that

$$F_n(y|x) \;=\; F_{\varepsilon|X}(y|x) + \frac{1}{f_X(x)}\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)(I\{\varepsilon_i \le y\} - F_{\varepsilon|X}(y|x)) + o_p(\frac{1}{\sqrt{nh}})$$

and the process $\sqrt{nh}(F_n(y|x) - F_{\varepsilon|X}(y|x)) - h^2H(y|x))$, $y \in \mathbb{R}$, weakly converges to a centered Gaussian process $\tilde{G}(y|x)$, $y \in \mathbb{R}$, with covariance structure

$$\mathrm{Cov}(\tilde{G}(y|x), \tilde{G}(z|x)) \;=\; \frac{1}{f_X(x)}\Big(F_{\varepsilon|X}(y \wedge z|x) - F_{\varepsilon|X}(y|x)F_{\varepsilon|X}(z|x)\Big)\int K^2(u)\,du.$$

Hence, we have shown that the errors $\varepsilon_1, \ldots, \varepsilon_n$ cannot be replaced by estimators $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ without changing the asymptotic distribution.

3 Empirical likelihood estimation

An empirical likelihood method distributes possibly different weights p_i to the ith observation, $i = 1, \ldots, n$. In our context this means to consider the following class of kernel estimators for the conditional distribution of ε, given $X = x$,

$$G_{\varepsilon|X}(y|x) \;=\; \frac{1}{h}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)p_i I\{\hat{\varepsilon}_i \le y\}\frac{1}{\tilde{f}_X(x)} \tag{3.1}$$

for some density K and positive bandwidth h, where the denominator

$$\tilde{f}_X(x) \;=\; \frac{1}{h}\sum_{j=1}^n K\Big(\frac{X_j - x}{h}\Big)p_j$$

assures that $G_{\varepsilon|X}(\cdot|x)$ is a distribution function. Under the conditions

$$p_i > 0\,\forall i = 1, .., n, \; \sum_{i=1}^n p_i = 1 \tag{3.2}$$

$\tilde{f}_X$ can be interpreted as an asymptotically unbiased estimator for the covariate density f_X. The weights are chosen such that the product $\prod_{i=1}^n p_i$ is maximized while fulfilling (3.2). This maximization problem (without further constraints) gives $p_i = \frac{1}{n}$, $i = 1, \ldots, n$, and thus $G_{\varepsilon|X}(\cdot|x)$ reduces to the simple kernel estimator $\hat{F}_n(\cdot|x)$ as considered in section 2. However, now we also take into account the condition

$$\sum_{i=1}^n p_i \, g(\hat{\varepsilon}_i, x) \, K\Big(\frac{X_i - x}{h}\Big) = 0 \tag{3.3}$$

which, by definition of the estimator $G_{\varepsilon|X}$ in (3.1), is in fact the empirical version $\int g(y,x)$ $G_{\varepsilon|X}(dy|x) = 0$ of our additional information (1.5). Maximization of the likelihood $\prod_{i=1}^n p_i$ means to minimize the Kullback-Leibler distance of the estimator $G_{\varepsilon|X}$ to the standard estimator $\hat{F}_n$, compare to Hall, Wolff & Yao (1999). See also Einmahl & McKeague (2003) or Antoine, Bonnal & Renault (2007) for empirical likelihood methods with conditional additional information.

For the observed sample we assume that

$$\min_{1\le i\le n} g_j(\hat{\varepsilon}_i, x) K\Big(\frac{X_i - x}{h}\Big) < 0 < \max_{1\le i\le n} g_j(\hat{\varepsilon}_i, x) K\Big(\frac{X_i - x}{h}\Big)$$

for all $j \in \{1, \ldots, k\}$ and the matrix $(nh)^{-1} \sum\limits_{i=1}^n K^2((X_i - x)/h) g(\hat{\varepsilon}_i, x) g^\top(\hat{\varepsilon}_i, x)$ to be positive definite. Those assumptions are needed for the existence of the unique solution of the empirical likelihood maximization. One obtains analogously to Qin and Lawless (1994) and Kiwitt, Nagel & Neumeyer (2008) the solution

$$p_i = p_i(x) = \frac{1}{n} \frac{1}{1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})}$$

of the maximization of $\prod_{i=1}^n p_i$ under (3.2) and (3.3). Here $\hat{\eta}_n(x)$ is the solution of

$$\sum_{i=1}^n p_i(x) \, g(\hat{\varepsilon}_i, x) \, K\Big(\frac{X_i - x}{h}\Big) = \sum_{i=1}^n \frac{g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})}{1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})} \; = \; 0$$

such that $\min_{i=1,\ldots,n}(1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})) > 1/n$. Inserting the weights into the definition of the estimator $G_{\varepsilon|X}$ finally gives our empirical likelihood estimator $\hat{F}_n^*$ for $F_{\varepsilon|X}$,

$$\hat{F}_n^*(y|x) = \sum_{i=1}^n \frac{\frac{1}{h} K(\frac{X_i - x}{h}) p_i(x)}{\frac{1}{h} \sum_{j=1}^n K(\frac{X_i - x}{h}) p_j(x)} I\{\hat{\varepsilon}_i \le y\}.$$

We will investigate whether the incorporation of the additional information (1.5) improves the estimate in comparison to the simple kernel estimator $\hat{F}_n$ considered in section 2. Another main motivation for the consideration of the empirical likelihood estimator is the bootstrap data generation for model (1.4). The detailed explanations are postponed to section 5.

Theorem 3.1 *(a) Under model (1.4) and the assumptions listed in this section and in section A we have the stochastic expansion*

$$\begin{aligned}\hat{F}_n^*(y|x) &= \hat{F}_n(y|x) - \frac{1}{f_X(x)}\Big(\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) g(\varepsilon_i, x)^\top \\ &\qquad - \frac{1}{n}\sum_{i=1}^n \varepsilon_i \int \frac{1}{h}K\Big(\frac{t-x}{h}\Big) g'(z,x)^\top \frac{1}{b}\tilde{K}\Big(\frac{X_i - t}{b}\Big) f_{\varepsilon|X}(z|t)\, d(z,t) - b^2\tilde{B}(x)^\top\Big) \\ &\qquad \times \Sigma^*(x)^{-1} E[g(\varepsilon, X) I\{\varepsilon \le y\} \mid X = x] + o_p(\frac{1}{\sqrt{nh}})\end{aligned}$$

uniformly with respect to $y \in \mathbb{R}$ for each $x \in (0,1)$, where the expansion for $\hat{F}_n(\cdot|x)$ is given in Theorem 2.1 (a),

$$\tilde{B}(x) = \frac{1}{2f_X(x)}((mf_X)''(x) - (mf_X'')(x)) \int \tilde{K}(u)u^2\, du E[g'(\varepsilon, x) \mid X = x],$$

$g'(y,x) = \frac{\partial g(y,x)}{\partial y}$ *and* $\Sigma^*(x) = E[g(\varepsilon, X) g^\top(\varepsilon, X) \mid X = x]$.
(b) Under model (1.4) and the assumptions listed in this section and in section A the process

$$G_n^*(y|x) = \sqrt{nh}\Big(\hat{F}_n^*(y|x) - F_{\varepsilon|X}(y|x) - h^2 H^*(y|x) - b^2 B^*(y|x)\Big)\ , \quad y \in \mathbb{R},$$

converges weakly to a centered Gaussian process $G^*(y|x)$, $y \in \mathbb{R}$, with covariance structure

$$\begin{aligned}\mathrm{Cov}(G^*(y|x), G^*(z|x)) &= \mathrm{Cov}(G(y|x), G(z|x)) + \frac{1}{f_X(x)} A(y|x)^\top S(x) A(z|x) \\ &\quad + \frac{1}{f_X(x)} R(y|x)^\top A(z|x) + \frac{1}{f_X(x)} R(z|x)^\top A(y|x),\end{aligned}$$

where

$$\begin{aligned}A(y|x) &= \Sigma^*(x)^{-1} E[g(\varepsilon, X) I\{\varepsilon \le y\} \mid X = x] \\ S(x) &= \Sigma^*(x) \int K^2(u)\, du + E[\varepsilon^2 \mid X = x] E[g'(\varepsilon, x) \mid X = x] E[g'(\varepsilon, x) \mid X = x]^\top k_2(\lambda) \\ &\quad - 2E[g(\varepsilon, x)\varepsilon \mid X = x] E[g'(\varepsilon, x) \mid X = x]^\top k_1(\lambda) \\ R(y|x) &= -E[I\{\varepsilon \le y\} g(\varepsilon, x) \mid X = x] \int K^2(u)\, du \\ &\quad + \Big(E[I\{\varepsilon \le y\}\varepsilon \mid X = x] E[g'(\varepsilon, x) \mid X = x] - E[\varepsilon g(\varepsilon, x) \mid X = x] f_{\varepsilon|X}(y|x)\Big) k_1(\lambda) \\ &\quad + E[\varepsilon^2 \mid X = x] E[g'(\varepsilon, x) \mid X = x] f_{\varepsilon|X}(y|x) k_2(\lambda)\end{aligned}$$

and for the bias constants we have

$$\begin{aligned}H^*(y|x) &= H(y|x) - \frac{1}{2f_X(x)} \int K(u)u^2\, du \frac{\partial^2 (E[g(\varepsilon, x)^\top \mid X = t] f_X(t))}{\partial t^2}\Big|_{t=x} \Sigma^*(x)^{-1} \\ &\qquad \times E[g(\varepsilon, X) I\{\varepsilon \le y\} \mid X = x] \\ B^*(y|x) &= B(y|x) + \tilde{B}(x)^\top \Sigma^*(x)^{-1} E[g(\varepsilon, X) I\{\varepsilon \le y\} \mid X = x].\end{aligned}$$

Here $\tilde{B}(x)$ and $\Sigma^*(x)$ are defined in (a), and the process G as well as $H(\cdot|x)$, $B(\cdot|x)$ and k_1, k_2 are defined in Theorem 2.1.

The proof of Theorem 3.1 is given in the appendix.

Remark 3.2 Similar to Remark 2.3 consider the case where the estimation is based on an iid-sample $\varepsilon_1, \ldots, \varepsilon_n$, and let F_n^* denote the corresponding empirical likelihood estimator. Then it follows from the proof of Theorem 3.1 that the process

$$\sqrt{nh}\Big(F_n^*(y|x) - F_{\varepsilon|X}(y|x) - h^2 H^*(y|x) - b^2 B(y|x)\Big)\ , \quad y \in \mathbb{R},$$

converges weakly to a centered Gaussian process with a variance

$$\mathrm{Var}(\tilde{G}(y|x)) - \frac{1}{f_X(x)}\Sigma^*(x)^{-1}(E[I\{\varepsilon \le y\}\varepsilon \mid X = x])^2 \int K(u)u^2\,du$$

that is uniformly smaller than the variance $\mathrm{Var}(\tilde{G}(y|x))$ as given in Remark 2.3 resulting from the standard estimator.

Remark 3.3 We assumed smoothness of the function g defining the additional information (1.5). However, results similar to Theorem 3.1 can be obtained for indicator functions like $g(\varepsilon, x) = I\{\varepsilon \le a(x)\} - b(x)$ for incorporation of information on the conditional quantiles [compare to Kiwitt, Nagel & Neumeyer (2008)].

Remark 3.4 The additional information (1.5) can include unknown finite-dimensional parameters in the form that

$$\exists\vartheta \text{ such that } E[g_\vartheta(\varepsilon, x) \mid X = x] = 0$$

[see also Qin & Lawless (1994) who consider estimation of an (unconditional) distribution function based on iid data where the (unconditional) additional information includes some unknown parameter which is also estimated by the empirical likelihood method]. In the typical case where ϑ can be estimated by a $\sqrt{n}$-consistent estimator $\hat{\vartheta}$ one uses the empirical likelihood estimation as explained before where in (3.3) the function g is replaced by $g_{\hat{\vartheta}}$. Theorem 3.1 remains valid with g replaced by g_ϑ, where ϑ denotes the "true" parameter (because the convergence rate of $\hat{\vartheta}$ to the true ϑ is faster than $\sqrt{nh}$).
Consider for example an assumed linear structure of the variance function $\mathrm{Var}(\varepsilon \mid X = x) = \sigma_\vartheta^2(x) = \vartheta_0 + \vartheta_1 x$ by defining $g_\vartheta(\varepsilon, x) = \varepsilon^2 - \sigma_\vartheta^2(x)$. The parameter $\vartheta = (\vartheta_0, \vartheta_1)^\top$ can be estimated by least-squares approach in a linear model of observations $Z = (\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)^\top$ with design matrix

$$D = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}.$$

We obtain the estimator

$$\begin{aligned}\hat{\vartheta} \;&=\; (D^\top D)^{-1} D^\top Z \;=\; \vartheta + \begin{pmatrix} 1 & \overline{X}_n \\ \overline{X}_n & \overline{X_n^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (\hat{\varepsilon}_i^2 - E[\varepsilon_i^2 \mid X_i]) \\ \frac{1}{n}\sum_{i=1}^n X_i(\hat{\varepsilon}_i^2 - E[\varepsilon_i^2 \mid X_i]) \end{pmatrix} \\ &= \vartheta + O_p(\frac{1}{\sqrt{n}}),\end{aligned}$$

where the rate can be shown analogously to Lemma B.2(ii) by Kiwitt, Nagel & Neumeyer (2008) [see also Müller, Schick & Wefelmeyer (2004)].

4 Comparison in asymptotic theory and simulations

In examples we will compare both considered estimators $\hat{F}_n(y|x)$ and $\hat{F}_n^*(y|x)$ in terms of the asymptotic biases and variances to investigate whether incorporating the additional information by the empirical likelihood yields an improved estimator.

