Empirical likelihood estimators for the error distribution in nonparametric regression models

Sebastian Kiwitt, Eva-Renate Nagel and Natalie Neumeyer*

Fakultät für Mathematik, Ruhr-Universität Bochum

44780 Bochum, Germany

*corresponding author, email: natalie.neumeyer@rub.de

October 13, 2005

Abstract

The aim of this paper is to show that existing estimators for the error distribution in nonparametric regression models can be improved when additional information about the distribution is included by the empirical likelihood method. The weak convergence of the resulting new estimator to a Gaussian process is shown and the performance is investigated by comparison of asymptotic mean squared errors and by means of a simulation study. As a by-product of our proofs we obtain stochastic expansions for smooth linear estimators based on residuals from the nonparametric regression model.

Short title: Empirical likelihood for regression errors

AMS Classification: 62G08, 62G05

Keywords and Phrases: empirical distribution function, empirical likelihood, error distribution, estimating function, nonparametric regression, Owen estimator

1 Introduction

Since a few decades in statistical research nonparametric regression models have been investigated intensively. We consider such a model,

$$Y_i = m(X_i) + \varepsilon_i \;\; (i = 1, \dots, n),$$

with independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ and centered, unobserved, independent and identically distributed errors $\varepsilon_1, \ldots, \varepsilon_n$ (independent from the design points $X_1, \ldots, X_n$). In the last decades research focused mainly on nonparametric estimation of the regression function m and variance $\sigma^2 = E[\varepsilon_1^2]$ and corresponding hypotheses tests. Since a few years only there exist results on estimation of the distribution of the unobserved errors $\varepsilon_1, \ldots, \varepsilon_n$. For example, consistent estimators for the regression function and error distribution can be used to evaluate prediction intervals for future observations at some point x [see Akritas and Van Keilegom (2001)]. Further the empirical distribution function of estimated errors recently turned out to be valuable for goodness-of-fit tests concerning the regression or variance function, see Van Keilegom, González Manteiga and Sánchez Sellero (2004) and Dette and Van Keilegom (2005), or for testing the equality of regression functions in a two–sample problem, see Pardo-Fernández, Van Keilegom and González-Manteiga (2004) and Neumeyer and Dette (2005). We denote by F_n the (not available) empirical distribution function of unobserved errors. Classical results by Donsker (1952) show weak convergence of the empirical process $\sqrt{n}(F_n - F)$ to a Brownian bridge B with covariance structure

$$\text{(1.1)} \qquad \mathrm{Cov}(B(y), B(z)) \;=\; F(y \wedge z) - F(y)F(z).$$

Now, let $\hat{F}_n$ denote the empirical distribution function of nonparametrically estimated residuals, i.e.

$$\text{(1.2)} \qquad \hat{F}_n(y) \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{\hat{\varepsilon}_i \le y\}, \quad y \in \mathbb{R},$$

where the residuals are defined as $\hat{\varepsilon}_i = Y_i - \hat{m}(X_i)$, and $\hat{m}$ denotes the Nadaraya–Watson estimator for the regression function m. The asymptotic behavior of $\hat{F}_n$ has been investigated by Akritas and Van Keilegom (2001) and Cheng (2002). A smooth version of $\hat{F}_n$ was considered by Müller, Schick and Wefelmeyer (2004c). Cheng (2004), Qin (1996), and Qin, Shi and Chai (1996) propose corresponding error density estimators. In the heteroscedastic model considered by Akritas and Van Keilegom (2001) the regression and variance function are defined as L-functionals depending on some score function J. Slight adaptations of Akritas and Van Keilegom's (2001) arguments for a homoscedastic regression model and score function $J \equiv 1$ show under some regularity assumptions (that are valid under the assumptions stated in section 2 of the presented paper) that the process $\sqrt{n}(\hat{F}_n(\cdot) - F(\cdot) - b(\cdot))$ converges weakly to a Gaussian process G with covariance structure

$$\text{(1.3)} \quad \mathrm{Cov}(G(y), G(z)) \;=\; F(y \wedge z) - F(y)F(z) + E[\varepsilon_1^2]f(y)f(z) + E[\varepsilon_1 I\{\varepsilon_1 \le y\}]f(z) + E[\varepsilon_1 I\{\varepsilon_1 \le z\}]f(y),$$

where f denotes the error density and the bias term is defined as

$$b(y) \;=\; h^2 f(y)\frac{1}{2}\int K(u)u^2\,du \int ((mf_X)''(x) - (mf_X'')(x))\,dx. \tag{1.4}$$

Here K denotes a kernel function and h a bandwidth used for the construction of the kernel estimator $\hat{m}$ and f_X denotes the density of the design points. We give a short derivation of this result in appendix A. One notices that the covariance structure of the asymptotic process G given in (1.3) differs from the covariance structure of the asymptotic process B, see (1.1). For Gaussian errors considering nonparametric residuals instead of true errors even results in a uniformly smaller asymptotic variance,

$$\mathrm{Var}(G(y)) \;\leq\; \mathrm{Var}(B(y)) \quad \forall y \in \mathbb{R},$$

but the estimation is biased then. We want to investigate whether the estimation of the error distribution can further be improved in terms of mean squared error when additional information is used. Our most important example for additional information is the centeredness of the errors, i. e. $E[\varepsilon_1] = 0$, that is required by the model but is not explicitely used in the estimation $\hat{F}_n$. Further examples for additional information are a known variance or median. Improvements of the estimator by including additional information could be obtained by the Empirical Likelihood method that was introduced by Owen (1988, 2001) and further developed by Hall and LaScala (1990), DiCiccio, Hall and Romano (1989, 1991), DiCiccio and Romano (1989), Hall (1990), Kitamura (1997), Einmahl and McKeague (2003), among many others. Qin and Lawless (1994) and Zhang (1997) considered the problem of estimating the distribution of an (observed) iid-sample $\varepsilon_1, \ldots, \varepsilon_n$ when auxiliary information is available in terms of

$$E[g(\varepsilon_1)] = \int g(y)\,dF(y) = 0,$$

where $g = (g_1, \ldots, g_k)^T : \mathbb{R} \to \mathbb{R}^k$ is a known function such that $E[g_j^2(\varepsilon_1)] < \infty$ $(j = 1, \ldots, k)$. This could be, for example, $g(\varepsilon) = \varepsilon$ for the model assumption of centered errors, $g(\varepsilon) = (\varepsilon, \varepsilon^2 - \sigma^2)^T$ for centered errors with a known variance σ^2 or $g(\varepsilon) = I\{\varepsilon \leq q\} - \frac{1}{2}$ for a priori information about the median q. The empirical likelihood estimator for the error distribution is

$$\tilde{F}_n(y) \;=\; \sum_{i=1}^{n} p_i I\{\varepsilon_i \leq y\}, \quad y \in \mathbb{R}, \tag{1.5}$$

where the weights $p_i \in (0, 1)$ are chosen such that the empirical likelihood

$$\prod_{i=1}^{n} p_i \tag{1.6}$$

is maximized under the constraints

$$\sum_{i=1}^{n} p_i = 1, \quad \int g(y)\, d\tilde{F}_n(y) = \sum_{i=1}^{n} p_i g(\varepsilon_i) = 0. \tag{1.7}$$

Qin and Lawless (1994) and Zhang (1997) showed that in this setting the obtained empirical likelihood estimator has a uniformly smaller asymptotic variance than the empirical distribution function. For empirical likelihood and moment restrictions see also Kitamura (2001), Kitamura, Tripathi, Ahn (2004), and Bonnal and Renault (2004), among others.

The empirical likelihood method was applied in the context of estimation of the error distribution in linear models with fixed design and homoscedastic errors by Nagel (2002) using the additional information $E[\varepsilon_1] = 0$, i. e. $g(\varepsilon) = \varepsilon$, that is available from the model. In this context it depends on the method of parameter estimation whether the empirical likelihood method yields a smaller asymptotic variance than the empirical distribution function based on parametric residuals. A comprehensive study of the residual based empirical distribution functions in the linear model can be found in Koul (2002). Estimators for the error distribution in AR(1)-models including the centeredness assumption were considered by Genz (2004).

In the presented paper we propose a residual based empirical likelihood method for the error distribution in nonparametric regression models when auxiliary information is available. We develop asymptotic expansions for the empirical likelihood estimator, $\overline{F}_n$, in our context and prove weak convergence of the process $\sqrt{n}(\overline{F}_n(\cdot) - F(\cdot) - \overline{b}(\cdot))$ to a Gaussian process $\overline{G}$, where $\overline{b}$ denotes a bias term. We compare the resulting asymptotic mean squared error of the new estimator,

$$\frac{\mathrm{Var}(\overline{G}(y))}{n} + \overline{b}^2(y),$$

with the analogous term for the residual based empirical distribution function, $\mathrm{Var}(G(y))/n + b^2(y)$ defined in (1.3) and (1.4), in some examples. In the especially interesting case of $g(\varepsilon) = \varepsilon$ (i. e. including the model assumption of centered errors explicitely into the estimation) we obtain asymptotically the same variances but a considerable reduction of bias.

As a by-product of our proofs we regain asymptotic expansions for smooth linear statistics

$$\frac{1}{n}\sum_{i=1}^{n} g(\hat{\varepsilon}_i)$$

in a similar homoscedastic setting considered by Müller, Schick and Wefelmeyer (2004a) and prove analogous results for heteroscedastic models.

The paper is organized as follows. In section 2 we describe our (homoscedastic) model and explain regularity assumptions needed to obtain the main asymptotic results presented in section

3. In section 4 we consider the analogous procedures in a heteroscedastic setting. In section 5 some examples are illustrated in order to discuss the results and in section 6 a simulation study is presented. The proofs are deferred to an appendix.

2 Model and assumptions

We first consider a nonparametric homoscedastic regression model with independent observations

$$Y_i \;=\; m(X_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{2.1}$$

under the following assumptions.

(M1) The univariate design points $X_1, \dots, X_n$ are independent and identically distributed with distribution function F_X on compact support, say $[0,1]$. F_X has a twice continuously differentiable density f_X, such that $\inf_{x\in[0,1]} f_X(x) > 0$. The regression function m is twice continuously differentiable in $(0,1)$ with bounded derivatives.

(M2) The errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with distribution function F. They are centered, $E[\varepsilon_1] = 0$, with variance $\sigma^2 = \mathrm{Var}(\varepsilon_1) \in (0,\infty)$, and independent from the design points. F is continuously differentiable with bounded, everywhere positive density f.

(M3) There exist constants γ, C and $\beta > 0$ such that for all $z \in \mathbb{R}$ with $|z| \le \gamma$

$$|F(y+z) - F(y) - zf(y)| \;\le\; C|z|^{1+\beta}.$$

Note that assumption (M3) is satisfied with $\beta = 1$ under the stronger assumption that f is continuously differentiable with $\sup_{y\in\mathbb{R}} |f'(y)| < \infty$.

In order to estimate the distribution F of the unobserved errors, one builds nonparametric residuals

$$\hat\varepsilon_i \;=\; Y_i - \hat m(X_i), \quad i = 1, \dots, n,$$

where $\hat m(x)$ denotes the Nadaraya-Watson estimator [Nadaraya (1964), Watson (1964)] for $m(x)$, that is

$$\hat m(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{X_i - x}{h}\Big) Y_i \frac{1}{\hat f_X(x)} \tag{2.2}$$

with the kernel density estimator for $f_X(x)$,

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\Big(\frac{X_i - x}{h}\Big). \tag{2.3}$$

For the kernel estimators we need the following assumptions.

(K) Let K denote a symmetric density with compact support and $\int uK(u)\,du = 0$.

(H) Let $h = h_n$ be a sequence of bandwidths such that $nh^4 = O(1)$, $nh^{3+2\alpha}(\log(h^{-1}))^{-1} \to \infty$ (for some $\alpha > 0$) and $n^\beta h^{1+\beta}(\log(h^{-1}))^{-1-\beta} \to \infty$ (where β is defined in assumption (M3)) for $n \to \infty$.

Note that the last bandwidth condition can be omitted when $\frac{1+\beta}{\beta} \le 3 + 2\alpha$. This is always valid for $\beta \ge \frac{1}{2}$. The constant α has only relevance in the technics of the proof and can be chosen arbitrarily small.
We denote by $\hat{F}_n$ the empirical distribution function based on residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$ defined in (1.2). Assumptions (M1), (M2), (M3), (K) and (H) were already imposed by Akritas and Van Keilegom (2001) to show weak convergence of the residual based empirical process, $\sqrt{n}(\hat{F}_n - F - b)$ [where the bias b is defined in (1.4)].
We further assume that additional information about the error distribution is available. This auxiliary information is given in terms of assumption (A).

(A) $E[g(\varepsilon_1)] = \int g(y)f(y)\,dy = 0$, where $g = (g_1, \ldots, g_k)^T : \mathbb{R} \to \mathbb{R}^k$ is a known function such that $E[g_j^2(\varepsilon_1)] < \infty$ for $j = 1, \ldots, k$.

Example 2.1 The most important example to consider is $g(\varepsilon) = \varepsilon$ because the centeredness of the errors is a given model assumption. Further a priori information, for instance, of a zero median can be described by the function $g(\varepsilon) = I\{\varepsilon \le 0\} - 1/2$. When can be assumed, for example, that the variance is known and the third moment is zero one would define $g(\varepsilon) = (\varepsilon, \varepsilon^2 - \sigma^2, \varepsilon^3)^T$.

Throughout the paper the following assumptions on the function g are only assumed to be valid when stated explicitely.