First we consider the simple residual-based kernel estimator $\hat{F}_n$. Here for the asymptotic variance we have from Theorem 2.1 that

$$\begin{aligned}\mathrm{Var}(G(y|x)) \;=\; & \frac{1}{f_X(x)}\Big(F_{\varepsilon|X}(y|x) - F_{\varepsilon|X}^2(y|x)\Big)\int K^2(u)\,du + \frac{1}{f_X(x)}\sigma^2(x) f_{\varepsilon|X}^2(y|x) k_2(\lambda) \\ & + \frac{2}{f_X(x)} U_1(y|x) f_{\varepsilon|X}(y|x) k_1(\lambda) \qquad (4.1)\end{aligned}$$

with the notation $U_1(y|x) = E[\varepsilon I\{\varepsilon \le y\} \mid X = x]$ and $\sigma^2(x) = E[\varepsilon^2 \mid X = x]$. For example with uniformly distributed covariates $X \sim U[0,1]$ and normally distributed errors $\varepsilon \sim N(0, \sigma^2(x))$ one obtains

$$\mathrm{Var}(G(y|x)) \;=\; \Phi\Big(\frac{y}{\sigma(x)}\Big)\Phi\Big(\frac{-y}{\sigma(x)}\Big)\int K^2(u)\,du + \varphi^2\Big(\frac{y}{\sigma(x)}\Big)(k_2(\lambda) - 2k_1(\lambda))$$

because $U_1(y|x) = -\sigma^2(x) f_{\varepsilon|X}(y|x)$. Here Φ and φ denote the standard normal distribution and density function, respectively. A straightforward calculation shows that the integrated variance is equal to

$$\int \mathrm{Var}(G(y|x))\,dy \;=\; \frac{\sigma(x)}{2\sqrt{\pi}}\Big(\int K^2 + \int (K - K * K)^2\Big). \qquad (4.2)$$

Under those assumptions with $K = \tilde{K}$ the bias terms from Theorem 2.1 reduce to

$$B(y|x) = \frac{1}{2\sigma(x)}\varphi\Big(\frac{y}{\sigma(x)}\Big) m''(x) \int K(u) u^2\,du.$$

and

$$H(y|x) = \frac{1}{2}\Big(\frac{1}{\sigma^2(x)}\varphi'\Big(\frac{y}{\sigma(x)}\Big) - \frac{\sigma'(x)}{\sigma^2(x)}\varphi\Big(\frac{y}{\sigma(x)}\Big)\Big)\int K(u) u^2\,du.$$

The integrated squared bias is

$$\int (h^2 H(y|x) + b^2 B(y|x))^2 \, dy = \frac{(\int K(u)u^2 \, du)^2}{8\sqrt{\pi}\sigma(x)} \Big(\Big(b^2 m''(x) - h^2 \frac{\sigma'(x)}{\sigma(x)} \Big)^2 + \frac{h^4}{2\sigma^2(x)} \Big). \quad (4.3)$$

Example 1. Let $g(\varepsilon, x) = \varepsilon$, i. e. we explicitly incorporate the information that the errors are centered into the estimation. This is no additional information, but given by the model definition (1.4). Although the estimator $\hat{m}$ implicitly already uses the information of centered errors, the explicit use of this information can improve both the asymptotic variance and the bias; the latter especially in cases where the regression estimation bias is large. For the asymptotic variance of the empirical likelihood estimator we obtain with above notations that

$$\operatorname{Var}(G^*(y|x)) = \operatorname{Var}(G(y|x)) + \frac{U_1^2(y|x)}{f_X(x)\sigma^2(x)}(k_2(\lambda) - k_2(0)) + 2\frac{U_1(y|x) f_\varepsilon(y|x)}{f_X(x)}(k_2(\lambda) - k_1(\lambda)),$$

where $\operatorname{Var}(G(y|x))$ is given in (4.1) and both variances are equal for $\lambda = 0$. Further $k_2(\lambda) - k_2(0) \leq 0$ and $U_1(y|x) \leq 0$. E. g., for normal distribution we always have an improvement as then

$$\operatorname{Var}(G^*(y|x)) = \operatorname{Var}(G(y|x)) - \frac{1}{f_X(x)} \varphi^2 \Big(\frac{y}{\sigma(x)} \Big) \int (K(u) - (K * \tilde{K}_\lambda)(u))^2 \, du.$$

In that example with uniformly distributed covariates integrating with respect to y one obtains as overall variance improvement

$$\int (\operatorname{Var}(G^*(y|x)) - \operatorname{Var}(G(y|x))) \, dy = -\frac{\sigma(x)}{2\sqrt{\pi}} \int (K(u) - (K * \tilde{K}_\lambda)(u))^2 \, du,$$

where $\int \operatorname{Var}(G(y|x)) \, dy$ is given in (4.2).

We now consider the bias of order b^2 that is due to the estimation of the regression function m. For the constant defined in Theorem 3.1 we have here

$$B^*(y|x) \;=\; \frac{1}{2f_X(x)}((mf_X)''(x) - (mf_X'')(x)) \int \tilde{K}(u)u^2 \, du \Big(f_{\varepsilon|X}(y|x) + \frac{U_1(y|x)}{\sigma^2(x)} \Big).$$

Especially for normally distributed errors here one obtains that $B^*(y|x) = 0$ and hence that by the empirical likelihood method the bias of order b^2 cancels completely. For the bias that stems from the kernel approach for estimating the conditional distribution we have $H^*(y|x) = H(y|x)$ which follows directly from the definition in Theorem 3.1 for $g(\varepsilon, x) = \varepsilon$. Hence the empirical likelihood method in that case does not change the bias of order h^2. A calculation of the integrated squared bias yields, e. g., for uniformly distributed covariates, normally distributed errors and $K = \tilde{K}$ that

$$\int (h^2 H^*(y|x) + b^2 B^*(y|x))^2 \, dy = h^4 \int H^2(y|x) \, dy = h^4 \frac{(\int K(u)u^2 \, du)^2}{8\sqrt{\pi}\sigma^3(x)} \Big(\frac{1}{2} + (\sigma'(x))^2 \Big)$$

and a comparison with (4.3) shows that the empirical likelihood method improves upon the bias iff $b^2(m''(x))^2 \geq 2h^2 m''(x)\sigma'(x)/\sigma(x)$ which is the case for constant variances $\sigma^2(x) \equiv \sigma^2$, for instance.

Example 2. Let $g(\varepsilon, x) = \varepsilon^2 - \sigma^2(x)$ for some known function $\sigma^2(\cdot)$. Due to $E[g'(\varepsilon, x) \mid X = x] = 0$ and $E[\varepsilon g(\varepsilon, x) \mid X = x] = 0$ for error distributions with vanishing third moment we have

$$\operatorname{Var}(G^*(y|x)) \;=\; \operatorname{Var}(G(y|x)) - \frac{\int K^2(u)\,du}{f_X(x)} \frac{U_2^2(y|x)}{\eta(x)}$$

with the notations $U_2(y|x) = E[(\varepsilon^2 - \sigma^2(x)) I\{\varepsilon \leq y\} \mid X = x]$ and $\eta(x) = E[(\varepsilon^2 - \sigma^2(x))^2 \mid X = x]$. The asymptotic variance of the empirical likelihood estimator is smaller than the asymptotic variance of the simple kernel estimator uniformly in y. For instance, for uniformly distributed covariates and normally distributed errors we obtain $\eta(x) = 2\sigma^4(x)$, $U_2(y|x) = -\sigma^2(x) y f_{\varepsilon|X}(y|x)$ and thus

$$\operatorname{Var}(G^*(y|x)) \;=\; \operatorname{Var}(G(y|x)) - \frac{\int K^2(u)\,du}{2\sigma^2(x)} y^2 \varphi^2\Big(\frac{y}{\sigma(x)}\Big)$$

with an overall improvement of

$$\int (\operatorname{Var}(G^*(y|x)) - \operatorname{Var}(G(y|x)))\,dy = -\frac{\sigma(x)}{8\sqrt{\pi}} \int K^2(u)\,du.$$

Note further that $\tilde{B}(x) = 0$ in Theorem 3.1 and hence $B^*(y|x) = B(y|x)$ and there is no change in the bias arising from estimation of the regression function. Furthermore we have

$$H^*(y|x) \;=\; H(y|x) - \frac{1}{2 f_X(x)} \int K(u) u^2\,du \Big((\sigma^2 f_X)''(x) - (\sigma^2 f_X'')(x)\Big) \frac{U_2(y|x)}{\eta(x)}.$$

For uniformly distributed covariates and normally distributed errors we have

$$H^*(y|x) \;=\; H(y|x) + \frac{1}{4} \int K(u) u^2\,du \frac{(\sigma^2)''(x)}{\sigma^3(x)} y \varphi\Big(\frac{y}{\sigma(x)}\Big)$$

and (with $K = \tilde{K}$)

$$\begin{aligned} &\int (h^2 H^*(y|x) + b^2 B^*(y|x))^2\,dy \\ &= \frac{(\int K(u) u^2\,du)^2}{8\sqrt{\pi}\sigma(x)} \bigg(\Big(b^2 m''(x) - h^2 \frac{\sigma'(x)}{\sigma(x)}\Big)^2 + \frac{h^4}{8\sigma^2(x)} \Big(((\sigma^2)''(x))^2 - 2\Big)^2\bigg). \end{aligned}$$

A comparison with (4.3) shows that the empirical likelihood method improves upon the bias iff $((\sigma^2)''(x))^2 \leq 4(\sigma^2)''(x)$. As in this example σ^2 is known, this condition can be checked. It yields, e. g. that no bias improvement can be obtained for x-values with $(\sigma^2)''(x) < 0$.

Example 3. As the information $E[\varepsilon \mid X = x] = 0$ is "for free" we consider $g(\varepsilon) = (\varepsilon, \varepsilon^2 - \sigma^2(x))^\top$ in this example to see whether compared to Example 2 further improvements can be achieved. The improvement of the variance cumulates as here (for distributions with vanishing conditional third moment) one obtains

$$\begin{aligned}\mathrm{Var}(G^*(y|x)) &= \mathrm{Var}(G(y|x)) - \frac{\int K^2(u)\,du}{f_X(x)}\frac{U_2^2(y|x)}{\eta(x)} \\ &\quad + \frac{U_1^2(y|x)}{f_X(x)\sigma^2(x)}(k_2(\lambda) - k_2(0)) + 2\frac{U_1(y|x)f_\varepsilon(y|x)}{f_X(x)}(k_2(\lambda) - k_1(\lambda)),\end{aligned}$$

which for uniform covariates and normal errors reduces to

$$\begin{aligned}\mathrm{Var}(G^*(y|x)) &= \mathrm{Var}(G(y|x)) - \frac{\int K^2(u)\,du}{2\sigma^2(x)} y^2\varphi^2\Big(\frac{y}{\sigma(x)}\Big) \\ &\quad - \varphi^2\Big(\frac{y}{\sigma(x)}\Big)\int (K(u) - (K * \tilde{K}_\lambda)(u))^2\,du\end{aligned}$$

with an overall improvement of

$$\int (\mathrm{Var}(G^*(y|x)) - \mathrm{Var}(G(y|x)))\,dy = -\frac{\sigma(x)}{8\sqrt{\pi}}\Big(\int K^2(u)\,du + 4\int (K(u) - (K * \tilde{K}_\lambda)(u))^2\,du\Big).$$

For the bias one obtains B^* as in Example 1 and H^* as in Example 2 and hence for uniform covariates and normal errors that

$$\begin{aligned}\int (h^2 H^*(y|x) + b^2 B^*(y|x))^2\,dy &= h^4 \int (H^*(y|x))^2\,dy \\ &= h^4 \frac{(\int K(u)u^2\,du)^2}{16\sqrt{\pi}\sigma^3(x)}\Big(\frac{1}{4}\Big((\sigma^2)''(x) + 2\Big)^2 + 2(\sigma'(x))^2\Big).\end{aligned}$$

FIGURES 1–3 HERE

In figures 1–3 the good performance of the empirical likelihood estimator is demonstrated. Figure 1 corresponds to example 1, figure 2 to example 2 and figure 3 to example 3. In all three figures the first and third row show the asymptotic integrated squared bias (left panel), integrated variance (middle panel) and integrated mean squared error (right panel) as functions in x for the estimators $\hat{F}_n$ (dashed line) and $\hat{F}_n^*$ (solid line). The second and fourth row show for a fixed $x = 0.5$ the squared bias (left), variance (middle) and mean squared error (right) as functions in y (graphics for other x-values were similar and are thus not depicted). We took X to be uniformly distributed, ε to be normally distributed and $m(x) = x^2$. In all three figures the first two rows depict results for the variance function $\sigma^2(x) = e^{-3x}$, whereas the last two rows show results for $\sigma^2(x) = (1 + 0.5x)^2$. We chose equal bandwidths $h = b = n^{-1/5}$ such that the squared bias factors $b^4 = h^4$ are equal to the variance factor $(nh)^{-1}$. Those factors are neglected to obtain results independent of n.

Although the bias of the empirical likelihood estimator is not always smaller than that of the simple kernel estimator it can be seen from the graphics that the empirical likelihood method can improve the overall performance of the conditional distribution estimator by far.

We also investigated the small sample performance of both conditional distribution estimators in simulations. To this end we considered the above setting and examples for sample size $n = 100$, bandwidths $h = b = n^{-1/5}$ and Gaussian kernels. The results are based on 500 simulation runs. In figure 4 the estimated mean integrated squared error is shown as function in x and the estimated mean squared error is shown as function in y for fixed $x = 0.5$. We observe almost no difference between the performances of $\hat{F}_n$ and $\hat{F}_n^*$ for examples 1 (first row in figure 4) and 2 (second row), but for example 3 (third row) the empirical likelihood method clearly outperforms the simple kernel estimator. In the last row of figure 4 we also present results for examples $g(\varepsilon, x) = (\varepsilon, \varepsilon^3, \varepsilon^5)$ (first two panels) and $g(\varepsilon, x) = (\varepsilon, I\{\varepsilon \leq 0\} - 0.5)$ (last two panels), cf. Remark 3.3. Especially for the last example we observe a great improvement by the empirical likelihood method due to the knowlegde of the error distribution in one particular point.

FIGURE 4 HERE

5 Discussion of future applications and conclusion

5.1 Bootstrap for the general nonparametric regression model

In nonparametric regression asymptotic distributions of estimators and test statistics often depend on unknown features and estimation of these leads to slow convergence of asymptotic procedures. Hence, often resampling methods such as bootstrap are applied. For bootstrap a new data set $(X_i^\star, Y_i^\star)$, $i = 1, \ldots, n$, has to be generated. In regression usually one uses the same measurement points as in the original data set, i. e. $X_i^\star = X_i$ and defines bootstrap observations by $Y_i^\star = \tilde{m}(X_i) + \varepsilon_i^\star$. Here, $\tilde{m}$ denotes some suitable estimator for the regression function and new errors $\varepsilon_i^\star$ are generated with different methods.

In the homoscedastic model (1.2) usually residual bootstrap is applied, where the errors $\varepsilon_i^\star$ are drawn with replacement from residuals $\hat{\varepsilon}_j = Y_j - \hat{m}(X_j)$, $j = 1, \ldots, n$, i. e. they are generated from the empirical distribution function of residuals. In most cases the bootstrap errors are also centered with respect to the conditional distribution given the original data (by substracting the mean $n^{-1} \sum_{j=1}^n \hat{\varepsilon}_j$) to reflect the condition $E[\varepsilon] = 0$ for the bootstrap model, see Härdle & Bowman (1988). Note that the centering is not necessary to prove asymptotic validity of most bootstrap procedures, because $n^{-1} \sum_{j=1}^n \hat{\varepsilon}_j = o_p(n^{-1/2})$, see Müller, Schick & Wefelmeyer (2004), but can have advantages for finite sample sizes and is often recommended.

In the heteroscedastic model (1.3) the described bootstrap data generation does not reflect the original model and wild bootstrap is mostly applied, see Härdle & Mammen (1993). As alternative, heteroscedastic residual bootstrap can be applied, where the bootstrap errors are built as $\varepsilon_i^\star = \hat{\sigma}(X_i)\eta_i^\star$ and $\eta_i^\star$ is generated from the empirical distribution function of residuals $\hat{\eta}_j = (Y_j - \hat{m}(X_j))/\hat{\sigma}(X_j)$, $j = 1, \ldots, n$, standardized by substracting the mean $n^{-1}\sum_{j=1}^n \hat{\eta}_j$ and dividing by the standard deviation $(n^{-1}\sum_{k=1}^n(\hat{\eta}_k - n^{-1}\sum_{j=1}^n \hat{\eta}_j)^2)^{1/2}$ to reflect the conditions $E[\eta] = 0$, $E[\eta^2] = 1$ in the original model. See for instance Neumeyer (2008).

Both methods will not work in general in the model (1.4) considered in the paper at hand. This was already observed by Zhu, Fujikoshi & Naito (2001). Those authors consider a test for constant variance and a modification of the wild bootstrap for cases where model (1.3) is violated and $E[\varepsilon^4 \mid X = x]$ depends on x [i. e. they consider our general model (1.4)]. They show that in this case wild bootstrap does not work in general. They do not give a general solution to that problem, but a modification of the wild bootstrap, which works specific for their test statistic.

A general solution would be to generate $\varepsilon_i^\star$ for a given covariate $X_i = x$ from an estimator for the distribution $F_{\varepsilon|X}(\cdot|x)$. The kernel based estimator $\hat{F}_n(\cdot|x)$ from section 2 can be applied, for instance, and Theorem 2.1 will be helpful for proving asymptotic correctness of such bootstrap procedures. The centeredness of the error, i. e. $E[\varepsilon \mid X] = \int y F_{\varepsilon|X}(dy|X) = 0$ should be reflected in the bootstrap data generation as well. This can either be done by substracting the mean $\int y\hat{F}_n(dy|x)$ [i. e. a Nadaraya-Watson estimator based on the sample $(X_i, \hat{\varepsilon}_i)$, $i = 1, \ldots, n$] from the generated bootstrap errors, or by application of the proposed empirical likelihood method by setting $g(\varepsilon, x) = \varepsilon$.

When generating bootstrap samples for hypotheses testing, it is necessary to generate data, which fulfill the null hypothesis, see e. g. Shao & Tu (1995). If the null hypothesis can be written in the form of equation (1.5) this can be done by the empirical likelihood method. Consider for example a test for a parametric form of the variance function with the null hypothesis

$$H_0 : \exists \vartheta \in \Theta \text{ such that } \operatorname{Var}(Y \mid X) = \sigma_\vartheta^2(X), \tag{5.1}$$

where the parametric variance function σ_ϑ^2 is known except for a finite dimensional parameter ϑ. A test for constant variance (often called homoscedasticity) is a special case, where $\sigma_\vartheta^2 = \vartheta$ and $\Theta = \mathbb{R}^+$. To generate a (centered) bootstrap error $\varepsilon_i^\star$ for covariate X_i one would now sample from the empirical likelihood distribution function $\hat{F}_n^*(\cdot|X_i)$ as defined above by setting $g(\varepsilon, x) = (\varepsilon, \varepsilon^2 - \sigma_{\hat{\vartheta}}^2(x))^\top$, where ϑ beforehand has been estimated by $\hat{\vartheta}$ (see Remark 3.4). Doing so would be suitable to obtain an alternative version of the test by Zhu, Fujikoshi & Naito (2001) and would moreover generalize other procedures for testing hypothesis (5.1) from the literature, see the introduction for references. For instance, Dette & Hetzler (2009a) assume our general model (1.4) and suggest a test for (5.1) based on an empirical process of pseudo residuals. For application of the test, those authors consider a heteroscedastic residual bootstrap as described above and show its good performance in simulations for the heteroscedastic regression model (1.3). This bootstrap method presumably will fail when the conditional fourth moment of the error actually depends on the covariate, because this function explicitely appears in the asymptotic distribution of the test statistic, but the heteroscedastic residual bootstrap does not reflect this dependence correctly. The stochastic expansion as given in our Theorem 3.1 (a) will be helpful to prove asymptotic validity of the general bootstrap version of Dette & Hetzler's (2009a) test. However, this is beyond the scope of the paper at hand. Note that although in Dette & Hetzler (2009b) the same authors consider an asymptotically distribution-free version of their test [based on a martingale transformation due to Khmaladze (1981)] they show that for small sample sizes the bootstrap version has better power (those simulations are only for model (1.3), where their bootstrap version is valid).

In this paper we only consider mean regression models. Similar theory can be developed for quantile regression models . In such a model for a given $\alpha \in (0,1)$ the regression function m is defined as the conditional α-quantile of Y, given the covariate X, i. e. $P(Y \leq m(x) \mid X = x) = \alpha$. Hence, when writing such a model as $Y = m(X) + \varepsilon$ it is essential that the error distribution is allowed to depend on the covariate such as in model (1.4) for mean regression. The error ε has conditional α-quantile (given X) zero. Bootstrap samples for such a model can now be built as $Y_i^\star = \tilde{m}(X_i) + \varepsilon_i^\star$, where $\varepsilon_i^\star$ is generated from the empirical likelihood distribution $\hat{F}_n^*(\cdot|X_i)$ defined as above, based on residuals $\varepsilon_j = Y_j - \hat{m}(X_j)$ [here $\tilde{m}$ and $\hat{m}$ denote nonparametric quantile curve estimators, see for instance Koenker (2005) or Dette & Volgushev (2008)], and with $g(\varepsilon, x) = I\{\varepsilon \leq 0\} - \alpha$. This way one forces the bootstrap observations to fulfill the original model. Higher dimensional functions g can be chosen, when additional information has to be incorporated, for instance for the sake of generating observations that fulfill some null hypothesis. See also Härdle, Ritov & Song (2010) who construct confidence bands for the regression function by using a similar bootstrap where $\varepsilon_i^\star$ is generated from the quantile regression version of $\hat{F}_n(\cdot|X_i)$.

5.2 Empirical likelihood hypotheses testing

Hypothesis tests for validity of (1.5) can be based on a distance measure between the kernel estimator and the empirical likelihood estimator for the conditional error distribution, where the latter one is only consistent under the null hypothesis (1.5). When an L^2-distance is applied, e. g., we end with a consistent test statistic

$$nh \int (\hat{F}_n(y|x) - \hat{F}_n^*(y|x))^2 d(y,x) \tag{5.2}$$

The asymptotic distribution can be derived with similar methods as used in the proof of Theorem 3.1. Detailed asymptotic theory and investigation of the numerical performance are a future research project, but beyond the scope of the paper at hand.

The method can e. g. be used to test for a vanishing conditional third moment ($g(\varepsilon, x) = \varepsilon^3$) or median ($g(\varepsilon, x) = I\{\varepsilon \leq 0\} - 0.5$) or to test whether the conditional fourth error moment depends on the covariate or not [$g_\vartheta(\varepsilon, x) = \varepsilon^4 - \vartheta$ with pre-estimated $\vartheta = \hat{\vartheta}$; this hypothesis is important for the tests explained in section 5.1, see e. g. Zhu, Fujikoshi & Naito (2001)]. More complicated hypotheses such as (1.5) with $g(\varepsilon, x) = \varepsilon^2 - L(m(x))$ for dependence of the conditional variance from the conditional mean by some specified link function L are possible. The method can also be used to test for parametric variance structure (5.1) (with pre-estimated $\vartheta = \hat{\vartheta}$ as explained before) and so gives an alternative procedure to the ones discussed in section 5.1. For this hypothesis test statistic (5.2) is a generalization of Dette, Neumeyer & Van Keilegom's (2007) test from model (1.3) to the general model (1.4).

5.3 Conclusion

For a general nonparametric regression model we have suggested two different estimators for the conditional error distribution, given the covariate. The first estimator, $\hat{F}_n$, uses the typical kernel approach based on nonparametically estimated residuals. The second estimator, $\hat{F}_n^*$, is a modification of $\hat{F}_n$ and applies empirical likelihood weights. It can explicitly make use of the centeredness assumption on the errors as well as other additional information.
For both estimators we have given asymptotic expansions and weak convergence results. Although from a comparison of the estimators it could not be concluded that the empirical likelihood method always improves upon the estimation, we have seen its good performance in some examples.
The estimators $\hat{F}_n$ und $\hat{F}_n^*$ both will presumably be useful in bootstrap procedures for the general nonparametric repression model and can also be applied for testing of model assumptions.

A Technical assumptions

Let (X_i, Y_i) be independent and identically distributed such as (X, Y) from model (1.4) and let $\varepsilon_i = Y_i - m(X_i)$, $i = 1, \ldots, n$. Let $F_{\varepsilon|X}(\cdot|t)$ and $f_{\varepsilon|X}(\cdot|t)$ denote the conditional distribution and density function, respectively, of ε, given $X = t$. Our aim is to estimate $F_{\varepsilon|X}(\cdot|x)$ for some (fixed) $x \in (0, 1)$. Let in the following $\mathcal{U}$ denote a suitable neighborhood of x. Let further f_X denote the density of X with support $[0, 1]$ such that $\inf_{t\in[0,1]} f_X(t) > 0$. Let $(y, t) \mapsto f_{\varepsilon,X}(y, t) = f_{\varepsilon|X}(y|t)f_X(t)$ denote the joint density of (ε, X).

Model smoothness assumptions. We assume the regression function m as well as f_X to be twice continuously differentiable. We assume $f_{\varepsilon|X}(\cdot|t)$ to be twice continuously differentiable with respect to t and $t \mapsto (m \cdot f_X)''(t) - m(t)\, f_X''(t)$ to be Lipschitz-continuous. We assume $f_{\varepsilon|X}(y|t)$ to be continuously differentiable with respect to y for all t and $y \mapsto f_{\varepsilon|X}(y|x)$ to be of bounded variation. We assume $f_{\varepsilon|X}(y|t)$ as well as $\partial f_{\varepsilon|X}(y|t)/\partial t$, $\partial^2 f_{\varepsilon|X}(y|t)/\partial t^2$, $\partial f_{\varepsilon|X}(y|t)/\partial y$ to be bounded by some constant c uniformly for $y \in \mathbb{R}$ and $t \in \mathcal{U}$.

Assumptions on the kernels. Let the kernel K be twice continuously differentiable and symmetric with support $[-1,1]$ such that $\int uK(u)\,du = 0$, $K(-1) = K(1) = 0$, $\int K(u)\,du = 1$. Let the kernel $\tilde{K}$ fulfill the same assumptions.

Bandwidth assumptions. For the sequences of bandwidths $h = h_n$ and $b = b_n$ we assume for $n \to \infty$ that $nh^5 = O(1)$, $b/h \to \lambda \geq 0$, $(\log(b^{-1}))^2/(nhb^2) \to 0$, $nb^{3+3\alpha} \to \infty$ for some $\alpha > 0$, $(h/b)(\log(b^{-1}))^{\beta+1}/(nb)^{\beta} \to 0$ for β defined below, $(\log n)^{2+\gamma}/(nh^3) \to 0$ for some $\gamma > 0$.

Note that then Akritas & Van Keilegom's (2001) bandwidth condition $nb^{3+2\alpha}/\log(b^{-1}) \to \infty$ is valid and also $b^2 = O((nh)^{-1/2})$, $h^2 = O((nh)^{-1/2})$.

We assume suitable boundary modifications for the Nadaraya-Watson estimator $\hat{m}$ to obtain convergence in probability of $\sup_{t\in[0,1]} |\hat{m}(t) - m(t)|$, $\sup_{t\in[0,1]} |\hat{m}'(t) - m'(t)|$ and $\sup_{s,t} |\hat{m}'(s) - m'(s) - \hat{m}'(t) + m'(t)|\,/|s-t|^{\alpha}$ to zero [compare to Neumeyer (2009) and see, e.g. Müller (1984) or Härdle (1989, p. 130) for boundary corrections].