(G1) We assume that g_j is continuously differentiable and there exist constants γ, C and $\beta > 0$ such that

$$\Big| \int \big(g_j(y+z) - g_j(y) - zg_j'(y)\big)\, f(y)\,dy \Big| \le C|z|^{1+\beta}$$

for all $z \in \mathbb{R}$ with $|z| \le \gamma$, $j = 1, \ldots, k$. Moreover, let $E[|g_j'(\varepsilon_1)|] < \infty$, $j = 1, \ldots, k$.

[Without restriction we use the same constants γ, C and β as in assumption (M3).]

(G2) There exist constants δ, C such that for some positive $\kappa < 2(1+\alpha)$

$$\Big(E\Big[\sup_{\substack{z,\tilde{z}\in\mathbb{R}:|z|\leq\delta,\\ |\tilde{z}|\leq\delta,|z-\tilde{z}|\leq\xi}} (g_j(\varepsilon_1+z)-g_j(\varepsilon_1+\tilde{z}))^2\Big]\Big)^{1/2} \leq C\xi^{1/\kappa}$$

(where α is defined in assumption (H)), $j = 1,\ldots,k$.

Note that (G2) is, for example, satisfied for $\kappa = 1$ by Taylor's expansion when g_j is continuously differentiable and $E[\sup_{z\in\mathbb{R}:|z|\leq\delta}(g_j'(\varepsilon_1+z))^2] < \infty$ ($j = 1,\ldots,k$). Then $\kappa = 1 < 2(1+\alpha)$ is always valid ($\alpha > 0$). The smoothness assumptions (G1) and (G2) are similar to the assumptions imposed by Müller, Schick and Wefelmeyer (2004a, Assumption 1, p. 79, and 2004b, Assumption B1, p. 536) to obtain asymptotic results about smooth linear estimators based on nonparametric residuals, compare the discussion of the assumptions given there [Müller, Schick and Wefelmeyer (2004a, section 3)]. Note also that either with assumption (G2) or for the indicator function $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ it follows

$$\exists\delta > 0 \text{ such that } E\Big[\sup_{y\in\mathbb{R}:|y|\leq\delta}(g_j(\varepsilon_1+y)-g_j(\varepsilon_1))^2\Big] < \infty\ \forall j = 1,\ldots,k \tag{2.4}$$

and this condition will be used repeatedly during the proof.

Example 2.2 Functions $g(\varepsilon) = \varepsilon^k - c$ corresponding to moment assumptions fulfill assumptions (G1) and (G2) when $E[\varepsilon_1^{2k}] < \infty$. The same is valid for polynomials, for example, $g(\varepsilon) = \varepsilon^4 - c\varepsilon^2$, to account for a relation between second and fourth moment.

Remark 2.3 Assumptions (G1) and (G2) are mainly imposed to obtain stochastic expansions of $\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i)$. In addition to smooth functions g satisfying (G1) and (G2) or indicator functions $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ the theory can be developed for every function g, such that an expansion $\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^n (g(\varepsilon_i)+h(\varepsilon_i)) + o_P(\frac{1}{\sqrt{n}})$ is valid with some weak assumptions on the function h (compare Lemma B.1 (ii), (iii) in the appendix).

We further need the following assumptions to assure a unique solution in the maximization of the empirical likelihood.

(S1) We assume that $\min_{1\leq i\leq n} g_j(\hat{\varepsilon}_i) < 0 < \max_{1\leq i\leq n} g_j(\hat{\varepsilon}_i)$ for all $j = 1,\ldots,k$.

(S2) We assume that $\Sigma = E[g(\varepsilon_1)g(\varepsilon_1)^T]$ and $\sum_{i=1}^n g(\hat{\varepsilon}_i)g(\hat{\varepsilon}_i)^T$ are positive definite.

Note that assumption (S1) is valid in probability for an increasing sample size because of assumption (A) when the residuals $\hat{\varepsilon}_i$ are replaced by the true errors ε_i. Further, the first assumption of (S2) with Lemma B.1 (v) (in the appendix) implies the second assumption of (S2) for increasing sample size, in probability.

3 Main asymptotic results

The motivation of the empirical likelihood method is as explained in the introduction [compare (1.5)–(1.7)]. The estimator for the error distribution is

$$\overline{F}_n(y) \;=\; \sum_{i=1}^n p_i I\{\hat{\varepsilon}_i \le y\},$$

where the weights $p_i \in (0,1)$ are chosen such that $\prod_{i=1}^n p_i$ is maximized under the constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i g(\hat{\varepsilon}_i) = 0.$$

Analogously to Qin and Lawless (1994) we obtain the estimator

$$\overline{F}_n(y) \;=\; \frac{1}{n}\sum_{i=1}^n \frac{1}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)} I\{\hat{\varepsilon}_i \le y\}, \tag{3.1}$$

where $\hat{\eta}_n$ is defined as solution of the equation

$$\sum_{i=1}^n \frac{g(\hat{\varepsilon}_i)}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)} \;=\; 0 \tag{3.2}$$

while for all $i = 1,\dots,n$ it holds that $1+\hat{\eta}_n^T g(\hat{\varepsilon}_i) > \frac{1}{n}$. The following two propositions give stochastic expansions for the solution $\hat{\eta}_n$ as well as for the distribution estimator $\overline{F}_n$.

Proposition 3.1 *Under model (2.1) and assumptions (M1), (M2), (M3), (K), (H), (A), (S1), (S2) and with either $g(\varepsilon) = I\{\varepsilon \le a\} - b$ or g satisfying (G1), (G2) we have the expansion*

$$\hat{\eta}_n = \Sigma^{-1}\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i) + o_P(\frac{1}{\sqrt{n}}).$$

Proposition 3.2 *Under model (2.1) and assumptions (M1), (M2), (M3), (K), (H), (A), (S1), (S2) and with either $g(\varepsilon) = I\{\varepsilon \le a\} - b$ or g satisfying (G1), (G2) we have uniformly with respect to $y \in \mathbb{R}$,*

$$\overline{F}_n(y) = \hat{F}_n(y) - U(y)^T\hat{\eta}_n + o_P(\frac{1}{\sqrt{n}}),$$

where $\hat{F}_n$ denotes the residual based empirical distribution function defined in (1.2) and $U(y) = E[g(\varepsilon_1)I\{\varepsilon_1 \le y\}]$.

We state our main results for two different cases of additional information in the Theorems 3.3 and 3.7, namely smooth functions g that satisfy assumptions (G1) and (G2) resp. indicator functions that give quantile informations.

Theorem 3.3 *Under model (2.1) and assumptions (M1), (M2), (M3), (K), (H), (A), (G1), (G2), (S1) and (S2) we have uniformly in $y \in \mathbb{R}$ the expansion*

$$\overline{F}_n(y) = \overline{b}(y) + \frac{1}{n}\sum_{i=1}^{n}\Big[I\{\varepsilon_i \le y\} + f(y)\varepsilon_i - U(y)^T\Sigma^{-1}(g(\varepsilon_i) - E[g'(\varepsilon_1)]\varepsilon_i)\Big] + o_P(\frac{1}{\sqrt{n}})$$

where the bias term is defined as $\overline{b}(y) = h^2 B(f(y) + U(y)^T\Sigma^{-1}E[g'(\varepsilon_1)])$ *with*

$$B = \frac{1}{2}\int K(u)u^2\,du \int ((mf_X)''(x) - mf_X''(x))\,dx,$$

and $U(y)$, Σ *are defined in Proposition 3.2 and assumption (S2), respectively. The process* $\sqrt{n}(\overline{F}_n(\cdot) - F(\cdot) - \overline{b}(\cdot))$ *converges weakly to a centered Gaussian process* $\overline{G}$ *with covariance structure*

$$\begin{aligned}
\mathrm{Cov}(\overline{G}(y), \overline{G}(z)) \;=\; & F(y\wedge z) - F(y)F(z) + f(y)f(z)\sigma^2 \\
& + f(y)E[\varepsilon_1 I\{\varepsilon_1 \le z\}] + f(z)E[\varepsilon_1 I\{\varepsilon_1 \le y\}] \\
& + U(y)^T\Sigma^{-1}\Big(\Sigma - E[g'(\varepsilon_1)]E[\varepsilon_1 g^T(\varepsilon_1)] \\
& \qquad - E[\varepsilon_1 g(\varepsilon_1)]E[g'(\varepsilon_1)]^T + E[g'(\varepsilon_1)]\sigma^2 E[g'(\varepsilon_1)]^T\Big)\Sigma^{-1}U(z) \\
& - U^T(y)\Sigma^{-1}\Big(U(z) - E[g'(\varepsilon_1)]E[\varepsilon_1 I\{\varepsilon_1 \le z\}]\Big) \\
& - U^T(z)\Sigma^{-1}\Big(U(y) - E[g'(\varepsilon_1)]E[\varepsilon_1 I\{\varepsilon_1 \le y\}]\Big) \\
& - \Big(f(z)U(y)^T + f(y)U(z)^T\Big)\Sigma^{-1}\Big(E[\varepsilon_1 g(\varepsilon_1)] - E[g'(\varepsilon_1)]\sigma^2\Big).
\end{aligned}$$

The proof of Theorem 3.3 is given in appendix C. It is not true that for all distributions uniformly in y the asymptotic variance of the empirical likelihood estimator is smaller than the asymptotic variance of the residual based empirical distribution function as it is the case for an observed iid-sample $\varepsilon_1, \ldots, \varepsilon_n$. Different functions g and underlying distributions F have to be investigated. Also, bias and variance have to be taken into account simultaneously for the comparison as we will do in the discussion of the asymptotic results in section 5.

Remark 3.4 Note that under a stronger bandwidth condition $nh^4 = o(1)$ the bias term $\overline{b}(y)$ in Theorem 3.3 is negligible. It can be seen from the expansion stated in the Theorem that incorporating the auxiliary information about the error distribution does not lead to a smaller asymptotic variance in the case $g(\varepsilon) = E[g'(\varepsilon)]\varepsilon$ because then $\overline{F}_n = \hat{F}_n + o_P(n^{-1/2})$. Then, using the auxiliary information that the errors are centered, that is $g(\varepsilon) = \varepsilon$, does not change the variance asymptotically. A heuristic explanation for this phenomenon was given by Müller, Schick and Wefelmeyer (2004a) in the similar context of linear smooth residual based estimators by the statement that the mean zero information is already used for estimating $\hat{\varepsilon}_i$. However, the bias terms of order h^2 should also be taken into account under the less restrictive bandwidth condition $nh^4 = O(1)$. The bias changes from $b(y)$ defined in (1.4) to $\overline{b}(y)$ when using empirical likelihood and the latter term can be considerably smaller as will be discussed in section 5.

Corollary 3.5 *Under the assumptions of Theorem 3.3 we have* $\mathrm{Var}(\overline{G}(y)) = \mathrm{Var}(G(y))$ *for all* $y \in \mathbb{R}$ *if and only if* $g(\varepsilon) = c\varepsilon$ *for some* $c \in \mathbb{R}$.

Remark 3.6 As a by-product of the proof of Theorem 3.3 we obtain the following asymptotic expansion for smooth linear estimators based on nonparametric residuals,

$$\frac{1}{n}\sum_{i=1}^{n} g(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^{n}(g(\varepsilon_i) - E[g'(\varepsilon_1)]\varepsilon_i) - h^2 E[g'(\varepsilon_1)]B + o_P(\frac{1}{\sqrt{n}})$$

where B is defined in Theorem 3.3 [compare Lemma B.1 (ii) in the appendix]. This completes results given by Müller, Schick and Wefelmeyer (2004a, 2004b), who considered more restrictive bandwidth conditions to neglect the bias and used a leave-one-out local polynomial estimator for the regression function.

Next we state our asymptotic results for indicator functions g that include additional information about quantiles. The proof of Theorem 3.7 is given in appendix C.

Theorem 3.7 *Under model (2.1) and assumptions (M1), (M2), (M3), (K), (H), (A), (S1) and (S2) where* $g(\varepsilon) = I\{\varepsilon \le a\} - b$ *for some constants* a *and* $b = F(a) \notin \{0, 1\}$ $(k = 1)$ *we have uniformly in* $y \in \mathbb{R}$ *the expansion*

$$\overline{F}_n(y) = \overline{b}(y) + \frac{1}{n}\sum_{i=1}^{n}\Big[I\{\varepsilon_i \le y\} + f(y)\varepsilon_i - U(y)\frac{1}{b(1-b)}(I\{\varepsilon_i \le a\} - b + f(a)\varepsilon_i)\Big] + o_P(\frac{1}{\sqrt{n}})$$

where the bias term is $\overline{b}(y) = h^2 B(f(y) - \Sigma^{-1}U(y)f(a))$ *with* B *defined in Theorem 3.3. The process* $\sqrt{n}(\overline{F}_n(\cdot) - F(\cdot) - \overline{b}(\cdot))$ *converges weakly to a centered Gaussian process* $\overline{G}$ *with covariance structure*

$$\begin{aligned}
\mathrm{Cov}(\overline{G}(y), \overline{G}(z)) \; = \; & F(y \wedge z) - F(y)F(z) + f(y)f(z)\sigma^2 \\
& + f(y)E[\varepsilon_1 I\{\varepsilon_1 \le z\}] + f(z)E[\varepsilon_1 I\{\varepsilon_1 \le y\}] \\
& + (b(1-b))^{-2}U(z)U(y)\Big(b(1-b) + 2f(a)E[\varepsilon_1 I\{\varepsilon_1 \le a\}] + f^2(a)\sigma^2\Big) \\
& - (b(1-b))^{-1}U(z)\Big(F(a \wedge y) - bF(y) + f(a)E[\varepsilon_1 I\{\varepsilon_1 \le y\}]\Big) \\
& - (b(1-b))^{-1}U(y)\Big(F(a \wedge z) - bF(z) + f(a)E[\varepsilon_1 I\{\varepsilon_1 \le z\}\Big) \\
& - (b(1-b))^{-1}\Big(f(z)U(y) + f(y)U(z)\Big)\Big(E[\varepsilon_1 I\{\varepsilon_1 \le a\}] + f(a)\sigma^2\Big).
\end{aligned}$$

Remark 3.8 For the ease of presentation we stated results for k-dimensional smooth functions g in Theorem 3.3 and for one dimensional indicator functions in Theorem 3.7. Results can straightforwardly be generalized to k-dimensional vectors g of indicator functions for including information about k quantiles, or vectors with some smooth components and some indicator function components. Further, results can be generalized for all information functions g with expansions similar to those given in Lemma B.1 (ii) or (iii).

4 The heteroscedastic case

In this section we consider a nonparametric heteroscedastic regression model with independent observations

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n, \tag{4.1}$$

under assumptions (M1) from section 3 and (M) stated below. Assumption (M) is somewhat stronger than (M2), (M3) for the homoscedastic model, but was already needed to obtain Akritas and Van Keilegom's (2001) result in the heteroscedastic setting.

(M) The variance function σ^2 is bounded and twice continuously differentiable with bounded derivatives in $(0, 1)$ such that $\inf_{x\in[0,1]} \sigma^2(x) > 0$. The errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with distribution function F. They are centered, $E[\varepsilon_1] = 0$, with variance $\mathrm{Var}(\varepsilon_1) = 1$ and existing fourth moment, and independent from the design points. F is twice continuously differentiable with everywhere positive density f such that $\sup_{y\in\mathbb{R}} |yf(y)| < \infty$, $\sup_{y\in\mathbb{R}} |y^2 f'(y)| < \infty$.

We are going to investigate whether the empirical distribution function $\hat{F}_n$ [defined in (1.2)] of estimated residuals

$$\hat{\varepsilon}_i = \frac{Y_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)} \tag{4.2}$$

[with variance estimator $\hat{\sigma}^2$ defined below in (4.3)] can be improved by including additional information about the error distribution F. For the sake of brevity we restrict ourselves to the case of smooth information functions g satisfying assumptions (A), (S1), (S2) and the modified assumptions (G1'), (G2') stated below.

(G1') We assume that g_j is continuously differentiable and there exist constants γ, C and $\beta > 0$ such that

$$\Big| \int \big(g_j(y + z_1 + yz_2) - g_j(y) - g_j'(y)(z_1 + yz_2)\big) f(y)\, dy \Big| \le C \int |z_1 + yz_2|^{1+\beta} f(y)\, dy$$

for all $z_1, z_2 \in \mathbb{R}$ with $|z_1|, |z_2| \leq \gamma$, $j = 1, \ldots, k$. Moreover, let $E[|g_j'(\varepsilon_1)|] < \infty$, $E[|\varepsilon_1 g_j'(\varepsilon_1)|] < \infty$ for $j = 1, \ldots, k$, and $E[|\varepsilon_1|^{1+\beta}] < \infty$.

(G2') There exist constants δ, C such that for some positive $\kappa < 2(1+\alpha)$ $(j = 1, \ldots, k)$

$$\Big(E\Big[\sup_{\substack{|z_1|\leq\delta, |\tilde{z}_1|\leq\delta, |z_2|\leq\delta, |\tilde{z}_2|\leq\delta \\ |z_1-\tilde{z}_1|\leq\xi, |z_2-\tilde{z}_2|\leq\xi}} (g_j(\varepsilon_1 + z_1 + \varepsilon_1 z_2) - g_j(\varepsilon_1 + \tilde{z}_1 + \varepsilon_1 \tilde{z}_2))^2\Big]\Big)^{1/2} \leq C\xi^{1/\kappa}$$

(where α is defined in assumption (H)).

Our main interest lies in the information given by the model that the errors are centered and have variance one, i. e. $g(\varepsilon) = (\varepsilon, \varepsilon^2 - 1)^T$. The residuals $\hat{\varepsilon}_i$ defined in (4.2) are built with use of the Nadaraya–Watson estimator for m defined in (2.2) and the corresponding variance estimator,

$$\hat{\sigma}^2(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{X_i - x}{h}\Big)(Y_i - \hat{m}(x))^2 \frac{1}{\hat{f}_X(x)}, \tag{4.3}$$

where the kernel density estimator $\hat{f}_X$ is defined in (2.3) and we assume that (K) and (H) [with β from assumption (G1')] are valid. Under the stated assumptions, Akritas and Van Keilegom's (2001) results show that the process $\sqrt{n}(\hat{F}_n(\cdot) - F(\cdot) - b(\cdot))$ converges weakly to a Gaussian process G with covariance structure

$$\begin{aligned} \text{(4.4)}\quad \mathrm{Cov}(G(y), G(z)) = {} & F(y \wedge z) - F(y)F(z) \\ & + f(y)f(z) + E[\varepsilon_1 I\{\varepsilon_1 \leq y\}]f(z) + E[\varepsilon_1 I\{\varepsilon_1 \leq z\}]f(y) \\ & + \frac{1}{2} y f(y) E[(\varepsilon_1^2 - 1) I\{\varepsilon_1 \leq z\}] + \frac{1}{2} z f(z) E[(\varepsilon_1^2 - 1) I\{\varepsilon_1 \leq y\}] \\ & + \frac{1}{2} y f(y) f(z) E[\varepsilon_1^3] + \frac{1}{2} f(y) z f(z) E[\varepsilon_1^3] + \frac{1}{4} y f(y) z f(z) \mathrm{Var}(\varepsilon_1^2), \end{aligned}$$

where the bias term is defined as $b(y) = h^2(f(y)B_1 + yf(y)B_2)$ and

$$B_1 = \frac{1}{2}\int K(u)u^2\,du \int \frac{1}{\sigma(x)}((mf_X)''(x) - (mf_X'')(x))\,dx \tag{4.5}$$

$$B_2 = \frac{1}{2}\int K(u)u^2\,du \int \frac{1}{2\sigma^2(x)}\big((\sigma^2 f_X)''(x) - (\sigma^2 f_X'')(x) + 2(m'(x))^2 f_X(x)\big)\,dx \tag{4.6}$$

[see appendix A]. With use of the now differently defined residuals $\hat{\varepsilon}_i$ [see (4.2)] the empirical likelihood estimator $\overline{F}_n$ is defined as in (3.1) and under the stated assumptions Propositions 3.1 and 3.2 are valid as well [see appendix D]. The main difference to the results in the homoscedastic model arises from a different expansion for linear functionals of the residuals as was given in Remark 3.6 for the homoscedastic case. We formulate the following Proposition that generalizes results by Müller, Schick and Wefelmeyer (2004a) for the heteroscedastic case.

Proposition 4.1 *Assume model (4.1) is valid under assumptions (M), (M1), (G1'), (G2'), (K), (H) and (A). Then, for $j = 1, \ldots, k$,*

$$\frac{1}{n}\sum_{i=1}^{n} g_j(\hat{\varepsilon}_i) \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(g_j(\varepsilon_i) - E[g_j'(\varepsilon_1)]\varepsilon_i - \frac{1}{2}E[\varepsilon_1 g_j'(\varepsilon_1)](\varepsilon_i^2 - 1)\Big)$$
$$- h^2 E[g_j'(\varepsilon_1)]B_1 - h^2 E[\varepsilon_1 g_j'(\varepsilon_1)]B_2 + o_P(n^{-1/2}),$$

where B_1 and B_2 are defined in (4.5) and (4.6), respectively.

The proof of this Proposition is given in appendix D. Combining Propositions 3.1, 3.2 and 4.1 we obtain the main result of this section.

Theorem 4.2 *Under model (4.1) and assumptions (M), (M1), (K), (H), (A), (G1'), (G2'), (S1) and (S2) we have uniformly in $y \in \mathbb{R}$ the expansion*

$$\overline{F}_n(y) \;=\; \overline{b}(y) + \frac{1}{n}\sum_{i=1}^{n}\Big[I\{\varepsilon_i \le y\} + f(y)\varepsilon_i + \frac{1}{2}yf(y)(\varepsilon_i^2 - 1)$$
$$- U(y)^T\Sigma^{-1}\Big(g(\varepsilon_i) - E[g'(\varepsilon_1)]\varepsilon_i - \frac{1}{2}E[\varepsilon_1 g'(\varepsilon_1)](\varepsilon_i^2 - 1)\Big)\Big] + o_P(\frac{1}{\sqrt{n}})$$

where the bias term is

$$\overline{b}(y) = h^2\Big\{\Big(f(y) + U(y)^T\Sigma^{-1}E[g'(\varepsilon_1)]\Big)B_1 + \Big(yf(y) + U(y)^T\Sigma^{-1}E[\varepsilon_1 g'(\varepsilon_1)]\Big)B_2\Big\}$$

with B_1 and B_2 defined in (4.5) and (4.6), respectively. The process $\sqrt{n}(\overline{F}_n(\cdot) - F(\cdot) - \overline{b}(\cdot))$ converges weakly to a centered Gaussian process $\overline{G}$ with covariance structure

$$\begin{aligned}
&\mathrm{Cov}(\overline{G}(y), \overline{G}(z)) \\
=\;& \mathrm{Cov}(G(y), G(z)) + U(y)^T\Sigma^{-1}\Big(\Sigma - E[g'(\varepsilon_1)]E[\varepsilon_1 g^T(\varepsilon_1)] \\
&\quad - E[\varepsilon_1 g(\varepsilon_1)]E[g'(\varepsilon_1)]^T + E[g'(\varepsilon_1)]\sigma^2 E[g'(\varepsilon_1)]^T\Big)\Sigma^{-1}U(z) \\
&- \frac{1}{2}\Big(U(y)^T\Sigma^{-1}zf(z) + U(z)^T\Sigma^{-1}yf(y)\Big)\Big(E[g(\varepsilon_1)(\varepsilon_1^2 - 1)] - E[g'(\varepsilon_1)]E[\varepsilon_1(\varepsilon_1^2 - 1)]\Big) \\
&- U(y)^T\Sigma^{-1}\Big(U(z) - E[g'(\varepsilon_1)]E[\varepsilon_1 I\{\varepsilon_1 \le z\}] + f(z)E[\varepsilon_1 g(\varepsilon_1)] - f(z)\sigma^2 E[g'(\varepsilon_1)]\Big) \\
&- U(z)^T\Sigma^{-1}\Big(U(y) - E[g'(\varepsilon_1)]E[\varepsilon_1 I\{\varepsilon_1 \le y\}] + f(y)E[\varepsilon_1 g(\varepsilon_1)] - f(y)\sigma^2 E[g'(\varepsilon_1)]\Big) \\
&+ \frac{1}{4}U^T(y)\Sigma^{-1}E[\varepsilon_1 g'(\varepsilon_1)]E[(\varepsilon_1^2 - 1)^2]E[\varepsilon_1 g'(\varepsilon_1)]^T\Sigma^{-1}U(z) \\
&- \frac{1}{2}U^T(y)\Sigma^{-1}\Big(E[g(\varepsilon_1)(\varepsilon_1^2 - 1)] - E[g'(\varepsilon_1)]E[\varepsilon_1(\varepsilon_1^2 - 1)]\Big)E^T[\varepsilon_1 g'(\varepsilon_1)]\Sigma^{-1}U(z) \\
&- \frac{1}{2}U^T(z)\Sigma^{-1}\Big(E[g(\varepsilon_1)(\varepsilon_1^2 - 1)] - E[g'(\varepsilon_1)]E[\varepsilon_1(\varepsilon_1^2 - 1)]\Big)E^T[\varepsilon_1 g'(\varepsilon_1)]\Sigma^{-1}U(y)
\end{aligned}$$

$$+\frac{1}{4}\Big(U^T(y)\Sigma^{-1}zf(z)+U^T(z)\Sigma^{-1}yf(y)\Big)E[\varepsilon_1 g'(\varepsilon_1)]E[(\varepsilon_1^2-1)^2]$$
$$+\frac{1}{2}U^T(y)\Sigma^{-1}E[\varepsilon_1 g'(\varepsilon_1)]\Big(E[(\varepsilon_1^2-1)I\{\varepsilon_1\leq z\}]+f(z)E[\varepsilon_1^3]\Big)$$
$$+\frac{1}{2}U^T(z)\Sigma^{-1}E[\varepsilon_1 g'(\varepsilon_1)]\Big(E[(\varepsilon_1^2-1)I\{\varepsilon_1\leq y\}]+f(y)E[\varepsilon_1^3]\Big).$$

with $\mathrm{Cov}(G(y),G(z))$ *from (4.4).*

The example $g(\varepsilon)=(\varepsilon,\varepsilon^2-1)^T$ to include the model assumptions explicitly in the estimation is considered in the next section in Example 5.6. A sketch of the proof of Theorem 4.2 is given in appendix D.