Empirical likelihood regularity assumptions. Let $g = (g_1, \ldots, g_k)^\top : \mathbb{R}\times[0,1] \to \mathbb{R}^k$ be a known function such that (1.5) is valid. Let $\Sigma^*(x) = E[g(\varepsilon, X)g^\top(\varepsilon, X) \mid X = x]$ exist and be positive definite. We assume $t \mapsto E[g(\varepsilon, X) \mid X = t]$ to be continuous and $t \mapsto E[g(\varepsilon, X)\varepsilon \mid X = t]$ to be twice continuously differentiable. We assume for all $j \in \{1, \ldots, k\}$ that $g_j'(y,t) = \partial g_j(y,t)/\partial y$ and $g_j''(y,t) = \partial^2 g_j(y,x)/\partial y^2$ exist for all y and all $t \in \mathcal{U}$ such that the following integrals exist and are uniformly bounded with respect to $t \in \mathcal{U}$,

$$\int |g_j'(y,t)| f_{\varepsilon,X}(y,t)\,dy, \quad \int (g_j'(y,t))^2 |\partial f_{\varepsilon,X}(y,t)/\partial t|\,dy, \quad \int |g_j'(y,t)\partial^2 f_{\varepsilon,X}(y,t)/\partial t^2|\,dy,$$

$$\int |g_j(y,t)||\partial f_{\varepsilon,X}(y,t)/\partial t|\,dy, \quad \int g_j^2(y,t)|\partial^2 f_{\varepsilon,X}(y,t)/\partial t^2|\,dy, \quad \int (g_j'(y,t))^2 f_{\varepsilon,X}(y,t)\,dy.$$

We further assume $t \mapsto E[g_j'(\varepsilon, X) \mid X = t]$ and $t \mapsto E[g_j''(\varepsilon, X) \mid X = t]$ to be continuously differentiable. We assume the existence of some $\iota > 0$ such that

$$E\Big[\frac{1}{h}\Big(K\Big(\frac{X-x}{h}\Big)g_j(\varepsilon, x)\Big)^{2+\iota}\Big] = O(1).$$

We assume the existence of $C, \beta > 0$ such that

$$\begin{aligned} &\Big|\int \frac{1}{h}K\Big(\frac{u-x}{h}\Big)\left(g_j(y+\varphi(u), x) - g_j(y,x) - \varphi(u)g_j'(y)\right) f_{\varepsilon,X}(y,u)\,d(y,u)\Big| \\ &\leq\; C \cdot \int \Big|\frac{1}{h}K\Big(\frac{u-x}{h}\Big)\Big||\varphi(u)|^{1+\beta} f_X(u)\,du \end{aligned} \tag{A.1}$$

for all $j \in \{1, \ldots, k\}$ and $\varphi \in C_{\delta}^{1+\alpha}[0,1]$. Here, the constant β was already mentioned in the bandwidth assumptions, where also α is defined. Moreover, $C_{\delta}^{1+\alpha}[0,1]$ is defined as the space of all differentiable functions $\varphi : [0,1] \to \mathbb{R}$ such that

$$\max\Big(\sup_{x\in[0,1]} |\varphi(x)| + \sup_{x\in[0,1]} |\varphi'(x)|\Big) + \sup_{x,z\in[0,1]} \frac{|\varphi'(x)-\varphi'(z)|}{|x-z|^\alpha} \le \delta.$$

We assume the existence of some $\delta, C > 0$ and some $\kappa \in (0, 2(1+\alpha))$ such that

$$E\Big[\frac{1}{h}K^2\Big(\frac{X-x}{h}\Big) \sup_{\substack{y,z\in\mathbb{R}:|y|\le\delta,\\ |z|\le\delta,|y-z|\le\xi}} (g_j(\varepsilon_1+y,x) - g_j(\varepsilon_1+z,x))^2\Big] \le C\xi^{2/\kappa} \qquad \text{(A.2)}$$

$$E\Big[\frac{1}{h}K^4\Big(\frac{X-x}{h}\Big) \sup_{y\in\mathbb{R}:|y|\le\delta} (g_j(\varepsilon_1+y,x) - g_j(\varepsilon_1,x))^4\Big] = O(1)$$

for all $j \in \{1,\dots,k\}$.

B Proofs

We start with a lemma, which is similar to Lemma 19.24 by van der Vaart (1998, p. 280).

Lemma B.1 *Let $Z_1, Z_2, \dots$ be independent random variables with distribution P, and $(\mathcal{Y}\times\mathcal{T},\rho)$ a totally bounded semimetric space. Let*

$$G_n(y,t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(\varphi_{n,y,t}(Z_i) - \int \varphi_{n,y,t}\,dP\Big),$$

$y \in \mathcal{Y}$, $t \in \mathcal{T}$, be a sequence of empirical processes that converges weakly to a process $G(y,t)$, $y \in \mathcal{Y}$, $t \in \mathcal{T}$, and is asymptotically equicontinuous with respect to ρ. Let $\hat{t}_n = \hat{t}_n(Z_1,\dots,Z_n)$ be a random sequence such that $\lim_{n\to\infty} P(\hat{t}_n \in \mathcal{T}) = 1$ and $\hat{t}_n$ converges to $0 \in \mathcal{T}$ in probability. Let further $G(y,0) = 0$ almost surely for all $y \in \mathcal{Y}$, and

$$\sup_{y\in\mathcal{Y}} \rho((y,t_n),(y,0)) \to 0 \text{ whenever } t_n \to 0 \text{ in } \mathcal{T}. \qquad \text{(B.1)}$$

Then, $G_n(y,\hat{t}_n)$ converges to zero in probability uniformly with respect to $y \in \mathcal{Y}$.

Proof of Lemma B.1. We may assume for the proof that $\hat{t}_n \in \mathcal{T}$.

Now $(G_n(\cdot,\cdot),\hat{t}_n)$ converges weakly to $(G(\cdot,\cdot),0)$ by Slutsky's lemma [see Kosorok (2008), Theorem 7.15, p. 112]. Let $\mathbb{D} = \ell^\infty(\mathcal{Y}\times\mathcal{T})\times\mathcal{T}$ and $\mathbb{D}_0 = UC(\mathcal{Y}\times\mathcal{T})\times\{0\} \subset \mathbb{D}$, where $UC(\mathcal{Y}\times\mathcal{T})$ denotes the subset of maps in $\ell^\infty(\mathcal{Y}\times\mathcal{T})$, which are uniformly continuous with respect to ρ. From Theorem 7.19 in Kosorok (2008, p. 114) and our assumptions it follows that $P((G(\cdot,\cdot),0) \in \mathbb{D}_0) = 1$. Let further the map $g : \mathbb{D} \to \ell^\infty(\mathcal{Y})$ be defined by $g(x(\cdot,\cdot),t) = x(\cdot,t)$. Then, g is continuous in $\mathbb{D}_0$. To see this consider a sequence $(x_n(\cdot,\cdot),t_n)$ in $\mathbb{D}$ which converges to $(x(\cdot,\cdot),0) \in \mathbb{D}_0$, i. e.

$$\sup_{y\in\mathcal{Y},t\in\mathcal{T}} |x_n(y,t) - x(y,t)| \to 0, \qquad \text{(B.2)}$$

where x is uniformly continuous with respect to ρ, and $t_n \to 0$ in $\mathcal{T}$. Note that

$$\begin{aligned}\sup_{y\in\mathcal{Y}}|g(x_n(\cdot,\cdot),t_n)(y) - g(x(\cdot,\cdot),0)(y)| &= \sup_{y\in\mathcal{Y}}|x_n(y,t_n) - x(y,0)| \\ \leq \sup_{y\in\mathcal{Y}}|x_n(y,t_n) - x(y,t_n)| &+ \sup_{y\in\mathcal{Y}}|x(y,t_n) - x(y,0)|.\end{aligned}$$

Here, the first term on the right-hand side converges to zero by (B.2). Further because of (B.1) for each $\delta > 0$ there exists some n_0 such that for all $n \geq n_0$ the second term can be bounded by

$$\sup_{\substack{y\in\mathcal{Y},t\in\mathcal{T}\\ \rho((y,t),(y,0))<\delta}} |x(y,t) - x(y,0)|.$$

Now the latter term converges to zero for $\delta \to 0$ because of the assumed uniform continuity of x.

From the continuous mapping theorem [see Kosorok (2008), Theorem 7.7, p. 109] it follows that the process $g(G_n(\cdot,\cdot),\hat{t}_n) = G_n(\cdot,\hat{t}_n)$ converges weakly to $g(G(\cdot,\cdot),0) = G(\cdot,0) = 0$. The asserted uniform convergence in probability follows from Lemma 1.10.2 in van der Vaart and Wellner (2000). □

B.1 Proof of Theorem 2.1

(a) For $\hat{F}_n$ defined in (2.1) we use the decomposition

$$\hat{F}_n(y|x) - F_{\varepsilon|X}(y|x) = [F_n(y|x) - F_{\varepsilon|X}(y|x)] + [\hat{F}_n(y|x) - F_n(y|x)], \qquad \text{(B.3)}$$

where, by definition in (2.2),

$$F_n(y|x) - F_{\varepsilon|X}(y|x) = S_n(y)\Big(1 - \frac{\bar{f}_X(x) - f_X(x)}{\bar{f}_X(x)}\Big)$$

with

$$S_n(y) = \frac{1}{f_X(x)}\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)(I\{\varepsilon_i \leq y\} - F_{\varepsilon|X}(y|x)), \qquad \text{(B.4)}$$

and where $\bar{f}_X(x) = (nh)^{-1}\sum_{j=1}^n K(\frac{X_j - x}{h})$ is a uniformly consistent estimator for the density value $f_X(x) > 0$. From the proof of Theorem 2.1 (b) below we have that $S_n(y)$ is of order $O(h^2) + O_p((nh)^{-1/2}) = O_p((nh)^{-1/2})$. Hence it follows that

$$F_n(y|x) - F_{\varepsilon|X}(y|x) = S_n(y) + o_p(\frac{1}{\sqrt{nh}}) \qquad \text{(B.5)}$$

uniformly with respect to $y \in \mathbb{R}$.

Now we will consider the second process in the decomposition (B.3), i. e.

$$\begin{aligned}\hat{F}_n(y|x) - F_n(y|x) &= \frac{1}{\hat{f}_X(x)}\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)(I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\}) \\ &= \frac{1}{f_X(x)}\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)(I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\}) + o_p(\frac{1}{\sqrt{nh}}) \quad \text{(B.6)}\end{aligned}$$

uniformly with respect to $y \in \mathbb{R}$, where the replacement of the density estimator in the denominator follows similarly to the argumentation before. Let

$$\begin{aligned}R_n(y) &= \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big)(I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\}) \quad \text{(B.7)} \\ &\quad - \int \frac{1}{h}K\Big(\frac{z - x}{h}\Big)(I\{\epsilon + m(z) - \hat{m}(z) \le y\} - I\{\epsilon \le y\}) f_{\varepsilon,X}(\epsilon, z)\, d(\epsilon, z),\end{aligned}$$

then we will show in Lemma B.2 that $\sup_{y\in\mathbb{R}} |R_n(y)| = o_p(\frac{1}{\sqrt{nh}})$.

From this and (B.6), (B.7) it follows that

$$\begin{aligned}&\hat{F}_n(y|x) - F_n(y|x) \\ &= \int \frac{1}{h}K\Big(\frac{z - x}{h}\Big)(I\{\epsilon + m(z) - \hat{m}(z) \le y\} - I\{\epsilon \le y\})\frac{f_{\varepsilon,X}(\epsilon, z)}{f_X(x)}\, d(\epsilon, z) + o_p(\frac{1}{\sqrt{nh}}) \\ &= \int \frac{1}{h}K\Big(\frac{z - x}{h}\Big)\Big(F_{\varepsilon|X}(y - m(z) + \hat{m}(z)|z) - F_{\varepsilon|X}(y|z)\Big)\frac{f_X(z)}{f_X(x)}\, dz + o_p(\frac{1}{\sqrt{nh}}) \\ &= \int \frac{1}{h}K\Big(\frac{z - x}{h}\Big) f_{\varepsilon|X}(y|z)(\hat{m}(z) - m(z))\frac{f_X(z)}{f_X(x)}\, dz + o_p(\frac{1}{\sqrt{nh}})\end{aligned}$$

by Taylor's expansion. Inserting the definition of $\hat{m}$ we obtain

$$\begin{aligned}&\hat{F}_n(y|x) - F_n(y|x) \\ &= \frac{1}{f_X(x)}\int \frac{1}{h}K\Big(\frac{z - x}{h}\Big) f_{\varepsilon|X}(y|z)\frac{1}{n}\sum_{i=1}^n \frac{1}{b}\tilde{K}\Big(\frac{X_i - z}{b}\Big)\varepsilon_i\, dz \quad \text{(B.8)} \\ &\quad + \frac{1}{f_X(x)}\int \frac{1}{h}K\Big(\frac{z - x}{h}\Big) f_{\varepsilon|X}(y|z)\frac{1}{n}\sum_{i=1}^n \frac{1}{b}\tilde{K}\Big(\frac{X_i - z}{b}\Big)(m(X_i) - m(z))\, dz \\ &\quad + o_p(\frac{1}{\sqrt{nh}})\end{aligned}$$

where we have replaced the random denominator $\hat{f}_X(z)$ by $f_X(z)$. Further one can show by arguments that are standard in kernel estimation theory that uniformly with respect to $y \in \mathbb{R}$,

$$\begin{aligned}&\frac{1}{f_X(x)}\int \frac{1}{h}K\Big(\frac{z - x}{h}\Big) f_{\varepsilon|X}(y|z)\frac{1}{n}\sum_{i=1}^n \frac{1}{b}\tilde{K}\Big(\frac{X_i - z}{b}\Big)(m(X_i) - m(z))\, dz \\ &= \frac{b^2}{2}\frac{f_{\varepsilon|X}(y|x)}{f_X(x)}((mf_X)''(x) - (mf_X'')(x))\int \tilde{K}(u)u^2\, du + o_p(\frac{1}{\sqrt{nh}}), \quad \text{(B.9)}\end{aligned}$$

and the assertion now follows from (B.3), (B.4), (B.5) and (B.8), (B.9). □

(b) For the structure of the bias term note that

$$E\Big[\frac{1}{f_X(x)}\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i-x}{h}\Big)(I\{\varepsilon_i\le y\}-F_{\varepsilon|X}(y|x))\Big] \;=\; h^2H(y|x)+o(h^2)$$

by standard arguments of kernel estimation theory, and that $o(h^2)=o((nh)^{-1/2})$. Hence, from Theorem 2.1(a) we obtain that $G_n(y|x)=G_n^{(1)}(y|x)+G_n^{(2)}(y|x)+o_p(1)$ uniformly with respect to $y\in\mathbb{R}$, where $(k=1,2)$

$$G_n^{(k)}(y|x) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\varphi_{n,y}^{(k)}(\varepsilon_i,X_i)-E[\varphi_{n,y}^{(k)}(\varepsilon,X)]\Big)$$

and

$$\begin{aligned}\varphi_{n,y}^{(1)}(\varepsilon,X) \;&=\; \frac{1}{f_X(x)}\frac{1}{\sqrt{h}}K\Big(\frac{X-x}{h}\Big)(I\{\varepsilon\le y\}-F_{\varepsilon|X}(y|x))\\ \varphi_{n,y}^{(2)}(\varepsilon,X) \;&=\; \frac{1}{f_X(x)}\varepsilon\int\frac{1}{\sqrt{h}}K\Big(\frac{z-x}{h}\Big)f_{\varepsilon|X}(y|z)\frac{1}{b}\tilde{K}\Big(\frac{X-z}{b}\Big)\,dz.\end{aligned}$$

In the following we will sketch a proof of weak convergence of the processes $G_n^{(k)}(\cdot|x)$ $(k=1,2)$. From this follows asymptotic stochastic equicontinuity for both processes and, hence, for the sum $G_n^{(1)}(\cdot|x)+G_n^{(2)}(\cdot|x)$. Convergence of the finite dimensional distributions of $G_n^{(1)}(\cdot|x)+G_n^{(2)}(\cdot|x)$ is shown by standard arguments applying Cramér-Wold's device and the Central Limit Theorem (including a straightforward calculation of the covariances). From Theorem 2.1 in Kosorok (2008, p. 15) then follows the desired weak convergence of $G_n^{(1)}(\cdot|x)+G_n^{(2)}(\cdot|x)$ and thus of $G_n(\cdot|x)$.

To show weak convergence of the process $G_n^{(1)}(\cdot|x)$ consider the function class $\{\varphi_{n,y}^{(1)}\mid y\in\mathbb{R}\}=\Phi_n\cdot(\mathcal{F}-\mathcal{G})$ with envelope $\Phi_n(\epsilon,z)=K((X-x)/h)/(\sqrt{h}f_X(x))$, where $\mathcal{F}=\{\epsilon\mapsto I\{\varepsilon\le y\}\mid y\in\mathbb{R}\}$ and $\mathcal{G}=\{\epsilon\mapsto F_{\varepsilon|X}(y|x)\mid y\in\mathbb{R}\}$, and the class $\mathcal{F}-\mathcal{G}$ has the constant envelope 1. Then from the proof of Theorem 9.15 by Kosorok (2008) it follows that

$$N(\eta||\Phi_n||_{2,Q},\Phi_n\cdot(\mathcal{F}-\mathcal{G}),L_2(Q))\le\sup_{Q'}N(\eta,\mathcal{F}-\mathcal{G},L_2(Q'))$$

which is of polynomial growth in η because $\mathcal{F}$ and $\mathcal{G}$ are both VC-classes. An application of Theorem 2.11.22 by van der Vaart & Wellner (2000, p. 220) (where the index space $\mathbb{R}$ is equipped with the semimetric $\rho(s,t)=|F_{\varepsilon|X}(s|x)-F_{\varepsilon|X}(t|x)|$) yields the desired weak convergence.

To show weak convergence of the process $G_n^{(2)}(\cdot|x)$ note that the functions have the structure $\varphi_{n,y}^{(2)}(\varepsilon,X)=\int\psi_n(\varepsilon,X,z)f_{\varepsilon|X}(y|z)\,dz$, where the class $\mathcal{G}=\{z\mapsto f_{\varepsilon|X}(y|z)\mid y\in\mathbb{R}\}$ is a subset of the space $C_c^2[0,1]$ of twice differentiable functions on $[0,1]$, where the function as well as the derivatives are bounded by the constant c. From Theorem 2.7.1, van der

Vaart and Wellner (2000, p. 155), it follows that for the bracketing number of $\mathcal{G}$ with respect to the supremum norm it follows that $\log N_{[]}(\eta, \mathcal{G}, || \cdot ||_\infty) \leq K\eta^{-1/2}$ for some constant K. Those brackets $[\ell, u]$ in $\mathcal{G}$ are used to cover the class $\{\varphi_{n,y}^{(2)} \mid y \in \mathbb{R}\}$ by (the same number of) $\tilde{c}\eta$-brackets (for some constant $\tilde{c}$) of the form $[\int \psi_n(\varepsilon, X, z)\ell(z)\,dz, \int \psi_n(\varepsilon, X, z)u(z)\,dz]$ (consider $\varepsilon \geq 0$ here for simplicity). The assertion follows with an application of Theorem 2.11.23 by van der Vaart & Wellner (2000, p. 220). □

Lemma B.2 *For R_n defined in (B.7) we have under the assumptions of Theorem 2.1 that* $\sup_{y\in\mathbb{R}} |R_n(y)| = o_p(\frac{1}{\sqrt{nh}})$.

Proof of Lemma B.2. Consider the empirical process

$$H_n(y,t) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\varphi_{n,y,t}(\varepsilon_i, X_i) - \int \varphi_{n,y,t}(\epsilon, z) f_{\varepsilon,X}(\epsilon, z)d(\epsilon, z)\Big),$$

where

$$\varphi_{n,y,t}(\epsilon, z) \;=\; \frac{1}{\sqrt{h}}K\Big(\frac{z-x}{h}\Big)\Big[I\{\epsilon + t(z) \leq y\} - I\{\epsilon \leq y\}\Big]$$

indexed by $y \in \mathbb{R}$ and functions $t \in \mathcal{T} = C_\delta^{1+\alpha}[0,1]$ (the space is defined in appendix section A). The dependence of the function $\varphi_{n,y,t}$ on the sample size n arises from the bandwidths $h = h_n$. Note that $\hat{\varepsilon}_i = \varepsilon_i + \hat{t}_n(X_i)$ for $\hat{t}_n = m - \hat{m}$,

$$\sqrt{nh}R_n(y) \;=\; H_n(y, \hat{t}_n),$$

and Akritas & Van Keilegom (2001) show that $\lim_{n\to\infty} P(\hat{t}_n \in \mathcal{T}) = 1$. Further $\hat{t}_n$ converges in probability to $0 \in \mathcal{T}$, and $\varphi_{n,y,0} = 0$. To prove (B.14) we will apply Lemma B.1. To this end, first we show weak convergence of the process $H_n(y,t)$, $y \in \mathbb{R}$, $t \in \mathcal{T}$ to a centered Gaussian process $H(y,t)$, $y \in \mathbb{R}$, $t \in \mathcal{T}$, by an application of Theorem 2.11.23 of van der Vaart & Wellner (2000, p. 220). The function class $\mathcal{F}_n = \{\varphi_{n,y,t} \mid y \in \mathbb{R}, t \in \mathcal{T}\}$ has the square-integrable envelope

$$\Phi_n(\varepsilon, z) = \frac{1}{\sqrt{h}}K\Big(\frac{z-x}{h}\Big).$$

Hence, the first two conditions in (2.11.21) of the aforementioned reference follow. We further have pointwise convergence of the covariances and $\mathbb{R} \times \mathcal{T}$ is a totally bounded semimetric space with semimetric

$$\rho((y,t),(z,s)) \;=\; |F_{\varepsilon|X}(y|x) - F_{\varepsilon|X}(z|x)| + |F_{\varepsilon|X}(y - t(x)|x) - F_{\varepsilon|X}(z - s(x)|x)|.$$

We have validity of the third condition in (2.11.21), i. e.

$$\sup_{\rho((y,t),(z,s))<\delta_n} E\Big[(\varphi_{n,y,t}(\varepsilon, X) - \varphi_{n,z,s}(\varepsilon, X))^2\Big]$$

$$
\begin{aligned}
&= \sup_{\rho((y,t),(z,s))<\delta_n} \frac{1}{h} E\Big[K^2\Big(\frac{X-x}{h}\Big)\Big(I\{\varepsilon \le y - t(X)\} - I\{\varepsilon \le y\} - I\{\varepsilon \le z - s(X)\} + I\{\varepsilon \le z\}\Big)^2\Big] \\
&\le \sup_{\rho((y,t),(z,s))<\delta_n} 2\int \frac{1}{h} K^2\Big(\frac{u-x}{h}\Big)\Big(|F_{\varepsilon|X}(y-t(u)|u) - F_{\varepsilon|X}(z-s(u)|u)| \\
&\qquad\qquad + |F_{\varepsilon|X}(y|u) - F_{\varepsilon|X}(z|u)|\Big) f_X(u)\,du \\
&\le C\rho((y,t),(z,s)) + r_n \;=\; O(\delta_n) + r_n \;=\; o(1),
\end{aligned}
$$

by a change of variable $(u-x)/h = v$ and suitable bounding of remainder terms such that the sequence $r_n = o(1)$ does not depend on y, t, z or s. The function class $\mathcal{F}_n$ is the product of the envelope Φ_n and the difference $\mathcal{F}_1 - \mathcal{F}_2$ of function classes

$$
\mathcal{F}_1 = \{(\epsilon, z) \mapsto I\{\epsilon + t(z) \le y\} \mid y \in \mathbb{R}, t \in \mathcal{T}\}, \quad \mathcal{F}_2 = \{(\epsilon, z) \mapsto I\{\epsilon \le y\} \mid y \in \mathbb{R}, t \in \mathcal{T}\}.
$$

Because $\mathcal{F}_1 \subset \mathcal{F}_2$, in the following we only consider $\mathcal{F}_{n,1} = \Phi_n \mathcal{F}_1$. Let $[d_i^L, d_i^U]$, $i = 1, \ldots, m$ denote η^2-brackets for $C_\delta^{1+\alpha}[0,1]$ with respect to the supremum norm. For this we have $m = N_{[\,]}(\eta^2, \mathcal{T}, ||\cdot||_\infty) \le \exp(c\eta^{-2/(1+\alpha)})$ for some constant c by Theorem 2.7.1, van der Vaart and Wellner (2000). Let for given z the numbers $y_{i,k}^L$ give a partition of the line in segments having probability less or equal to η^2 with respect to $y \mapsto P(\epsilon + d_i^L(z) \le y \mid X = z)$ and the analogous definition is valid for $y_{i,k}^U$ with respect to $y \mapsto P(\epsilon + d_i^U(z) \le y \mid X = z)$. For given $z \in [0,1]$ and $y \in \mathbb{R}$ let now $y_{i,k_1}^L(z)$ be the largest of the $y_{i,k}^L$ that is less or equal to y and $y_{i,k_2}^U(z)$ the smallest of the $y_{i,k}^U(z)$ greater than or equal to y. For this we need $k_1, k_2 \in \{1, \ldots, K\}$, $K = O(\epsilon^{-2})$. The construction defines a bracket $[\ell, u]$ for $(\epsilon, z) \mapsto I\{\epsilon + t(z) \le y\}$ in the class $\mathcal{F}_1$ by considering

$$
\ell(\epsilon, z) = I\{\epsilon + d_i^L(z) \le y_{i,k_1}^L(z)\}, \quad u(\epsilon, z) = I\{\epsilon + d_i^U(z) \le y_{i,k_2}^U(z)\},
$$

By multiplication of those brackets with the envelope Φ_n we obtain brackets for the class $\mathcal{F}_{n,1}$. Those have $L_2(P)$-length

$$
\begin{aligned}
&\Big(\frac{1}{h}\int K^2(\frac{z-x}{h})(u(y,z) - \ell(y,z))^2 f_{\varepsilon|X}(y|z) f_X(z)\, d(y,z)\Big)^{1/2} \\
&\le \Big(\frac{1}{h}\int K^2(\frac{z-x}{h}) f_X(z)\, dz \sup_{s\in[0,1]} \int (u(y,s) - \ell(y,s))^2 f_{\varepsilon|X}(y|s)\, dy\Big)^{1/2} \\
&\le C\Big(\sup_{s\in[0,1]} \Big|F_{\varepsilon|X}(y_{i,k_2}^U(s) - d_i^U(s)|s) - F_{\varepsilon|X}(y_{i,k_1}^L(s) - d_i^L(s)|s)\Big|\Big)^{1/2} \\
&\le C\Big(\sup_{s\in[0,1]} \Big|F_{\varepsilon|X}(y(s) - d_i^U(s)|s) - F_{\varepsilon|X}(y(s) - d_i^L(s)|s)\Big| + \eta^2\Big)^{1/2}
\end{aligned}
$$

for some constant C by construction of the brackets for all $y(s) \in [y_{i,k_1}^L(s), y_{i,k_2}^U(s)]$. This further can be bounded by

$$
C\Big(\sup_{s\in[0,1] y\in\mathbb{R}} |f_{\varepsilon|X}(y|s)| \sup_{s\in[0,1]} |d_i^U(s) - d_i^L(s)| + \eta^2\Big)^{1/2} \le \tilde{C}\eta^2
$$

for some constant $\tilde{C}$. Because there exist positive constants c_1, c_2 such that $c_1 \leq ||\Phi_n||_{2,P} \leq c_2$ for all n we obtain for the class $\mathcal{F}_{n,1}$ that

$$N_{[]}(\eta||\Phi_n||_{2,P}, \mathcal{F}_{n,1}, L_2(P)) \leq N_{[]}(c\eta, \mathcal{F}, L_2(P)) = O(\eta^{-2}\exp(\tilde{c}\eta^{-2/(1+\alpha)}))$$

and the same bound is valid for the bracketing number of $\mathcal{F}_n$. Hence, we obtain that

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}(\eta||\Phi_n||_{2,P}, \mathcal{F}_n, L_2(P))}\, d\eta \to 0 \ \text{ for all } \ \delta_n \to 0$$

which is the remaining condition of Theorem 2.11.23 by van der Vaart & Wellner (2000, p. 220). With this theorem weak convergence of the process H_n to a centered Gaussian process and asymptotic equicontinuity with respect to ρ follows.