5 Discussion of the asymptotic results

In this section we compare the asymptotic mean squared error of the residual based empirical distribution function $\hat{F}_n$ with the mean squared error of the empirical likelihood estimator $\overline{F}_n$. We concentrate on the homoscedastic model first. Here we have

$$\begin{aligned}\mathrm{mse}(y) &= \frac{1}{n}\mathrm{Var}(G(y))+b^2(y)\\ &= \frac{1}{n}\Big(F(y)(1-F(y))+\sigma^2f^2(y)+2f(y)E[\varepsilon_1 I\{\varepsilon_1\leq y\}]\Big)\\ &\quad +h^4f^2(y)B^2\end{aligned}$$

[see (1.3), (1.4) in the introduction and B defined in Theorem 3.3]. First we consider the case of a smooth function g in the situation of Theorem 3.3. For the mean squared error of the new estimator we have

$$\begin{aligned}\overline{\mathrm{mse}}(y) &= \frac{1}{n}\mathrm{Var}(\overline{G}(y))+\overline{b}^2(y)\\ &= \frac{1}{n}\Big(F(y)(1-F(y))+\sigma^2f^2(y)+2f(y)E[\varepsilon_1 I\{\varepsilon_1\leq y\}]+W(y)\Big)\\ &\quad +h^4\Big(f(y)+U(y)^T\Sigma^{-1}E[g'(\varepsilon_1)]\Big)^2B^2\end{aligned}$$

where

$$\begin{aligned}W(y) &= U^T(y)\Sigma^{-1}\{\Sigma-E[g'(\varepsilon_1)]E[\varepsilon_1 g^T(\varepsilon_1)]-E[\varepsilon_1 g(\varepsilon_1)]E^T[g'(\varepsilon_1)]\\ &\quad +\sigma^2E[g'(\varepsilon_1)]E^T[g'(\varepsilon_1)]\}\Sigma^{-1}U(y)-2U^T(y)\Sigma^{-1}\{U(y)-E[g'(\varepsilon_1)]E[\varepsilon_1 I\{\varepsilon_1\leq y\}]\}\\ &\quad -2f(y)U^T(y)\Sigma^{-1}\{E[\varepsilon_1 g(\varepsilon_1)]-E[g'(\varepsilon_1)]\sigma^2\}.\end{aligned}$$

For the comparison we assume for the bandwidth used for both estimators that $h=cn^{-1/4}$ for some constant $c>0$ (such that $h^4=\frac{c^4}{n}$) and consider different examples of functions g

and distributions F. For the figures shown below we multiply the curves with n, hence, they are independent of the sample size. The value of B depends on the regression function, the design density and the choice of the kernel. We set $c^4B^2 = 1$ for the curves shown below. It is clear that the influence the bias has on the mean squared error changes with this constant and, hence, changes with the underlying regression function and the choice of kernel and bandwidth.

Example 5.1 The most important example is to include the moment assumption of centered residuals into the estimation. In the case $g(\varepsilon) = \varepsilon$ the asymptotic variance of both estimators is the same (compare Remark 3.4 and Corollary 3.5) and therefore it is sufficient to compare the bias terms. We have $\Sigma = E[\varepsilon_1^2] = \sigma^2$ and $g' \equiv 1$ and therefore $\overline{b}(y) = h^2B(f(y) + \frac{1}{\sigma^2}U(y))$ where $U(y) = E[\varepsilon_1 I\{\varepsilon_1 \leq y\}] \leq 0$ for all $y \in \mathbb{R}$. For example in the case of normally distributed errors we obtain $U(y) = -\sigma^2 f(y)$ and therefore $\overline{b}(y) = 0$ whereas $b(y) = h^2Bf(y) \neq 0$ for all $y \in \mathbb{R}$ (when $B \neq 0$). In this case the first order bias term cancels completely by the use of the empirical likelihood method. Also for examples of other distributions the new bias term can be considerably smaller. Figure 1 below shows curves for the squared bias, the variance and the mean squared error for normal distribution, student's t-distribution with three degrees of freedom and the double exponential distribution with density

$$f(x) = \frac{1}{\gamma} \exp\Big(- \frac{x-q}{\gamma}\Big) \times \exp\Big(- \exp\Big(- \frac{x-q}{\gamma}\Big)\Big)$$

for $\gamma = 3$ and $q = -\gamma c$ such that the expectation $\gamma c + q$ vanishes, where $c = 0.5772$ approximately, as skew distribution example (with $E[\varepsilon_1^3] \neq 0$). We observe a bias reduction for all three example distributions.

INCLUDE FIGURE 1 HERE.

Example 5.2 In the case $g(\varepsilon) = \varepsilon^2 - \sigma^2$ for known variance σ^2 we have $E[g'(\varepsilon_1)] = 0$ and therefore both estimators have the same bias $b(y) = \overline{b}(y)$. The new variance is

$$\begin{aligned}\mathrm{Var}(\overline{G}(y)) \;=\; & F(y)(1 - F(y)) + f^2(y)\sigma^2 + 2f(y)E[\varepsilon_1 I\{\varepsilon_1 \leq y\}] \\ & -U^2(y)\Sigma^{-1} - 2f(y)U(y)\Sigma^{-1}E[\varepsilon_1^3].\end{aligned}$$

For all distributions with vanishing third moment the new variance is uniformly smaller than $\mathrm{Var}(G(y))$ and we obtain an improved estimator. In particular, for normally distributed errors we have

$$\mathrm{Var}(\overline{G}(y)) \;=\; \mathrm{Var}(G(y)) - \frac{1}{2}y^2 f^2(y).$$

Figure 2 below shows squared bias, variance and mean squared error curves for normal distribution, student's t-distribution with five degrees of freedom and the double exponential distribution defined in Example 5.1.

INCLUDE FIGURE 2 HERE.

Example 5.3 The results obtained in Example 5.2 for a known variance can further be improved by including the information that the errors are centered and defining $g(\varepsilon) = (\varepsilon, \varepsilon^2 - \sigma^2)^T$. Denoting $\mu_3 = E[\varepsilon_1^3]$, $\mu_4 = E[(\varepsilon_1^2 - \sigma^2)^2]$ and $U_1(y) = E[\varepsilon_1 I\{\varepsilon_1 \leq y\}]$, $U_2(y) = E[(\varepsilon_1^2 - \sigma^2) I\{\varepsilon_1 \leq y\}]$ we obtain

$$\begin{aligned}
\operatorname{Var}(\overline{G}(y)) &= F(y)(1 - F(y)) + f^2(y)\sigma^2 + 2f(y)E[\varepsilon_1 I\{\varepsilon_1 \leq y\}] \\
&\quad + \frac{1}{(\sigma^2\mu_4 - \mu_3^2)^2}\Big(U_1^2(y)\mu_4\mu_3^2 + U_2^2(y)(2\sigma^2\mu_3 - \sigma^4\mu_4) \\
&\quad - 2U_1(y)U_2(y)\mu_3^3\Big) + \frac{2}{\sigma^2\mu_4 - \mu_3^2} f(y)\Big(U_1(y)\mu_3^2 - U_2(y)\mu_3\sigma^2\Big) \\
&= \operatorname{Var}(G(y)) - U_2^2(y)\frac{1}{\mu_4}
\end{aligned}$$

where the last line only holds for distributions with vanishing third moment. In this case the variance reduces to the variance in Example 5.2. For the bias term we have

$$\overline{b}(y) = h^2 B\Big(f(y) + \frac{1}{\sigma^2\mu_4 - \mu_3^2}(U_1(y)\mu_4 - U_2(y)\mu_3)\Big).$$

For distributions with $\mu_3 = 0$ the bias is the same as in Example 5.1; in particular, it is zero for normal distributions. For all distributions with vanishing third moment we therefore obtain an estimator with both smaller bias and smaller asymptotic variance. Figure 3 shows the corresponding squared bias, variance and mse curves for normal distribution, student's t-distribution with five degrees of freedom and double exponential distribution. For all three distributions we observe a considerably smaller mean squared error compared to Example 5.2.

INCLUDE FIGURE 3 HERE.

Next, we consider an example corresponding to additional information about quantiles according to Theorem 3.7.

Example 5.4 For $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ we obtain for distributions with zero median ($a = 0, b = 0.5$) the variance

$$\operatorname{Var}(\overline{G}(y)) = F(y)(1 - F(y)) + f^2(y)\sigma^2 + 2f(y)E[\varepsilon_1 I\{\varepsilon_1 \leq y\}]$$

$$+4\Big[4U^2(y)\{\frac{1}{4}+2f(0)E[\varepsilon_1 I\{\varepsilon_1\le 0\}]+f^2(0)\sigma^2\}$$
$$-2U(y)(F(0\wedge y)-\frac{1}{2}F(y)+f(0)E[\varepsilon_1 I\{\varepsilon_1\le y\}])$$
$$-2f(y)U(y)\{E[\varepsilon_1 I\{\varepsilon_1\le 0\}]+f(0)\sigma^2\}\Big]$$

with $U(y)=F(0\wedge y)-\frac{1}{2}F(y)\in[0,0.25]$ for all $y\in\mathbb{R}$. The bias is

$$\overline{b}(y) \;=\; h^2B(f(y)-4U(y)f(0))=h^2B(f(y)-4[F(0\wedge y)-\frac{1}{2}F(y)]f(0)).$$

Figure 4 below shows curves for the squared bias, the variance and the mean squared error for normal distribution, student's t-distribution with three degrees of freedom (both with $a=0$ and $b=0.5$) and double exponential distribution (with $a=0$ and $b=0.570371$).

INCLUDE FIGURE 4 HERE.

Example 5.5 We investigate in this example whether including the centeredness information gives better results compared to Example 5.4, that is, we consider $g(\varepsilon)=(\varepsilon, I\{\varepsilon\le 0\}-F(0))^T$, where $F(0)$ is known. This situation is not covered by Theorems 3.3 and 3.7, but results can be derived in a complete analogous way. We obtain for the asymptotic variance,

$$\mathrm{Var}(\overline{G}(y)) \;=\; \mathrm{Var}(G(y))+V^2(y)\mathrm{Var}(G(0))-2V(y)\mathrm{Cov}(G(y),G(0))$$

where $\mathrm{Var}(G(y))$ is defined in (1.3),

$$V(y) \;=\; \frac{\sigma^2U_2(y)-U_1(0)U_1(y)}{\sigma^2F(0)(1-F(0))-U_1^2(0)}$$

and $U_1(y)=E[\varepsilon_1 I\{\varepsilon_1\le y\}]$, $U_2(y)=E[(I\{\varepsilon_1\le 0\}-F(0))I\{\varepsilon_1\le y\}]=F(y\wedge 0)-F(0)F(y)$. For the bias we have

$$\overline{b}(y) \;=\; h^2B\Big(f(y)+\frac{U_1(y)\{F(0)(1-F(0))+U_1(0)f(0)\}-U_2(y)\{U_1(0)+\sigma^2f(0)\}}{\sigma^2F(0)(1-F(0))-U_1^2(0)}\Big)$$

For standard normally distributed errors we have again $\overline{b}(y)=0$. Figure 5 below shows curves for the squared bias, the variance and the mean squared error for normal distribution, student's t-distribution with three degrees of freedom and the double exponential distribution. We obtain uniformly smaller mse's for $\overline{F}_n$ compared with Example 5.4.

INCLUDE FIGURE 5 HERE.

Finally, we consider one example for the heteroscedastic model.

Example 5.6 We consider the information function $g(\varepsilon) = (\varepsilon, \varepsilon^2 - 1)^T$ to include the model assumption of centered errors with unit variance into the error distribution estimation. This is comparable to Example 5.1 in the homoscedastic setting and analogously the asymptotic variance curves of $\hat{F}_n$ and $\overline{F}_n$ coincide whereas the new estimator has a smaller bias. In Figure 6 the squared bias, variance and mse curves are displayed for normal, student's t-distribution with five degrees of freedom and double exponential distribution (all distributions standardized such that they are centered with unit variance).

INCLUDE FIGURE 6 HERE.

6 Small sample performance

In this section we compare the performances of the two distribution estimators for finite samples by means of a simulation study. We concentrate on the homoscedastic model (2.1) with regression function $m(x) = 3x^2$ and uniformly in $[0, 1]$ distributed design points. As error distributions we use the three different distributions already considered in section 5, namely standard normal distribution, student's t-distribution and the double exponential distribution. For the regression estimator we use the standard normal kernel and we consider bandwidths $h = c/n^{1/4}$ according to the theoretical bandwidth conditions, for different suitable values of the constant c between 0.1 and 3. Displayed in Figures 7 and 8 are values of the mean integrated squared error $E[\int(G_n(y) - F(y))^2 dy]$ for $G_n = \hat{F}_n$ and $G_n = \overline{F}_n$, estimated from 1000 replications, where the integral is approximated using 100 grid points in an appropriate interval chosen as $[-3, 3]$ for the normal distribution, $[-4, 4]$ for the t-distribution and $[-7, 7]$ for the double exponential distribution. For calculating $\hat{\eta}_n$ defined in (3.2) we used the Bisection method for one-dimensional g and the multivariate Newton Raphson procedure otherwise [both described in Press, Teukolsky, Vetterling and Flannery (2002), p. 357 and p. 383].

For Example 5.1, i. e. including the centeredness information, results are displayed in Figure 7. The sample size is $n = 50$ in the left column and $n = 100$ in the right column. In the first row the error distribution is standard normal. The solid curve (curve 1) displays the mean integrated squared error, or MISE, for the residual based empirical distribution function $\hat{F}_n$, whereas the dashed curve (curve 2) displays the corresponding results for the new estimator $\overline{F}_n$. We always obtain better results, i. e. a smaller MISE, for the new estimator, although for some choices of bandwidths the values are very close. For the symmetric distributions it is interesting to see the effect that for an increasing bandwidth the difference between the performances of the two estimators increases. This is according to the theory because for an increasing bandwidth the effect of the bias on the MISE increases and in this example the new estimator has a considerably smaller asymptotic bias. The results displayed in the second and third row of Figure 7 correspond to the student's t-distribution with three degrees of freedom and the double exponential distribution, respectively. We obtain a smaller MISE for the new estimator in all cases.