From the defintion of ρ and continuity of $F_{\varepsilon|X}$ it follows that $\rho((y, t_n), (y, 0)) \to 0$ uniformly in y whenever $t_n \to 0$. Hence, Lemma B.1 is applicable and yields uniform convergence of $H_n(\cdot, \hat{t}_n)$ to zero in probability. □

B.2 Proof of Theorem 3.1

Proposition B.3 *Under the assumptions of Theorem 3.1 one has*

$$\begin{aligned}
\hat{\eta}_n(x) \;=\; & \Sigma^{-1}(x)\Big[\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) g(\varepsilon_i, x) \\
& - \frac{1}{n}\sum_{i=1}^n \varepsilon_i \int \frac{1}{h}K\Big(\frac{z-x}{h}\Big) g'(y,x)\frac{1}{b}\tilde{K}\Big(\frac{X_i - z}{b}\Big) f_{\varepsilon|X}(y|z)\, d(y,z) \\
& - \frac{b^2}{2}((mf_X)''(x) - (mf_X'')(x)) \int g'(y,x) f_{\varepsilon|X}(y|x)\, dy \int \tilde{K}(u)u^2\, du\Big] + o_p(\frac{1}{\sqrt{nh}}),
\end{aligned}$$

where $g'(y,x) = \frac{\partial\, g(y,x)}{\partial\, y}$ *and* $\Sigma(x) = E[g(\varepsilon, X)g^\top(\varepsilon, X) \mid X = x] f_X(x) \int K^2(u)\, du$.

Proof of Proposition B.3. The coefficient $\hat{\eta}_n(x)$ was defined via the following equality,

$$\begin{aligned}
0 \;=\; & \frac{1}{n}\sum_{i=1}^n \frac{g(\hat{\varepsilon}_i, x)\frac{1}{h}K(\frac{X_i - x}{h})}{1 + \hat{\eta}_n^\top(x) g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})} \\
=\; & \frac{1}{nh}\sum_{i=1}^n g(\hat{\varepsilon}_i, x)\, K\Big(\frac{X_i - x}{h}\Big) - \frac{1}{nh}\sum_{i=1}^n g(\hat{\varepsilon}_i, x) g^\top(\hat{\varepsilon}_i, x) K^2\Big(\frac{X_i - x}{h}\Big)\hat{\eta}_n(x) \\
& + \frac{1}{nh}\sum_{i=1}^n \frac{g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})\, (\,\hat{\eta}_n^\top(x) g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})\,)^2}{1 + \hat{\eta}_n^\top(x) g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})} \\
=\; & \frac{1}{nh}\sum_{i=1}^n g(\hat{\varepsilon}_i, x)\, K\Big(\frac{X_i - x}{h}\Big) - \Sigma_n(x)\hat{\eta}_n(x) + o_p(\frac{1}{\sqrt{nh}}), \qquad \text{(B.10)}
\end{aligned}$$

where the last equality can be shown analogously to the proof of Proposition 3.2 by Kiwitt, Nagel & Neumeyer (2008), and

$$\Sigma_n(x) \;=\; \frac{1}{nh}\sum_{i=1}^{n} g(\hat{\varepsilon}_i, x) g^\top(\hat{\varepsilon}_i, x) K^2\Big(\frac{X_i - x}{h}\Big) \;=\; \Sigma(x) + o_p(1). \tag{B.11}$$

Combining (B.10) and (B.11) one now obtains

$$\hat{\eta}_n(x) = \Sigma^{-1}(x)\frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\, g(\hat{\varepsilon}_i, x) + o_p(\frac{1}{\sqrt{nh}}). \tag{B.12}$$

Now let $j \in \{1, \ldots, k\}$. For

$$\begin{aligned} R_n \;=&\; \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) g_j(\hat{\varepsilon}_i, x) - \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) g_j(\varepsilon_i, x) \\ &- \int \frac{1}{h} K\Big(\frac{z - x}{h}\Big) g_j'(y, x)(m(z) - \hat{m}(z)) f_{\varepsilon,X}(y, z)\, d(y, z) \end{aligned} \tag{B.13}$$

we will show that

$$R_n \;=\; o_p(\frac{1}{\sqrt{nh}}). \tag{B.14}$$

To this end, we will apply Lemma B.1 to the empirical process

$$G_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Big(\varphi_{n,t}(\varepsilon_i, X_i) - \int \varphi_{n,t}(y, z) f_{\varepsilon,X}(y, z) d(y, z)\Big),$$

where

$$\varphi_{n,t}(\varepsilon, z) \;=\; \frac{1}{\sqrt{h}} K\Big(\frac{z - x}{h}\Big)\Big[g_j(\varepsilon + t(z), x) - g_j(\varepsilon, x)\Big]$$

indexed by functions $t \in \mathcal{T} = C_\delta^{1+\alpha}[0, 1]$ defined as in the proof of Lemma B.2. The dependence from the sample size n arises from the bandwidths $h = h_n$. Note that $\hat{\varepsilon}_i = \varepsilon_i + \hat{t}_n(X_i)$ for $\hat{t}_n = m - \hat{m}$,

$$\sqrt{nh} R_n \;=\; G_n(\hat{t}_n),$$

and $\lim_{n\to\infty} P(\hat{t}_n \in \mathcal{T}) = 1$. Further $\hat{t}_n$ uniformly converges to zero almost surely and $\varphi_{n,0} = 0$. First we will show weak convergence of the process $G_n(t)$, $t \in \mathcal{T}$, to a Gaussian process $G(t)$, $t \in \mathcal{T}$. By definition of $\mathcal{T}$,

$$\Phi_n(\varepsilon, z) = \frac{1}{\sqrt{h}}\Big|K\Big(\frac{z - x}{h}\Big)\Big| \sup_{y\in\mathbb{R}:\, |y|\le\delta} |g_j(\varepsilon + y, x) - g_j(\varepsilon, x)|$$

is an envelope for the function class $\mathcal{F}_n = \{\varphi_{n,t} \mid t \in \mathcal{T}\}$ with finite second moment. Weak convergence of the process $(G_n(t))_{t\in\mathcal{T}}$ can now be shown by applying Theorem 2.11.23 by van der Vaart & Wellner (2000). To this end $\mathcal{T}$ is equipped with the semimetric ρ defined by the supremum norm, and one uses the bracketing

$$\Big[\frac{1}{\sqrt{h}}K(\frac{z-x}{h})\Big((g_j(\varepsilon+c_k(z),x)-g_j(\varepsilon,x))-\sup_{\substack{z,\tilde{z}\in\mathbb{R}:|z|\leq\delta,\\ |\tilde{z}|\leq\delta,|z-\tilde{z}|\leq\tilde{\epsilon}}}(g_j(\varepsilon+z,x)-g_j(\varepsilon+\tilde{z},x))\Big),$$
$$\frac{1}{\sqrt{h}}K(\frac{z-x}{h})\Big((g_j(\varepsilon+c_k(z),x)-g_j(\varepsilon,x))+\sup_{\substack{z,\tilde{z}\in\mathbb{R}:|z|\leq\delta,\\ |\tilde{z}|\leq\delta,|z-\tilde{z}|\leq\tilde{\epsilon}}}(g_j(\varepsilon+z,x)-g_j(\varepsilon+\tilde{z},x))\Big)\Big],$$

$k=1,\ldots,K$, for $\mathcal{F}_n$ with $L^2(P)$-length ϵ. Here $\tilde{\epsilon}=\epsilon^{\kappa/2}(4C^2||K||_\infty||f_X||_\infty)^{-\kappa/2}$ and $K=N(\tilde{\epsilon},\mathcal{T},||\cdot||_\infty)$, with κ and C from assumption (A.2) and where $c_1,\ldots,c_K$ denote the centers of balls with $||\cdot||_\infty$-radius $\tilde{\epsilon}$ used to cover $\mathcal{T}$. Using those brackets one can show that $\log N_{[]}(\epsilon||\Phi_n||_{2,P},\mathcal{F}_n,L_2(P))=O(\log N(\tilde{\epsilon},\mathcal{T},||\cdot||_\infty))=O(\epsilon^{-\kappa/(1+\alpha)})$, where the last equality follows from Theorem 2.7.1 by van der Vaart & Wellner (2000, p. 154). Details are omitted for the sake of brevity. Weak convergence of $G_n(\hat{t}_n)$ to $G(0)=0$ [and hence, (B.14)] now follows from Lemma B.1 (where $\mathcal{Y}$ has only one element).

From (B.12), (B.13) and (B.14) we obtain that

$$\begin{aligned}\hat{\eta}_n(x) &= \Sigma^{-1}(x)\Big[\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{X_i-x}{h}\Big)\,g(\varepsilon_i,x)\\ &\quad+\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)(g(y+(m-\hat{m})(z),x)-g(y,x))f_{\varepsilon,X}(y,z)\,d(y,z)\Big]+o_p(\frac{1}{\sqrt{nh}}).\end{aligned}$$

The assertion of the Proposition now follows from ($j\in\{1,\ldots,k\}$)

$$\begin{aligned}&\int\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)(g_j(y+(m-\hat{m})(z),x)-g_j(y,x))f_{\varepsilon,X}(y,z)\,dz\,dy\\ &=\int\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)g_j'(y,x)(m-\hat{m})(z)f_{\varepsilon,X}(y,z)\,dz\,dy+o_p(\frac{1}{\sqrt{nh}})\end{aligned}$$

[by assumption (A.1) and the bandwidth condition $h(\log b^{-1})^{1+\beta}/(n^\beta b^{\beta+1})=o(1)$], inserting the definition of $\hat{m}$,

$$\begin{aligned}&\int\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)g_j'(y,x)(m-\hat{m})(z)f_{\varepsilon,X}(y,z)\,dz\,dy\\ &=\ -\frac{1}{n}\sum_{i=1}^n\varepsilon_i\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)g_j'(y,x)\frac{1}{b}\tilde{K}\Big(\frac{X_i-z}{b}\Big)\frac{f_{\varepsilon,X}(y,z)}{\hat{f}_X(z)}\,d(y,z)\\ &\quad-\frac{1}{n}\sum_{i=1}^n\int\frac{1}{h}K\Big(\frac{z-x}{h}\Big)g_j'(y,x)\frac{1}{b}\tilde{K}\Big(\frac{X_i-z}{b}\Big)(m(X_i)-m(z))\frac{f_{\varepsilon,X}(y,z)}{\hat{f}_X(z)}\,d(y,z)\\ &\quad+o_p(\frac{1}{\sqrt{nh}}),\end{aligned}$$

replacing the density estimtor $\hat{f}_X$ in the denominator by the density f_X and calculations of expectations and variances as are typical for proofs in the context of kernel estimation. □

Proposition B.4 *Under the assumptions of Theorem 3.1 one has*

$$\hat{F}_n^*(y|x) - \hat{F}_n(y|x) \;=\; -\hat{\eta}_n^\top(x) E[g(\varepsilon, X) I\{\varepsilon \le y\} \mid X = x] \int K^2(u)\,du + o_p(\frac{1}{\sqrt{nh}})$$

uniformly with respect to $y \in \mathbb{R}$ for each $x \in (0,1)$.

Proof of Proposition B.4.
Let $\tilde{f}_X(x) = \sum_{j=1}^n h^{-1} K(\frac{X_j - x}{h}) p_j(x)$ and $\bar{f}_X(x) = (nh)^{-1} \sum_{j=1}^n K(\frac{X_j - x}{h})$ denote the density estimators. Then, by definition,

$$\begin{aligned}
&\hat{F}_n^*(y|x) - \hat{F}_n(y|x) \\
&= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\Big(\frac{X_i - x}{h}\Big) I\{\hat{\varepsilon}_i \le y\} \Big(\frac{1}{\tilde{f}_X(x)} \frac{1}{1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})} - \frac{1}{\bar{f}_X(x)} \Big) \\
&= A_n(y) + B_n(y) + C_n(y),
\end{aligned}$$

where

$$\begin{aligned}
A_n(y) &= -\hat{\eta}_n(x)^\top \frac{1}{nh} \sum_{i=1}^n K^2\Big(\frac{X_i - x}{h}\Big) I\{\hat{\varepsilon}_i \le y\} g(\hat{\varepsilon}_i, x) \frac{1}{\tilde{f}_X(x)} \\
B_n(y) &= \frac{1}{nh} \sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) I\{\hat{\varepsilon}_i \le y\} \frac{\bar{f}_X(x) - \tilde{f}_X(x)}{\bar{f}_X(x) \tilde{f}_X(x)} \\
C_n(y) &= \frac{1}{nh} \sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) I\{\hat{\varepsilon}_i \le y\} \frac{1}{\tilde{f}_X(x)} \frac{(\hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h}))^2}{1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_i, x) K(\frac{X_i - x}{h})}.
\end{aligned}$$

Now, by Proposition B.3 we have that $\hat{\eta}_n(x) = O_p((nh)^{-1/2})$ and this can be used to show that $C_n(y) = o_p((nh)^{-1/2})$ and

$$\begin{aligned}
A_n(y) &= -\hat{\eta}_n(x)^\top E\Big[K^2\Big(\frac{X - x}{h}\Big) I\{\varepsilon \le y\} g(\varepsilon, x)\Big] \frac{1}{f_X(x)} + o_p(\frac{1}{\sqrt{nh}}) \\
&= -\hat{\eta}_n(x)^\top \int K^2\Big(\frac{z - x}{h}\Big) E[I\{\varepsilon \le y\} g(\varepsilon, x) \mid X = z] \frac{f_X(z)}{f_X(x)}\,dz + o_p(\frac{1}{\sqrt{nh}}) \\
&= -\hat{\eta}_n(x)^\top \int K^2(u)\,du\, E[I\{\varepsilon \le y\} g(\varepsilon, x) \mid X = x] + o_p(\frac{1}{\sqrt{nh}}).
\end{aligned}$$

Uniformity of these results in $y \in \mathbb{R}$ can be obtained by some simple estimations and an application of Theorem 37 by Pollard (1984, p. 34) to bound the remainder term. For the sake of brevity technical details are omitted.

Further, by definition of $\bar{f}_X$ and $\tilde{f}_X$ with $\hat{F}_n$ from (2.1),

$$B_n(y) \;=\; \hat{F}_n(y|x) \frac{1}{\tilde{f}_X(x)} \frac{1}{nh} \sum_{j=1}^n K\Big(\frac{X_j - x}{h}\Big) \Big(1 - \frac{1}{1 + \hat{\eta}_n(x)^\top g(\hat{\varepsilon}_j, x) K(\frac{X_j - x}{h})}\Big)$$

$$
\begin{aligned}
&= \hat{\eta}_n(x)^\top \hat{F}_n(y|x) \frac{1}{nh} \sum_{j=1}^{n} K^2\Big(\frac{X_j - x}{h}\Big) g(\hat{\varepsilon}_j, x) \frac{1}{\tilde{f}_X(x)} + o_p(\frac{1}{\sqrt{nh}}) \\
&= \hat{\eta}_n(x)^\top \hat{F}_n(y|x) \Big(E\Big[\frac{1}{h} K^2\Big(\frac{X - x}{h}\Big) g(\varepsilon, x)\Big] \frac{1}{\tilde{f}_X(x)} + o_p(1)\Big) + o_p(\frac{1}{\sqrt{nh}}) \\
&= O_p(\frac{1}{\sqrt{nh}})(F_{\varepsilon|X}(y|x) + o_p(1))\Big(0 + o(1) + o_p(1)\Big) + o_p(\frac{1}{\sqrt{nh}}) \\
&= o_p(\frac{1}{\sqrt{nh}})
\end{aligned}
$$

uniformly with respect to $y \in \mathbb{R}$ by Theorem 2.1, Proposition B.3 and (1.5). □

Proof of Theorem 3.1.
(a) The expansion follows directly from Propositions B.3 and B.4. □

(b) The bias structure follows directly from part (a) together with Theorem 2.1 and the following calculation. Note that for $\phi(t|x) = E[g(\varepsilon, x) \mid X = t] f_X(t)$ we have $\phi(x|x) = 0$ by (1.5), and hence

$$
E\Big[\frac{1}{h} K\Big(\frac{X - x}{h}\Big) g(\varepsilon, x)\Big] \;=\; \int \frac{1}{h} K\Big(\frac{t - x}{h}\Big) \phi(t|x)\, dt \;=\; \frac{h^2}{2} \int K(u) u^2\, du \frac{\partial^2 \phi(t|x)}{\partial t^2}\Big|_{t=x} + o(h^2)
$$

from which the formula for $H^*(\cdot|x)$ in the Theorem follows.

Weak convergence follows similarly to the proof of Theorem 2.1 (b) and the covariance structure follows by tedious, but straightforward calculations. □

C References

M. Akritas and I. van Keilegom (2001). *Nonparametric estimation of the residual distribution.* Scand. J. Statist. 28, 549–567.

B. Antoine, H. Bonnal and E. Renault (2007). *On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood.* J. Econometrics 138, 431–487.

F. Cheng (2002). *Consistency of error density and distribution function estimators in nonparametric regression* Stat. Probab. Lett. 59, 257–270.

H. Dette (2002). *A consistent test for heteroscedasticity in nonparametric regression based on the kernel method.* J. Statist. Plann. Infer. 103, 311–329.

H. Dette and B. Hetzler (2009a). *A simple test for the parametric form of the variance function in nonparametric regression.* Ann. Instit. Statist. Math. 61, 861–886.

H. Dette and B. Hetzler (2009b). *Khmaladze Transformation of Integrated Variance Processes with Applications to Goodness-of-Fit Testing.* Math. Meth. of Statist. 18, 97–116.

H. Dette and A. Munk (1998). *Testing heteroscedasticity in nonparametric regression.* J. R. Statist. Soc. B 60, 693–708.

H. Dette, N. Neumeyer and I. Van Keilegom (2007). *A new test for the parametric form of the variance function in non-parametric regression.* J. R. Statist. Soc. B 69, 903–917.

H. Dette and S. Volgushev (2008). *Non-crossing non-parametric estimates of quantile curves.* J. R. Statist. Soc. B 70, 609–627.

T. DiCiccio, P. Hall and J. Romano (1991). *Empirical Likelihood is Bartlett-Correctable.* Ann. Statist. 19, 1053–1061.

S. Efromovich (2005). *Estimation of the density of regression errors.* Ann. Statist. 33, 2194–2217.

S. Efromovich (2007). *Adaptive estimation of error density in nonparametric regression with small sample size.* J. Statist. Plann. Infer. 137, 363–378.

J. Einmahl and I. McKeague (2003). *Empirical likelihood based hypothesis testing.* Bernoulli 9, 267–290.

J. Einmahl and I. Van Keilegom (2008a). *Tests for independence in nonparametric regression.* Statist. Sinica 18, 601–616.

J. Einmahl and I. Van Keilegom (2008b). *Specification tests in nonparametric regression.* J. Econometrics 143, 88–102.

P. Hall, R. C. L. Wolff and Q. Yao (1999). *Methods for Estimating a Conditional Distribution Function.* J. Am. Stat. Assoc. 94, 154–163.

W. Härdle (1989). *Applied nonparametric regression.* Cambridge University Press, New York.

W. Härdle and A. Bowman (1988). *Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands.* J. Amer. Stat. Ass. 83, 102–110.

W. Härdle and E. Mammen (1993). *Comparing Nonparametric Versus Parametric Regression Fits.* Ann. Statist. 21, 1926–1947.

W. Härdle and J.S. Marron (1991). *Bootstrap simultaneous error bars for nonparametric regression.* Ann. Statist. 19, 778–796.

W. Härdle, Y. Ritov and S. Song (2010). *Partial Linear Quantile Regression and Bootstrap Confidence Bands.* preprint available at `http://sfb649.wiwi.hu-berlin.de/fedc/discussionPapers_en.php`

L. Horwáth and B. S. Yandell (1988). *Asymptotics of Conditional Empirical Processes.* J. Mult. Anal. 26, 184–206.

E. V. Khmaladze (1981). *Martingale approach in the theory of goodness-of-fit test.* Th. Probab. Appl. 26, 240–257.

Y. Kitamura (1997). *Empirical Likelihood Methods with weakly dependent processes.* Ann. Statist. 25, 2084–2102.

S. Kiwitt, E.-R. Nagel and N. Neumeyer (2008). *Empirical Likelihod Estimators for the Error Distribution in Nonparametric Regerssion Models.* Math. Meth. Statist. 17, 241–260.

R. Koenker (2005). *Quantile Regressions.* Cambridge University Press.

M. R. Kosorok (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer.

H. Koul and W. Song (2010). *Conditional variance model checking.* J. Statist. Planning Infer. 140, 1056–1072.

H. Liero (2003). *Testing homoecedasticity in nonparametric regression.* J. Nonparametr. Statist. 15, 31–51.

E. Molanes Lopez, I. Van Keilegom and N. Veraverbeke (2009). *Empirical likelihood for non-smoth criterion functions.* Scand. J. Stat. 36, 413–432.

H.-G. Müller (1984). *Boundary Effects in Nonparametric Curve Estimation.* Compstat 1984, Physica Verlag, Heidelberg, 84–89.

U. U. Müller, A. Schick and W. Wefelmeyer (2004). *Estimating linear functionals of the error distribution in nonparametric regression.* J. Statist. Plann. Inf. 119, 75–93.

U. U. Müller, A. Schick and W. Wefelmeyer (2007). *Estimating the error distribution function in semiparametric regression.* Statist. Decisions 25, 1–18.

U. U. Müller, A. Schick and W. Wefelmeyer (2009). *Estimating the error distribution function in semiparametric regression with multivariate convariates.* Stat. Probab. Lett. 79, 957–964.

É. A. Nadaraya (1964). *On non–parametric estimates of density functions and regression curves.* J. Probab. Appl. 10, 186–190.

N. Neumeyer (2008). *A bootstrap version of the residual-based smooth empirical distribution function.* J. Nonparam. Statist. 20, 153–174.

N. Neumeyer (2009). *Testing independence in nonparametric regression.* J. Mult. Anal. 100, 1551–1566.

A. B. Owen (1988). *Empirical Likelihood ratio confidence intervals for a single functional.* Biometrika 75, 2, 237–249.

A. B. Owen (2001). *Empirical Likelihood.* Chapman & Hall/CRC.

D. Pollard (1984). *Convergence of Stochastic Processes.* Springer.

J. Qin and J. Lawless (1994). *Empirical likelihood and general estimating equations.* Ann. Statist. 22, 300–325.

J. Shao and D. Tu (1995). *The Jackknife and bootstrap.* Springer.

W. Stute (1986). *On almost Sure Convergence of Conditional Empirical Distribution Functions.* Ann. Probab. 14, 891–901.

W. Stute, W. Gonzalez Manteiga and M. Presedo Quindimil (1998). *Bootstrap Approximations in Model Checks for Regression.* J. Am. Statist. Assoc. 93, 141–149.

A. W. van der Vaart (1998). *Asymptotic Statistics.* Cambridge University Press.

A. W. van der Vaart and J. A. Wellner (2000). *Weak Convergence and Empirical Processes.* Springer Series in Statistics.

G. S. Watson (1964). *Smooth Regression Analysis.* Sankhya A 26, 359–372.

L. Zhu, Y. Fujikoshi and K. Naito (2001). *Heteroscedasticity checks for regression models.* Science in China (Series A) 44, 1236–1252.

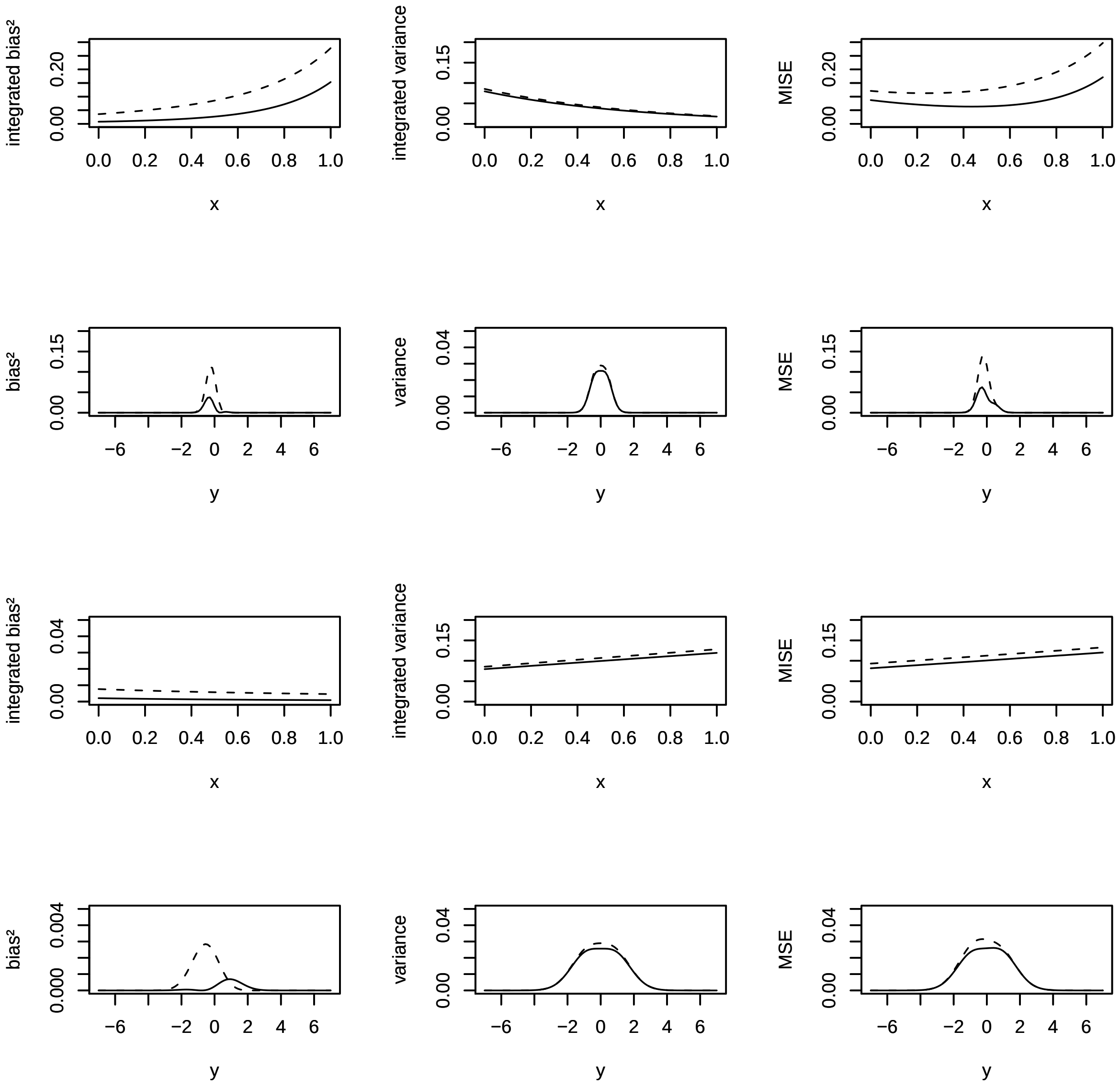

Figure 1: For the example $g(\varepsilon, x) = \varepsilon$ the figure shows comparisons of the integrated squared bias (left panel), variance (middle panel) and integrated mean squared error (right panel) as functions in x (first and third row) as well as the squared bias (left), variance (middle) and mean squared error (right) for fixed $x = 0.5$ as function in y (second and fourth row) for $\hat{F}_n$ (dashed lines) and $\hat{F}_n^*$ (solid lines). The results are for $X \sim U[0,1]$, $\varepsilon \sim N(0, \sigma^2(x))$ and $m(x) = x^2$. The variance function is $\sigma^2(x) = e^{-3x}$ in the first two rows and $\sigma^2(x) = (1 + 0.5x)^2$ in the last two rows.