INCLUDE FIGURE 7 HERE.

In Figure 8 the MISE curves for Example 5.3 are displayed in the left panel, where the sample size is $n = 100$ and we consider standard normal, t_5 and double exponential distribution. For standard normally and t–distributed errors the empirical likelihood methods yields a great improvement in terms of MISE. For double exponential distribution the new estimator has a smaller MISE only for bandwidth constants greater than $c = 0.5$ and for this distribution larger bandwidths should be recommended. In the right panel of Figure 8 results for Example 5.5 for sample size $n = 50$ and normal, t_3 and double exponential distribution are shown. In all cases the new estimator has a smaller MISE.

INCLUDE FIGURE 8 HERE.

We also implemented an approximation of $\hat{\eta}_n$ from the asymptotic expansion given in Proposition 3.2, where only the dominating term is used. In most cases this gave even better results, but it failed to work in Example 5.3. Therefore, the results are not displayed and we do recommend to use the Bisection or Newton Raphson procedure to obtain $\hat{\eta}_n$ instead of using the approximation.

Acknowledgements. The authors would like to thank Ingrid Van Keilegom and Holger Dette for helpful discussions. The financial support by the Deutsche Forschungsgemeinschaft (SFB 475) is gratefully acknowledged. Part of this paper was written while the third author visited the Australian National University in Canberra and this author would like to thank the members of the Mathematical Science Institute, in particular Peter Hall, for their hospitality.

A Appendix: Akritas and Van Keilegom's process

The homoscedastic model

We give a short derivation of the results stated in the introduction, in particular (1.4) and (1.3). From the proof of Theorem 1 by Akritas and Van Keilegom (2001, p. 555) it follows that in the case of homoscedasticity (where no estimation of a variance function σ is needed) and for score function $J = I[0,1]$ the function φ defined in Theorem 1 (p. 552) has the form

$$\varphi(x,z,y) \;=\; -f(y)\int (I\{z \le v\} - F(v|x))\,dv.$$

Here, $F(v|x) = F(v - m(x))$ is the conditional distribution of Y_i given $X_i = x$ and we obtain

$$\begin{aligned}\varphi(x,z,y) \;&=\; -f(y)\Big(\int_z^\infty (1 - F(v|x))\,dv - \int_{-\infty}^z F(v|x)\,dv\Big)\\ &=\; -f(y)(m(x) - z).\end{aligned}$$

Therefore, $\varphi(X_i, Y_i, y) = f(y)\varepsilon_i$ and (1.3) follows from Theorem 2 by Akritas and Van Keilegom (2001, p. 552). The bias formula (1.4) is deduced from the bias of the Nadaraya–Watson [Nadaraya (1964), Watson (1964)] regression estimator and the expansion

$$\hat F_n(y) - F(y) \;=\; F_n(y) - F(y) + f(y)\int (\hat m(x) - m(x)) f_X(x)\,dx + o_P(n^{-1/2})$$

where F_n denotes the empirical distribution function of the true errors [see p. 555 of Akritas and Van Keilegom (2001)].

The heteroscedastic model

In the heteroscedastic model (4.1) the function φ from Theorem 1 by Akritas and Van Keilegom (2001) is equal to

$$\varphi(x,z,y) \;=\; -\frac{f(y)}{\sigma^2(x)}\Big(\sigma(x)(m(x) - z) - ym^2(x) + ym(x)z + \frac{1}{2}\sigma^2(x)y + \frac{1}{2}m^2(x)y - \frac{1}{2}yz^2\Big)$$

and this yields $\varphi(X_i, Y_i, y) = f(y)(\varepsilon_i + \frac{y}{2}(\varepsilon_i^2 - 1))$. The bias $b(y)$ is deduced from the expansion

$$\begin{aligned}\hat F_n(y) - F(y) \;&=\; F_n(y) - F(y) + f(y)\int \frac{\hat m(x) - m(x)}{\sigma(x)} f_X(x)\,dx\\ &\quad + yf(y)\int \frac{\hat\sigma(x) - \sigma(x)}{\sigma(x)} f_X(x)\,dx + o_P(n^{-1/2})\\ &=\; F_n(y) - F(y) + f(y)\int \frac{\hat m(x) - m(x)}{\sigma(x)} f_X(x)\,dx\\ &\quad + yf(y)\int \frac{\hat\sigma^2(x) - \sigma^2(x)}{2\sigma^2(x)} f_X(x)\,dx + o_P(n^{-1/2})\end{aligned}$$

[compare p. 555, 564, Akritas and Van Keilegom (2001)] by inserting the definitions of $\hat{m}$ in (2.2) and $\hat{\sigma}^2$ in (4.3).

B Appendix: Auxiliary results

Lemma B.1 *Assume model (2.1) with assumptions (M1), (M2), (K), (H) and (A).*

(i) *With either $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ for some constants a and b ($k = 1$) or g satisfying (G1) and (G2) we have $\max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i)| = o_P(\sqrt{n})$ ($j = 1, \dots, k$).*

(ii) *Under the additional assumptions (G1), (G2) we have the expansion $\frac{1}{n}\sum_{i=1}^n g_j(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^n (g_j(\varepsilon_i) - E[g_j'(\varepsilon_1)]\varepsilon_i) - h^2 E[g_j'(\varepsilon_1)]B + o_P(\frac{1}{\sqrt{n}})$ ($j = 1, \dots, k$) where B is defined in Theorem 3.3.*

(iii) *For $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ for some constants a and b ($k = 1$) we have the expansion $\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^n g(\varepsilon_i) + f(a)\varepsilon_i + h^2 f(a)B + o_P(\frac{1}{\sqrt{n}})$ where B is defined in Theorem 3.3.*

(iv) *With either $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ for some constants a and b ($k = 1$) or g satisfying (G1) and (G2) we have $\frac{1}{n}\sum_{i=1}^n (g_j(\hat{\varepsilon}_i) - g_j(\varepsilon_i))^2 = o_P(1)$ ($j = 1, \dots, k$).*

(v) *With either $g(\varepsilon) = I\{\varepsilon \leq a\} - b$ for some constants a and b ($k = 1$) or g satisfying (G1) and (G2) we have $\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i) g^T(\hat{\varepsilon}_i) = \Sigma + o_P(1)$.*

Proof. (i). Under the assumptions, condition (2.4) is valid. We have from the triangle inequality

$$\max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i)| \;\leq\; \max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i) - g_j(\varepsilon_i)| + \max_{i=1,\dots,n} |g_j(\varepsilon_i)| \tag{B.1}$$

and consider the second term on the right hand side first. For every $\epsilon > 0$ one obtains

$$\begin{aligned} P\Big(\max_{i=1,\dots,n} |g_j(\varepsilon_i)| > \epsilon\sqrt{n}\Big) \;&\leq\; nE[I\{|g_j(\varepsilon_1)| > \epsilon\sqrt{n}\}] \\ &\leq\; \frac{1}{\epsilon^2} E\Big[|g_j(\varepsilon_1)|^2 I\{|g_j(\varepsilon_1)|^2 \geq \epsilon^2 n\}\Big] \;=\; o(1) \end{aligned} \tag{B.2}$$

with the assumption $E[g_j^2(\varepsilon_1)] < \infty$. For the first term on the right hand side of (B.1) we further have

$$\begin{aligned} P\Big(\max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i) - g_j(\varepsilon_i)| > \epsilon\sqrt{n}\Big) \;\leq\; & P\Big(\max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i) - g_j(\varepsilon_i)| > \epsilon\sqrt{n}, \max_{i=1,\dots,n} |\varepsilon_i - \hat{\varepsilon}_i| \leq \delta\Big) \\ & + P\Big(\max_{i=1,\dots,n} |\varepsilon_i - \hat{\varepsilon}_i| > \delta\Big) \end{aligned}$$

[with δ from condition (2.4)]. The last probability converges to zero because

$$\text{(B.3)} \qquad \max_{i=1,\dots,n} |\varepsilon_i - \hat{\varepsilon}_i| \le \sup_{x\in[0,1]} |\hat{m}(x) - m(x)| = o(1)$$

almost surely by assumptions (M1), (M2), (K) and (H). Furthermore we have

$$P\Big(\max_{i=1,\dots,n} |g_j(\hat{\varepsilon}_i) - g_j(\varepsilon_i)| > \epsilon\sqrt{n}, \max_{i=1,\dots,n} |\varepsilon_i - \hat{\varepsilon}_i| \le \delta\Big) \;\le\; P\Big(\max_{i=1,\dots,n} |h_j(\varepsilon_i)| > \epsilon\sqrt{n}\Big)$$

where $h_j(x) = \sup_{y\in\mathbb{R}:|y|\le\delta} |g_j(x+y) - g_j(x)|$. Analogous to argumentation (B.2) we obtain the assertion from (2.4), that is $E[h_j^2(\varepsilon_1)] < \infty$.

(ii). This statement corresponds to Theorem 2 by Müller, Schick and Wefelmeyer (2004a) when using a leave-one-out local polynomial estimator for the regression function. We present a different proof nevertheless because some arguments of the proof are used to show (iv) and (v). Our proof uses some ideas of the proof of Lemma 1, Akritas and Van Keilegom (2001). First, we show weak convergence of the empirical process

$$\text{(B.4)} \qquad G_n(\tilde{g}_j) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\tilde{g}_j(\varepsilon_i, X_i) - \int\int \tilde{g}_j(y,x) f(y) f_X(x)\, dy\, dx\Big)$$

indexed by functions $\tilde{g}_j \in \mathcal{G}_j$, where

$$\text{(B.5)} \qquad \mathcal{G}_j \;=\; \Big\{\tilde{g}_j : \mathbb{R}\times[0,1] \to \mathbb{R}, \tilde{g}_j(\varepsilon, x) = g_j(\varepsilon + h(x)) - g_j(\varepsilon) \;\Big|\; h \in \mathcal{H}\Big\}.$$

The smooth function class $\mathcal{H} = C_\delta^{1+\alpha}[0,1]$ is defined in van der Vaart and Wellner (1996, p. 154) [see also the proof of Lemma D.1: all continuous functions $h : [0,1] \to I\!R$ that fulfill (D.1) build the class $C_\rho^{1+\alpha}[0,1]$]. Note that for $n \to \infty$ the function $m - \hat{m}$ is an element of $C_\delta^{1+\alpha}[0,1]$ with probability one [confer Akritas and Van Keilegom (2001)] and that

$$\frac{1}{n}\sum_{i=1}^{n} g_j(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^{n} g_j(\varepsilon_i + (m - \hat{m})(X_i)).$$

The function class $\mathcal{G}_j$ has a square integrable envelope by (2.4) because $\sup_{x\in[0,1]} |h(x)| \le \delta$ for $h \in \mathcal{H}$. In order to show that $\mathcal{G}_j$ is Donsker we prove that the bracketing integral is finite, i. e.

$$\text{(B.6)} \qquad \int_0^\infty \sqrt{\log N_{[\,]}(\xi, \mathcal{G}_j, L_2(P))}\, d\xi < \infty,$$

where P denotes the distribution of (ε_1, X_1). Let $\xi > 0$ and define $\tilde{\xi} = (\xi/(2C))^\kappa$ with constant C defined in assumption (G2). We have from Theorem 2.7.1 in van der Vaart and Wellner (1996, p. 155) for the covering numbers of $\mathcal{H}$ with respect to the supremum norm,

$$\text{(B.7)} \qquad \log N(\tilde{\xi}, \mathcal{H}, ||\cdot||_\infty) \le K \xi^{-\kappa/(1+\alpha)}$$

for some constant K. Further, for $\tilde{\xi} \geq \delta$ the covering number is one, choosing the center $h_1 \equiv 0$ and noting that $\sup_{x\in[0,1]} |h(x)| \leq \delta$ for $h \in \mathcal{H}$. Let $h_1, \ldots, h_\lambda$ $[\lambda = N(\tilde{\xi}, \mathcal{H}, ||\cdot||_\infty)]$ denote the centers of a covering for $\mathcal{H}$ with radius $\tilde{\xi}$ with respect to the supremum norm. Let $h \in \mathcal{H}$ and $||h - h_i||_\infty < \tilde{\xi}$. Then a bracket for $\tilde{g}_j(\varepsilon, h) = g_j(\varepsilon + h(x)) - g_j(\varepsilon)$ is given by

$$\Big[g_j(\varepsilon + h_i(x)) - g_j(\varepsilon) - \tilde{g}_j^*(\varepsilon, x)\ ,\ g_j(\varepsilon + h_i(x)) - g_j(\varepsilon) + \tilde{g}_j^*(\varepsilon, x)\Big]$$

where $\tilde{g}_j^*(\varepsilon, x) = \sup |g_j(\varepsilon + z) - g_j(\varepsilon + \tilde{z})|$ and the supremum is built over $|z| \leq \delta, |\tilde{z}| \leq \delta, |z - \tilde{z}| \leq \tilde{\xi}$. The bracket has $L_2(P)$ length less or equal to ξ by assumption (G2). We have $N_{[\,]}(\xi, \mathcal{G}_j, L_2(P)) = \lambda$ brackets and (B.6) follows from (B.7) and the assumption $\frac{\kappa}{2(1+\alpha)} < 1$ [the integral only has to be evaluated in $(0, 2C\delta^{1/\kappa})$].

We have shown weak convergence of the process $G_n(\tilde{g}_j)$ defined in (B.4) and insert the random function $\hat{g}_j(\varepsilon, x) = g_j(\varepsilon + (m - \hat{m})(x)) - g_j(\varepsilon)$ now. We have

$$\text{(B.8)} \qquad \int \hat{g}_j^2\, dP = \int\int \Big(g_j(y + (m - \hat{m})(x)) - g_j(y)\Big)^2 f_X(x) f(y)\, dx\, dy \;=\; o_P(1)$$

by the dominated convergence theorem using (2.4) and the convergence $\int (g_j\,(y + (m - \hat{m})(x)) - g_j(y))^2 f_X(x)\, dx \longrightarrow 0$ for each fixed y (follows by Taylor's expansion and the almost sure uniform convergence of $\hat{m} - m$ to zero, because g_j' is continuous and therefore bounded in a neighbourhood of the fixed y). By (B.8), applying Lemma 19.24 of van der Vaart (1998, p. 280) we obtain

$$\begin{aligned} G_n(\hat{g}_j) \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(g_j(\varepsilon_i + (m - \hat{m})(X_i)) - g_j(\varepsilon_i) \\ &\quad - \int\int (g_j\,(y + (m - \hat{m})(x)) - g_j(y)) f_X(x) f(y)\, dx\, dy\Big) = o_P(1). \end{aligned}$$

By assumptions (G1) and (H) using $\sup_{x\in[0,1]} |\hat{m}(x) - m(x)| = O((n^{-1}h^{-1}\log h^{-1})^{1/2})$ a.s. [confer Prop. 3, Akritas and Van Keilegom (2001)] we obtain the expansion

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n g_j(\varepsilon_i + (m - \hat{m})(X_i)) \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n g_j(\varepsilon_i) \\ &\quad + \sqrt{n} \int (m - \hat{m})(x) f_X(x)\, dx \int g_j'(y) f(y)\, dy + o_P(1). \end{aligned}$$

The assertion now follows by inserting the definition of $\hat{m}$ in (2.2) and tedious but simple calculations of expectations and variances using assumptions (K) and (H).

(iii). This follows like Theorem 1 of Akritas and Van Keilegom (2001) in a homoscedastic model, compare appendix A.

(iv). The proof uses results from the proofs of (ii) and (iii). The function class $\mathcal{G}_j$ defined in (B.5) is Donsker (in either case for g) and therefore $\mathcal{G}_j^2$ is Glivenko–Cantelli class in probability [van der Vaart and Wellner (1996, p. 194, Lemma 2.10.14)]. We obtain

$$\sup_{h\in\mathcal{H}}\Big|\frac{1}{n}\sum_{i=1}^n\Big(g_j(\varepsilon_i+h(X_i))-g_j(\varepsilon_i)\Big)^2-\int\int\Big(g_j(y+h(x))-g_j(y)\Big)^2 f_X(x)f(y)\,dx\,dy\Big| \;=\; o_p(1)$$

and therefore

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\Big(g_j(\varepsilon_i+(m-\hat m)(X_i))-g_j(\varepsilon_i)\Big)^2\\ =\;&\int\int\Big(g_j(y+(m-\hat m)(x))-g_j(y)\Big)^2 f_X(x)f(y)\,dx\,dy+o_P(1).\end{aligned}$$

The assertion follows from (B.8) in the case of a smooth function g and the analogous statement for the indicator function $g(\varepsilon)=I\{\varepsilon\le a\}-b$. The latter is obtained by

$$\int\int(g\,(y+(m-\hat m)(x))-g(y))^2 f_X(x)f(y)\,dx\,dy \;=\; \int|F(a+(\hat m-m)(x))-F(a)|f_X(x)\,dx,$$

assumption (M3) and $\sup_{x\in[0,1]}|\hat m(x)-m(x)|=o(1)$ almost surely.

(v). For $j,\ell\in\{1,\dots,k\}$ we have to show that

$$\frac{1}{n}\sum_{i=1}^n g_j(\hat\varepsilon_i)g_\ell(\hat\varepsilon_i) \;=\; \int g_j(y)g_\ell(y)f(y)\,dy+o_P(1).$$

The proof is similar to the proof of (iv). The function classes

$$\tilde{\mathcal{G}}_j \;=\; \{(\varepsilon,x)\mapsto g_j(\varepsilon+h(x))\mid h\in\mathcal{H}\}$$

are Donsker ($j=1,\dots,k$) (where $\mathcal{H}=C_\delta^{1+\alpha}[0,1]$). From van der Vaart and Wellner (1996, p. 204, Problem 8) follows that $\tilde{\mathcal{G}}_j\tilde{\mathcal{G}}_\ell$ is Glivenko–Cantelli class in probability. From this follows

$$\sup_{h\in\mathcal{H}}\Big|\frac{1}{n}\sum_{i=1}^n g_j(\varepsilon_i+h(X_i))g_\ell(\varepsilon_i+h(X_i))-E[g_j(\varepsilon_1+h(X_1))g_\ell(\varepsilon_1+h(X_1))]\Big| \;=\; o_P(1)$$

and therefore we have with $\hat\varepsilon_i=\varepsilon_i+(m-\hat m)(X_i)$

$$\Big|\frac{1}{n}\sum_{i=1}^n g_j(\hat\varepsilon_i)g_\ell(\hat\varepsilon_i)-\int\int g_j\,(y+(m-\hat m)(x))\,g_\ell\,(y+(m-\hat m)(x))\,f_X(x)f(y)\,dx\,dy\Big| \;=\; o_P(1).$$

From $\sup_{x\in[0,1]}|\hat m(x)-m(x)|=o(1)$ almost surely we obtain the assertion similar to (B.8) and the considerations at the end of the proof of (iv). $\square$

Lemma B.2 *Under assumptions (M1), (M2), (K), (H), (A), (S1) and (S2) we have*

(i) $||\hat{\eta}_n|| = O_P(1/\sqrt{n})$

(ii) $\max\limits_{i=1,\ldots,n} \Big|\dfrac{1}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\Big| = O_P(1).$

Proof. (i). $\hat{\eta}_n$ is the solution of equation (3.2). From this we obtain the estimation

$$\begin{aligned}
0 &= \Big|\Big|\frac{1}{n}\sum_{i=1}^n \frac{g(\hat{\varepsilon}_i)}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\Big|\Big| \\
&= \Big|\Big|\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i) - \frac{1}{n}\sum_{i=1}^n \hat{\eta}_n^T g(\hat{\varepsilon}_i)\frac{g(\hat{\varepsilon}_i)}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\Big|\Big| \\
&\geq ||\hat{\eta}_n||\frac{1}{n}\sum_{i=1}^n ||g(\hat{\varepsilon}_i)||^2 \frac{1}{1+||\hat{\eta}_n|| \max\limits_{j=1,\ldots,k}\max\limits_{i=1,\ldots,n}|g_j(\hat{\varepsilon}_i)|} - \Big|\Big|\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i)\Big|\Big|
\end{aligned}$$

and it follows that

$$||\hat{\eta}_n|| \leq \Big|\Big|\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i)\Big|\Big|\Big(\frac{1}{n}\sum_{i=1}^n ||g(\hat{\varepsilon}_i)||^2 - \Big|\Big|\frac{1}{n}\sum_{i=1}^n g(\hat{\varepsilon}_i)\Big|\Big| \max_{j=1,\ldots,k}\max_{i=1,\ldots,n}|g_j(\hat{\varepsilon}_i)|\Big)^{-1}.$$

Now from Lemma B.1 (iv) we have that $\frac{1}{n}\sum_{i=1}^n ||g(\hat{\varepsilon}_i)||^2 = \frac{1}{n}\sum_{i=1}^n g_1^2(\hat{\varepsilon}_i)+\ldots+g_k^2(\hat{\varepsilon}_i)$ converges in probability to $E[g_1^2(\varepsilon_1)+\ldots+g_k^2(\varepsilon_1)] > 0$ by assumption (S2). From Lemma B.1 (i) and (ii) resp. (iii) follows the assertion.

(ii). In order to prove the assertion we show $\max_{1\leq i\leq n}|1-(1+\hat{\eta}_n^T g(\hat{\varepsilon}_i))^{-1}| = o_P(1)$. By Lemma B.1(i) and Lemma B.2(i) we have $P(||\hat{\eta}_n|| \max_{1\leq i\leq n, 1\leq j\leq k}|g_j(\hat{\varepsilon}_i)| \geq 1) = o(1)$ and, further, for all $\epsilon > 0$,

$$\begin{aligned}
&P\Big(\max_{1\leq i\leq n}\Big|\frac{\hat{\eta}_n^T g(\hat{\varepsilon}_i)}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\Big| > \epsilon\ ,\ ||\hat{\eta}_n|| \max_{1\leq i\leq n, 1\leq j\leq k}|g_j(\hat{\varepsilon}_i)| < 1\Big) \\
\leq\ &P\Big(\frac{||\hat{\eta}_n|| \max_{1\leq i\leq n,1\leq j\leq k}|g(\hat{\varepsilon}_i)|}{1-||\hat{\eta}_n|| \max_{1\leq i\leq n,1\leq j\leq k}|g(\hat{\varepsilon}_i)|} > \epsilon\Big) \;=\; P\Big(||\hat{\eta}_n|| \max_{1\leq i\leq n,1\leq j\leq k}|g(\hat{\varepsilon}_i)| > \frac{\epsilon}{1+\epsilon}\Big) \;=\; o(1).
\end{aligned}$$

□

Lemma B.3 *Under model (2.1) and assumptions (M1), (M2), (M3), (K), (H) and (A) we have* $\sup\limits_{y\in\mathbb{R}} \dfrac{1}{n}\sum\limits_{i=1}^n |I\{\hat{\varepsilon}_i \leq y\} - I\{\varepsilon_i \leq y\}| = o_P(1).$

Proof. The assertion is shown in two steps. The first step consists in showing $\frac{1}{n}\sum_{i=1}^n |I\{\hat{\varepsilon}_i \leq y\} - I\{\varepsilon_i \leq y\}| = o_P(1)$ for fixed $y \in \mathbb{R}$. To this end note that $|I\{\varepsilon_i \leq y\} - I\{\hat{\varepsilon}_i \leq y\}| \leq$

$I\{y-\tau<\varepsilon_i\le y+\tau\}$ for $\tau>0$ for all $i=1,\dots,n$, $y\in\mathbb{R}$ whenever $\max_{1\le i\le n}|\hat\varepsilon_i-\varepsilon_i|\le\tau$. Hence, for any $\epsilon>0$ we have

$$P\Big(\frac{1}{n}\sum_{i=1}^n|I\{\hat\varepsilon_i\le y\}-I\{\varepsilon_i\le y\}|>\epsilon\Big)$$

$$\le P\Big(\Big|\frac{1}{n}\sum_{i=1}^n I\{y-\tau<\varepsilon_i\le y+\tau\}-(F(y+\tau)-F(y-\tau))\Big|>\epsilon/2\Big) \tag{B.9}$$

$$+P(|F(y+\tau)-F(y-\tau)|>\epsilon/2)+P(\max_{1\le i\le n}|\hat\varepsilon_i-\varepsilon_i|>\tau). \tag{B.10}$$

The term (B.9) converges to zero almost surely according to the strong law of large numbers. Because F is continuous the first term in term (B.10) is equal to zero for some sufficiently small τ. The second term in (B.10) converges to zero because of (B.3).
After showing the assertion for fixed y we include the supremum by a standard argument. The distribution function F is continued with $F(-\infty):=0, F(\infty):=1$. Then the real line is segmented into $-\infty=y_0<y_1<\ldots<y_{N-1}<y_N=\infty$ such that $|F(y_k)-F(y_{k-1})|<\epsilon/2$ for $k=1,\dots,N$. For each y exists a $k\in\{1,\dots,N\}$ such that $y\in[y_{k-1},y_k)$. The assertion follows using $|I\{\varepsilon_i\le y\}-I\{\hat\varepsilon_i\le y\}|\le|I\{\hat\varepsilon_i\le y_k\}-I\{\varepsilon_i\le y_{k-1}\}|+|I\{\hat\varepsilon_i\le y_{k-1}\}-I\{\varepsilon_i\le y_k\}|$ for an estimation of $\frac{1}{n}\sum_{i=1}^n|I\{\varepsilon_i\le y\}-I\{\hat\varepsilon_i\le y\}|$ through expressions with fixed $y_k, 1\le k\le N$, using the first part of the proof. □

C Appendix: Proofs of main results

Proof of Proposition 3.1

$\hat\eta_n$ is a solution of equation (3.2) and this yields

$$0=\frac{1}{n}\sum_{i=1}^n g(\hat\varepsilon_i)-\frac{1}{n}\sum_{i=1}^n g(\hat\varepsilon_i)g(\hat\varepsilon_i)^T\hat\eta_n+\frac{1}{n}\sum_{i=1}^n(\hat\eta_n^Tg(\hat\varepsilon_i))^2\frac{g(\hat\varepsilon_i)}{1+\hat\eta_n^Tg(\hat\varepsilon_i)}. \tag{C.1}$$