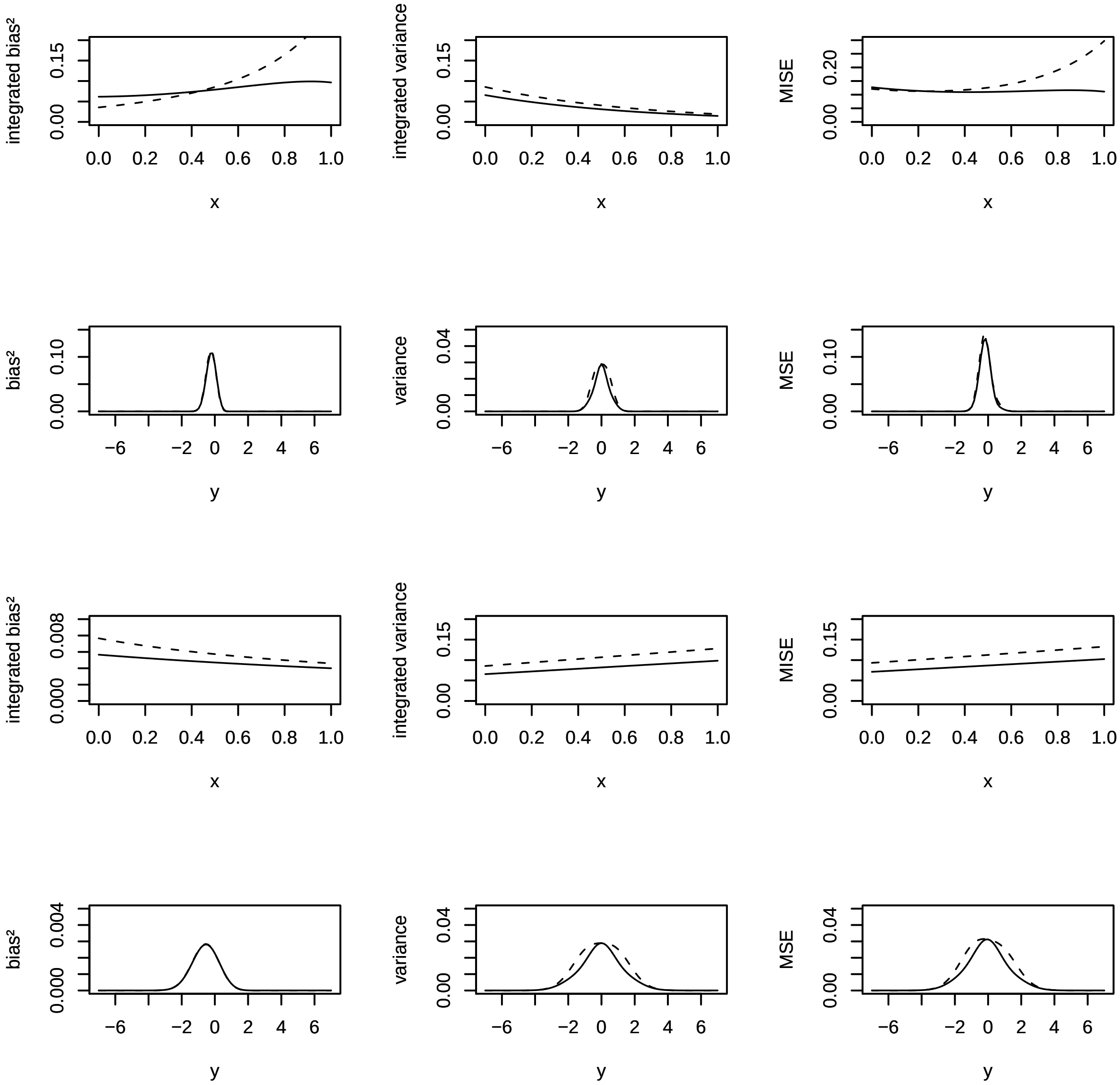

Figure 2: The figure shows curves as described in the caption of figure 1, but for $g(\varepsilon, x) = \varepsilon^2 - \sigma^2(x)$.

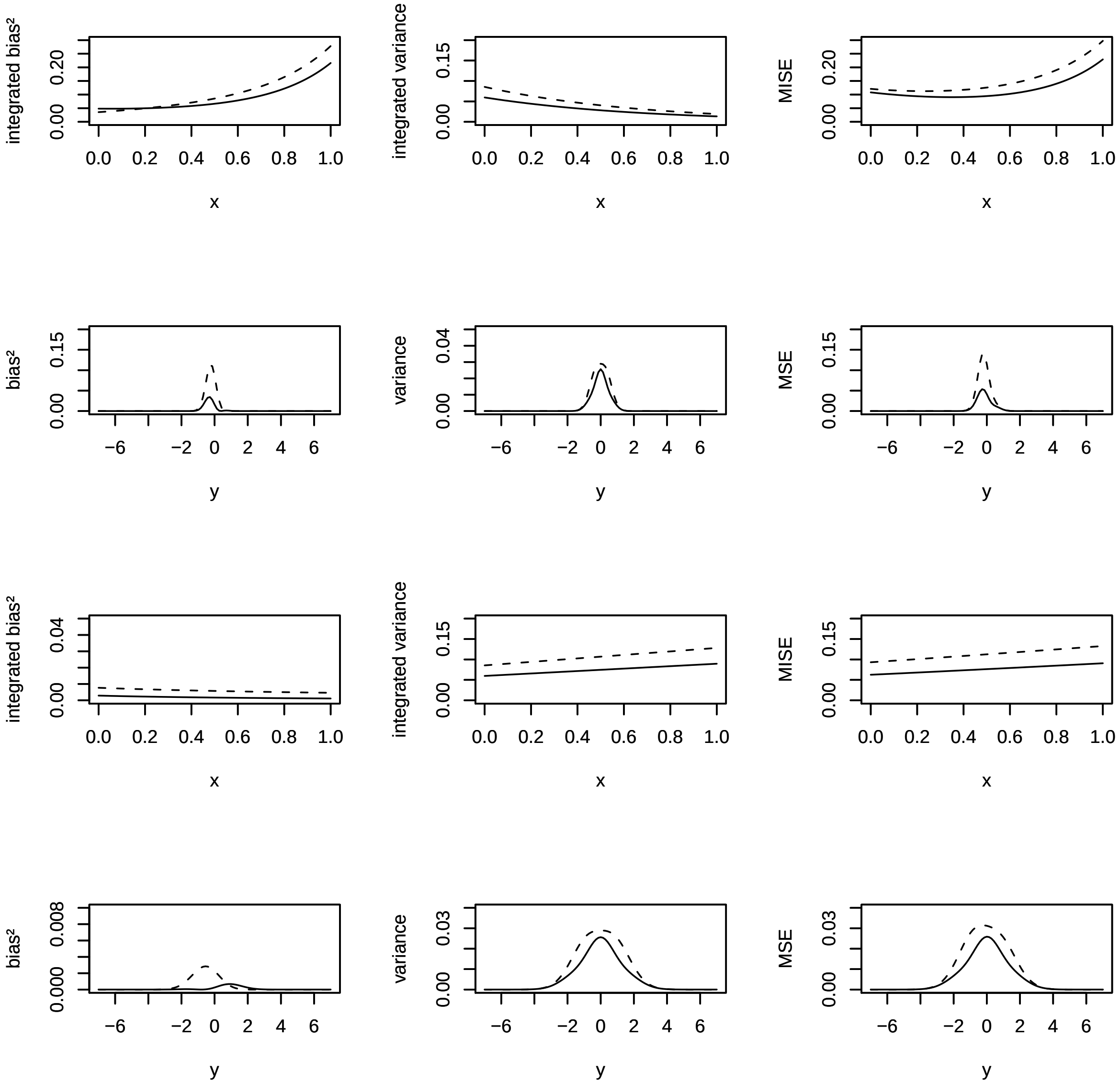

Figure 3: The figure shows curves as described in the caption of figure 1, but for $g(\varepsilon, x) = (\varepsilon, \varepsilon^2 - \sigma^2(x))^\top$.

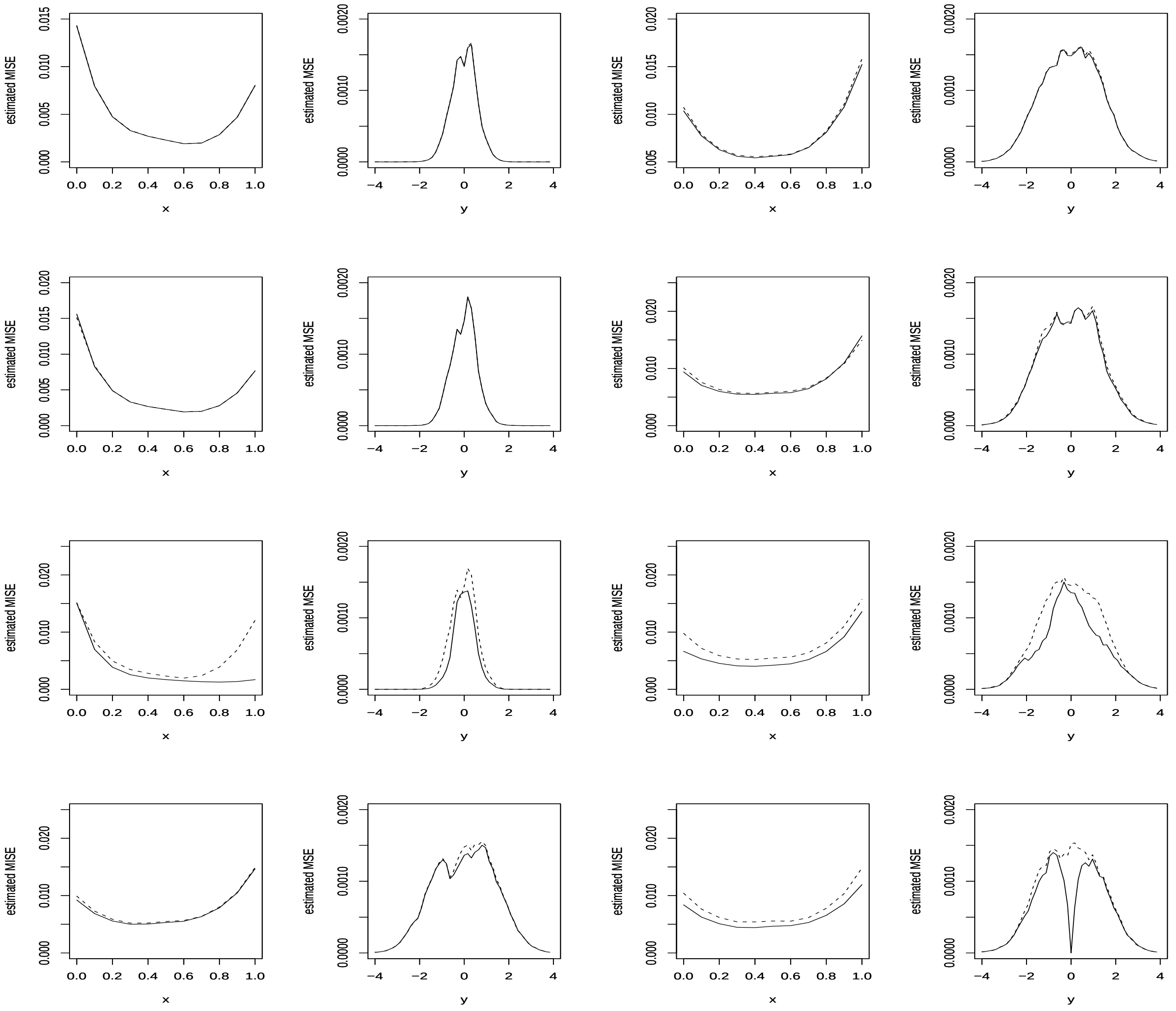

Figure 4: The figure shows results from simulations for the examples $g(\varepsilon, x) = \varepsilon$ (first row), $g(\varepsilon, x) = \varepsilon^2 - \sigma^2(x)$ (second row), $g(\varepsilon, x) = (\varepsilon, \varepsilon^2 - \sigma^2(x))^\top$ (third row), $g(\varepsilon, x) = (\varepsilon, \varepsilon^3, \varepsilon^5)$ (last row, first two panels), and $g(\varepsilon, x) = (\varepsilon, I\{\varepsilon \leq 0\} - 0.5)$ (last row, last two panels). For sample size $n = 100$ data were generated according to $X \sim U[0, 1]$, $\varepsilon \sim N(0, \sigma^2(x))$ and $m(x) = x^2$, where $\sigma^2(x) = e^{-3x}$ (first two columns in first three rows) and $\sigma^2(x) = (1+0.5x)^2$ (last two columns in first three rows, and in the last row). The panels in the first and third column show the simulated mean integrated squared error as function in x, whereas the remaining panels show the mean squared error as function in y for fixed $x = 0.5$. The dashed lines correspond to $\hat{F}_n$ and the solid lines to $\hat{F}_n^*$. The results are based on 500 simulation runs.