The last term can be estimated as follows,

$$\Big|\Big|\frac{1}{n}\sum_{i=1}^n(\hat\eta_n^Tg(\hat\varepsilon_i))^2\frac{g(\hat\varepsilon_i)}{1+\hat\eta_n^Tg(\hat\varepsilon_i)}\Big|\Big|$$

$$\le ||\hat\eta_n||^2\frac{1}{n}\sum_{i=1}^n||g(\hat\varepsilon_i)||^2\max_{j=1,\dots,k}\max_{i=1,\dots,n}|g_j(\hat\varepsilon_i)|\max_{i=1,\dots,n}\Big|\frac{1}{1+\hat\eta_n^Tg(\hat\varepsilon_i)}\Big|=o_P(\frac{1}{\sqrt{n}})$$

using Lemma B.2 (i), (ii), Lemma B.1 (i) and (v). In the second term of (C.1) we can replace $\frac{1}{n}\sum_{i=1}^n g(\hat\varepsilon_i)g(\hat\varepsilon_i)^T$ by Σ according to Lemma B.1, (v). The assertion follows by isolating $\hat\eta_n$. □

Proof of Proposition 3.2

We use the following expansion similar to the beginning of the proof of Prop. 3.1,

$$\overline{F}_n(y) - \hat{F}_n(y) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \hat{\eta}_n^T g(\hat{\varepsilon}_i)\Big(1 - \hat{\eta}_n^T g(\hat{\varepsilon}_i) + \frac{(\hat{\eta}_n^T g(\hat{\varepsilon}_i))^2}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\Big) I\{\hat{\varepsilon}_i \le y\}.$$

To prove the assertion of the Proposition we have to show uniformly with respect to $y \in \mathbb{R}$

$$\text{(C.2)} \qquad \frac{1}{n}\sum_{i=1}^{n} g(\hat{\varepsilon}_i) I\{\hat{\varepsilon}_i \le y\} \;=\; U(y) + o_P(1)$$

$$\text{(C.3)} \qquad \frac{1}{n}\sum_{i=1}^{n} (\hat{\eta}_n^T g(\hat{\varepsilon}_i))^2 I\{\hat{\varepsilon}_i \le y\} \;=\; o_P(\frac{1}{\sqrt{n}})$$

$$\text{(C.4)} \qquad \frac{1}{n}\sum_{i=1}^{n} \frac{(\hat{\eta}_n^T g(\hat{\varepsilon}_i))^3}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)} I\{\hat{\varepsilon}_i \le y\} \;=\; o_P(\frac{1}{\sqrt{n}}).$$

To show (C.2) we use the expansion $\frac{1}{n}\sum_{i=1}^{n} g(\hat{\varepsilon}_i) I\{\hat{\varepsilon}_i \le y\} = A_n(y) + B_n(y) + C_n(y)$, where

$$\begin{aligned}
A_n(y) \;&=\; \frac{1}{n}\sum_{i=1}^{n} g(\varepsilon_i) I\{\varepsilon_i \le y\} \\
B_n(y) \;&=\; \frac{1}{n}\sum_{i=1}^{n} (g(\hat{\varepsilon}_i) - g(\varepsilon_i)) I\{\hat{\varepsilon}_i \le y\} \\
C_n(y) \;&=\; \frac{1}{n}\sum_{i=1}^{n} g(\varepsilon_i)(I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\}).
\end{aligned}$$

$A_n(y)$ converges uniformly to $U(y)$ almost surely, because the class $\{\varepsilon \mapsto g(\varepsilon) I\{\varepsilon \le y\} \mid y \in \mathbb{R}\}$ is VC-subgraph [see van der Vaart and Wellner (1996), Lemma 2.6.18 (vi), p. 147] and therefore forms a Glivenko–Cantelli class. Further we have uniformly in $y \in \mathbb{R}$

$$\begin{aligned}
||B_n(y)|| \;&\le\; \Big(\frac{1}{n}\sum_{i=1}^{n} ||g(\hat{\varepsilon}_i) - g(\varepsilon_i)||^2\Big)^{1/2} \Big(\frac{1}{n}\sum_{i=1}^{n} I\{\hat{\varepsilon}_i \le y\}\Big)^{1/2} \\
&\le\; \Big(\frac{1}{n}\sum_{i=1}^{n} (g_1(\hat{\varepsilon}_i) - g_1(\varepsilon_i))^2 + \ldots + (g_k(\hat{\varepsilon}_i) - g_k(\varepsilon_i))^2\Big)^{1/2} = o_P(1)
\end{aligned}$$

by Lemma B.1 (v). Furthermore,

$$\begin{aligned}
||C_n(y)|| \;&\le\; \Big(\frac{1}{n}\sum_{i=1}^{n} ||g(\varepsilon_i)||^2\Big)^{1/2} \Big(\frac{1}{n}\sum_{i=1}^{n} (I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\})^2\Big)^{1/2} \\
&=\; O_P(1)\Big(\frac{1}{n}\sum_{i=1}^{n} |I\{\hat{\varepsilon}_i \le y\} - I\{\varepsilon_i \le y\}|\Big)^{1/2} \;=\; o_P(1)
\end{aligned}$$

by Lemma B.3. The next assertion, (C.3), is valid because

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}_n^T g(\hat{\varepsilon}_i))^2 I\{\hat{\varepsilon}_i \leq y\} \\
\leq\ & ||\hat{\eta}_n||^2 \frac{1}{n}\sum_{i=1}^{n}||g(\hat{\varepsilon}_i)g^T(\hat{\varepsilon}_i)|| = O_P(n^{-1})O_P(1) = o_P(n^{-1/2})
\end{aligned}$$

according to Lemma B.1(iv) and Lemma B.2. The assertion (C.4) is valid because it holds that

$$\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{\eta}_n^T g(\hat{\varepsilon}_i))^3}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)} I\{\hat{\varepsilon}_i \leq y\}| \\
\leq\ & ||\hat{\eta}_n||^3 \max_{1\leq i\leq n, 1\leq j\leq k}|g_j(\hat{\varepsilon}_i)| \max_{1\leq i\leq n}\left|\frac{1}{1+\hat{\eta}_n^T g(\hat{\varepsilon}_i)}\right| \frac{1}{n}\sum_{i=1}^{n}||g(\hat{\varepsilon}_i)g^T(\hat{\varepsilon}_i)|| = o_P(n^{-1})
\end{aligned}$$

using Lemma B.2 (i) and (ii), Lemma B.1 (i) and (v). □

Proof of Theorem 3.3

The expansion of the process follows from Propositions 3.2, 3.1 and Lemma B.1 (ii). Because sums of Donsker classes are Donsker [van der Vaart and Wellner (1996), p. 192, Ex. 2.10.7.] we only need to show that the following function classes $\mathcal{F}_\ell$ ($\ell = 1, 2$) are Donsker,

$$\begin{aligned}
\mathcal{F}_1 &= \{\varepsilon \mapsto I\{\varepsilon \leq y\} - F(y) \mid y \in \mathbb{R}\} \\
\mathcal{F}_2 &= \{\varepsilon \mapsto f(y)\varepsilon - U(y)^T\Sigma^{-1}(g(\varepsilon) - E[g'(\varepsilon_1)]\varepsilon) \mid y \in \mathbb{R}\}.
\end{aligned}$$

$\mathcal{F}_1$ is Donsker by classical results. $\mathcal{F}_2$ is a subset of the at most $(k+1)$-dimensional vector space $\{\varepsilon \mapsto c_0\varepsilon + \sum_{j=1}^{k} c_j h_j(\varepsilon) \mid c_0, \ldots, c_k \in \mathbb{R}\}$ (with $h_j(\varepsilon) = g_j(\varepsilon) - E[g_j'(\varepsilon_1)]\varepsilon$) and is therefore a VC-class [van der Vaart and Wellner (1996), p. 146, Lemma 2.6.15]. Pointwise separability of $\mathcal{F}_2$ can be shown by a standard argument considering the countable subclass indexed by rational $y \in \mathbb{Q}$. Moreover, $\mathcal{F}_2$ has a square integrable envelope by assumptions (M2) and (A) and is therefore Donsker [van der Vaart and Wellner (1996), p. 141].

For the calculation of the covariances denote $H_i(y) := I\{\varepsilon_i \leq y\} - F(y) + f(y)\varepsilon_i - U(y)^T\Sigma^{-1}(g(\varepsilon_i) - E[g'(\varepsilon_1)]\varepsilon_i)$ so that $\sqrt{n}(\overline{F}_n(y) - F(y) - \overline{b}(y)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} H_i(y) + o_p(1)$ and $E[H_1(y)] = 0$. For the covariance one obtains

$$\begin{aligned}
&\mathrm{Cov}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} H_i(y), \frac{1}{\sqrt{n}}\sum_{j=1}^{n} H_j(z)) = E[H_1(y)H_1(z)] \\
=\ & E[(I\{\varepsilon_1 \leq y\} - F(y))(I\{\varepsilon_1 \leq z\} - F(z))] + f(y)f(z)E[\varepsilon_1^2] \\
&+U(y)^T\Sigma^{-1}E[(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)^T]U(z)\Sigma^{-1}
\end{aligned}$$

$$+E[(I\{\varepsilon_1 \le y\} - F(y))f(z)\varepsilon_1] + E[(I\{\varepsilon_1 \le z\} - F(z))f(y)\varepsilon_1]$$
$$-E[(I\{\varepsilon_1 \le y\} - F(y))U(z)^T\Sigma^{-1}(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)]$$
$$-E[(I\{\varepsilon_1 \le z\} - F(z))U(y)^T\Sigma^{-1}(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)]$$
$$-E[f(y)\varepsilon_1 U(z)^T\Sigma^{-1}(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)]$$
$$-E[f(z)\varepsilon_1 U(y)^T\Sigma^{-1}(g(\varepsilon_1) - E[g'(\varepsilon_1)]\varepsilon_1)]$$

which coincides with the asserted asymptotic covariance in Theorem 3.3. □

Proof of Theorem 3.7

Using Propositions 3.2, 3.1 and Lemma B.1 (iii) the proof follows analogously to the proof of Theorem 3.3. □

D Appendix: Proofs for the heteroscedastic model

Let $K_1, K_2 > 0$ denote constants such that $2K_1 \le \sigma(x) \le K_2/2$ for all $x \in [0,1]$ and the derivatives σ' and σ'' are also bounded by $K_2/2$ according to assumption (M). Then it follows that $P(K_1 \le \hat\sigma(x) \le K_2$ for all $x \in [0,1])$ converges to one. From the argumentation in Akritas and Van Keilegom (2001) we have that $\hat m - m$ and $\hat\sigma - \sigma$ are elements of $C_\rho^{1+\alpha}[0,1]$ for all $\rho > 0$ for $n \to \infty$ with probability one. A suitable choice of ρ yields that $(\hat m - m)/\hat\sigma$ and $(\hat\sigma - \sigma)/\hat\sigma$ are elements of $\mathcal{H} = C_\delta^{1+\alpha}[0,1]$ with constant δ from assumption (G2'), for $n \to \infty$ with probability one. This is assured by the next Lemma and is needed in the proof of Proposition 4.1.

Lemma D.1 *Let $\delta > 0$ and $K_1, K_2 > 0$ such that $s \in C_{K_2}^{1+\alpha}[0,1]$ with $\inf_{x\in[0,1]} |s(x)| \ge K_1$. Then there exists some $\rho > 0$ such that $\frac{h}{s} \in C_\delta^{1+\alpha}[0,1]$ for all $h \in C_\rho^{1+\alpha}[0,1]$.*

Proof. Let $h \in C_\rho^{1+\alpha}[0,1]$. Then from the definition of the function class we have that

$$\max\Big(\sup_{x\in[0,1]} |h(x)|, \sup_{x\in[0,1]} |h'(x)|\Big) + \sup_{x,y\in[0,1]} \frac{|h'(x) - h'(y)|}{|x-y|^\alpha} \le \rho \tag{D.1}$$

[van der Vaart and Wellner (1996, p. 154)] and the analogous inequality for function s and constant K_2. From this and the boundedness of s by K_1 from above it follows with technical but straightforward estimations omitted for the sake of brevity that

$$\max\Big(\sup_{x\in[0,1]} \Big|\frac{h}{s}(x)\Big|, \sup_{x\in[0,1]} \Big|(\frac{h}{s})'(x)\Big|\Big) + \sup_{x,y\in[0,1]} \frac{|(\frac{h}{s})'(x) - (\frac{h}{s})'(y)|}{|x-y|^\alpha} \le c\rho$$

with some constant c only dependent on K_1, K_2. The assertion follows by the choice $\rho = \delta/c$ from the definition of $C_\delta^{1+\alpha}[0,1]$. □

Proof of Proposition 4.1

The proof follows the lines of the proof of Lemma B.1 (ii). The function classes $\mathcal{G}_j$ [compare (B.5)] is now defined as

$$\mathcal{G}_j = \Big\{\tilde{g}_j : \mathbb{R} \times [0,1] \to \mathbb{R}, \tilde{g}_j(\varepsilon, x) = g_j(\varepsilon + h_1(x) + \varepsilon h_2(x)) - g_j(\varepsilon) \;\Big|\; h_1, h_2 \in \mathcal{H}\Big\},$$

where $\mathcal{H} = C_\delta^{1+\alpha}[0,1]$ with constant δ from assumption (G2'). We show in the following that $\mathcal{G}_j$ is Donsker. To this end let for $\xi > 0$, as in the proof of Lemma B.1 (ii), $h_1, \dots, h_\lambda$ [$\lambda = N(\tilde{\xi}, \mathcal{H}, ||\cdot||_\infty)$] denote the centers of a supremum-norm covering for $\mathcal{H}$ with radius $\tilde{\xi} = (\xi/(2C))^\kappa$ with constant C from assumption (G2'). Then we have $N_{[]}(\xi, \mathcal{G}_j, L_2(P)) = \lambda^2$ brackets

$$\Big[g_j(\varepsilon + h_\ell(x) + \varepsilon h_k(x)) - g_j(\varepsilon) - \tilde{g}_j^*(\varepsilon, x)\ ,\ g_j(\varepsilon + h_\ell(x) + \varepsilon h_k(x)) - g_j(\varepsilon) + \tilde{g}_j^*(\varepsilon, x)\Big],$$

$\ell, k \in \{1, \dots, \lambda\}$, where $\tilde{g}_j^*(\varepsilon, x) = \sup |g_j(\varepsilon + z_1 + \varepsilon z_2) - g_j(\varepsilon + \tilde{z}_1 + \varepsilon \tilde{z}_2)|$ and the supremum is built over $|z_1|, |z_2|, |\tilde{z}_1|, |\tilde{z}_2| \le \delta, |z_1 - \tilde{z}_1|, |z_2 - \tilde{z}_2| \le \tilde{\xi}$. Each bracket has $L_2(P)$ length less or equal to ξ by assumption (G2'). The bracketing integral (B.6) is finite with the same reasoning as in the proof of Lemma B.1 (ii).

Now, from Lemma D.1 we have that the probability that $(\hat{m} - m)/\hat{\sigma} \in \mathcal{H}$ and $(\hat{\sigma} - \sigma)/\hat{\sigma} \in \mathcal{H}$ converges to one. The rest of the proof follows as the proof of Lemma B.1 (ii) leading to the expansion

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n g_j(\hat{\varepsilon}_i) &= \frac{1}{n}\sum_{i=1}^n g_j\Big(\varepsilon_i + \frac{(m - \hat{m})(x)}{\hat{\sigma}(x)} + \varepsilon_i \frac{(\sigma - \hat{\sigma})(x)}{\hat{\sigma}(x)}\Big) \\
&= \frac{1}{n}\sum_{i=1}^n g_j(\varepsilon_i) + o_P(n^{-1/2}) \\
&\quad + \int\int \Big(g_j\Big(y + \frac{(m - \hat{m})(x)}{\hat{\sigma}(x)} + y\frac{(\sigma - \hat{\sigma})(x)}{\hat{\sigma}(x)}\Big) - g_j(y)\Big) f_X(x) f(y)\, dx\, dy.
\end{aligned}$$

By assumption (G1') we further obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n g_j(\hat{\varepsilon}_i) &= \frac{1}{n}\sum_{i=1}^n g_j(\varepsilon_i) + \int\int g_j'(y) \frac{m(x) - \hat{m}(x)}{\hat{\sigma}(x)} f_X(x) f(y)\, dx\, dy \\
&\quad + \int\int g_j'(y) y \frac{\sigma(x) - \hat{\sigma}(x)}{\hat{\sigma}(x)} f_X(x) f(y)\, dx\, dy + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n g_j(\varepsilon_i) - E[g_j'(\varepsilon_1)] \int \frac{\hat{m}(x) - m(x)}{\sigma(x)} f_X(x)\, dx \\
&\quad - E[\varepsilon_1 g_j'(\varepsilon_1)] \int \frac{\hat{\sigma}^2(x) - \sigma^2(x)}{2\sigma^2(x)} f_X(x)\, dx + o_P(n^{-1/2}),
\end{aligned}$$

where the last equality follows from the uniform almost sure convergence of $\hat{\sigma}$ to σ with rate $O((n^{-1}h^{-1}\log h^{-1})^{1/2})$ [confer Prop. 3, Akritas and Van Keilegom (2001)], the bandwidth conditions (H), and $\hat{\sigma} - \sigma = (\hat{\sigma}^2 - \sigma^2)/(2\sigma) - (\hat{\sigma} - \sigma)^2/(2\sigma)$, where the last term results in a negligible remainder. The rest of the proof follows by inserting the definitions of $\hat{m}$ from (2.2), $\hat{\sigma}^2$ from (4.3), and some straightforward calculations of expectations and variances. □

Proof of Theorem 4.2

The validity of Propositions 3.1 and 3.2 in the heteroscedastic model (4.1) under the assumptions of Theorem 4.2 can be shown following the steps of the proofs for the homoscedastic case. To this end one shows that Lemmas B.1 (i), (vi), (v), B.2 and B.3 hold as well under these assumptions. The proofs are analogous, using the following estimation for Lemmas B.1 (i) and B.3 and noting that (G2) follows from (G2'). We have

$$\max_{i=1,\dots,n} |\varepsilon_i - \hat{\varepsilon}_i| \le \sup_{x\in[0,1]} \left|\frac{\hat{m}(x) - m(x)}{\hat{\sigma}(x)}\right| + \max_{i=1,\dots,n} |\varepsilon_i| \sup_{x\in[0,1]} \left|\frac{\hat{\sigma}(x) - \sigma(x)}{\hat{\sigma}(x)}\right| = o_P(1)$$

where the last equality follows from $\max_{i=1,\dots,n} |\varepsilon_i| = o_P(n^{1/4})$ under assumption (M), the uniform rates of convergence, $\sup_{x\in[0,1]} |\hat{m}(x) - m(x)| = O((n^{-1}h^{-1}\log h^{-1})^{1/2})$ and $\sup_{x\in[0,1]} |\hat{\sigma}(x) - \sigma(x)| = O((n^{-1}h^{-1}\log h^{-1})^{1/2})$ a.s. [confer Prop. 3, Akritas and Van Keilegom (2001)], the boundedness of σ from zero, and the bandwidth conditions (H).
The rest of the proof is exactly the same as the proof of Theorem 3.3. □

E References

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

H. Bonnal, E. Renault (2004). *On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood.* Technical report. http://ideas.repec.org/p/cir/cirwor/2004s-18.html

F. Cheng (2004). *Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression.* J. Statist. Plann. Inference 119, 95–107.

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

H. Dette and I. Van Keilegom (2005). *A new test for the parametric form of the variance function in nonparametric regression.* Technical report.
http://www.rub.de/mathematik3/preprint.htm

M. D. Donsker (1952). *Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems.* Ann. Math. Statist. 23, 277–281.

T. DiCiccio, J.P. Romano (1989). *On adjustments based on the signed root of the empirical likelihood ratio statistic.* Biometrika 76, 447–456.

T. DiCiccio, P. Hall and J.P. Romano (1989). *Comparison of parametric and empirical likelihood functions.* Biometrika 76, 465–476.

T. DiCiccio, P. Hall and J.P. Romano (1991). *Empirical likelihood is Bartlett-correctable.* Annals of Statistics 19, 1053–1061.

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

M. Genz (2004). *Anwendungen des Empirischen Likelihood-Schätzers der Fehlerverteilung in AR(1)-Prozessen.* Dissertation, Justus-Liebig-Universität Giessen. (in German).
http://geb.uni-giessen.de/geb/volltexte/2005/2069/pdf/GenzMichael-2004-12-03.pdf

P. Hall, B. LaScala (1990). *Methodology and algorithms of empirical likelihood.* Internat. Statist. Rev. 58, 109–127.

P. Hall (1990). *Pseudo-likelihood Theory for Empirical Likelihood.* Ann. Statist. 18, 121–140.

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

Y. Kitamura (2001). *Asymptotic optimality of empirical likelihood for testing moment restrictions.* Econometrica 69, 1661–1672.

Y. Kitamura, G. Tripathi, H. Ahn (2004). *Empirical likelihood-based inference in conditional moment restriction models.* Econometrica 72, 1667–1714.

H. L. Koul (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*, second edition. Springer, New York.

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

U. Müller, A. Schick, W. Wefelmeyer (2004b). *Estimating functionals of the error distribution in parametric and nonparametric regression.* J. Nonparam. Statist. 16, 525–548.

U. Müller, A. Schick, W. Wefelmeyer (2004c). *Estimating the error distribution function in nonparametric regression.* Technical report, http://www.mi.uni-koeln.de/∼wefelm/preprints.html

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

E.-R. Nagel (2002). *Der Empirical-Likelihood-Schätzer für die Fehlerverteilung im linearen Regressionsmodell.* Diploma dissertation, supervisor: E. Häusler, Justus-Liebig-Universität Gießen (in German).

N. Neumeyer, H. Dette (2005). *A note on one-sided nonparametric analysis of covariance by ranking residuals.* Mathematical Methods of Statistics 14, 80–104.

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.

J. C. Pardo-Fernández, I. Van Keilegom, W. González-Manteiga (2004). *Comparison of regression curves based on the estimation of the error distribution.* Discussion Paper 0416. Institut de Statistique (Université catholique de Louvain), Louvain-la-Neuve.

W. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (2002). *Numerical Recipes in C++, The Art of Scientific Computing.* Cambridge University Press, 2nd edition.

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

G. Qin (1996). *Consistent Residuals Density Estimation in Nonparametric Regression Under m(n)-Dependent Sample.* Journal of Mathematical Research and Exposition 16, 1996, no 4, 5005–516.

G. Qin, S. Shi, G. Chai (1996) *Asymptotics of the residual density estimation in nonparametric regression under m(n)-dependence.* Applied Mathematics A Journal of Chinese Universities Ser. B 11, 1996, no 1, 59–76.

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

A. W. van der Vaart and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer, New York.

I. Van Keilegom, W. González–Manteiga, C. Sánchez Sellero (2004). *Goodness-of-fit tests in parametric regression based onestimation the error distribution.* preprint. Institut de Statistique (Université catholique de Louvain), Louvain-la-Neuve.

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

B. Zhang (1997). *Estimating a distribution function in the presence of auxiliary information.* Metrika 46, 221–244.

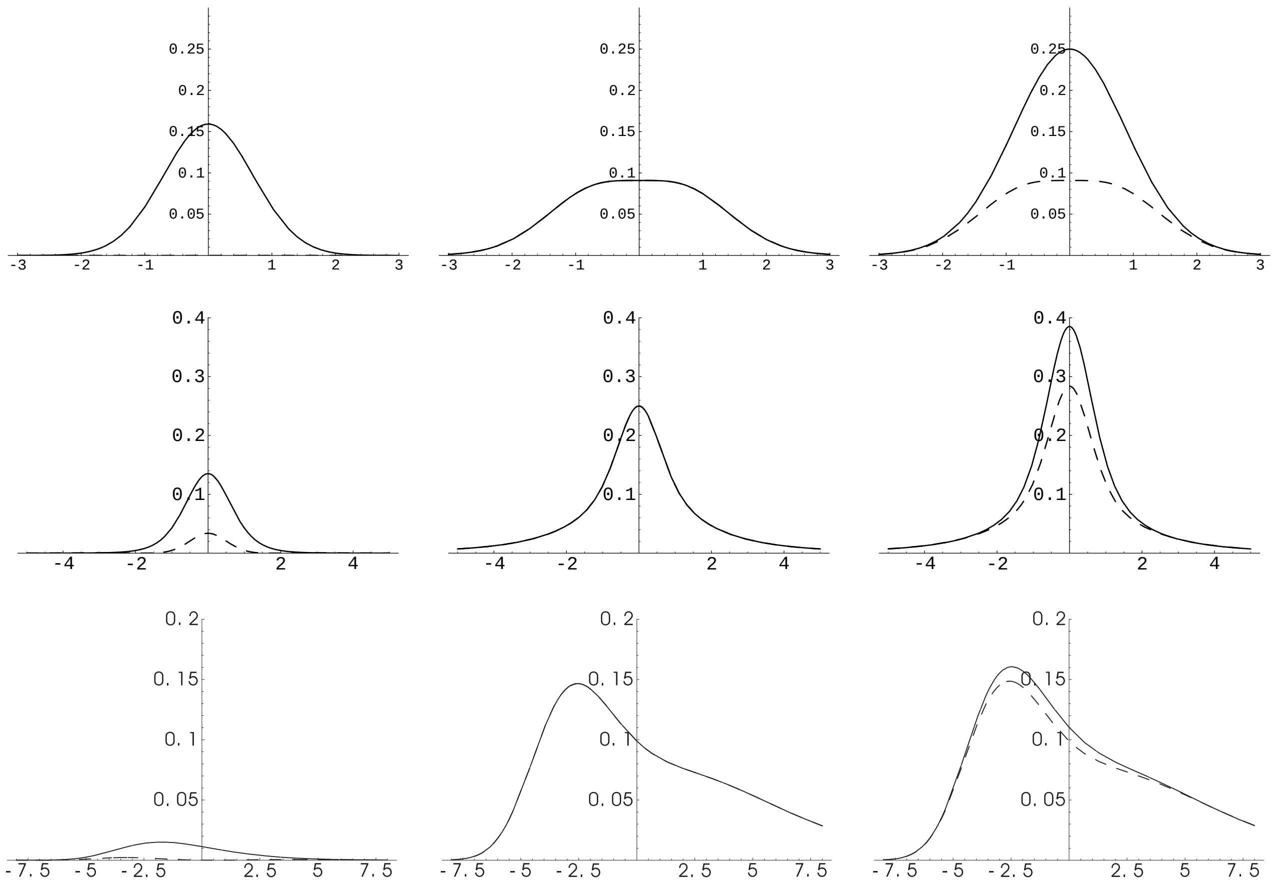

Figure 1: The figure shows curves for squared bias, variance and mean squared error for Example 5.1, i. e. $g(\varepsilon) = \varepsilon$. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. The first row corresponds to standard normally distributed errors. Here, the dashed bias curve vanishes. The second row shows results for student's t-distributed errors with three degrees of freedom. In the third row the errors are double exponentially distributed. In all three examples the two variance curves are identical.

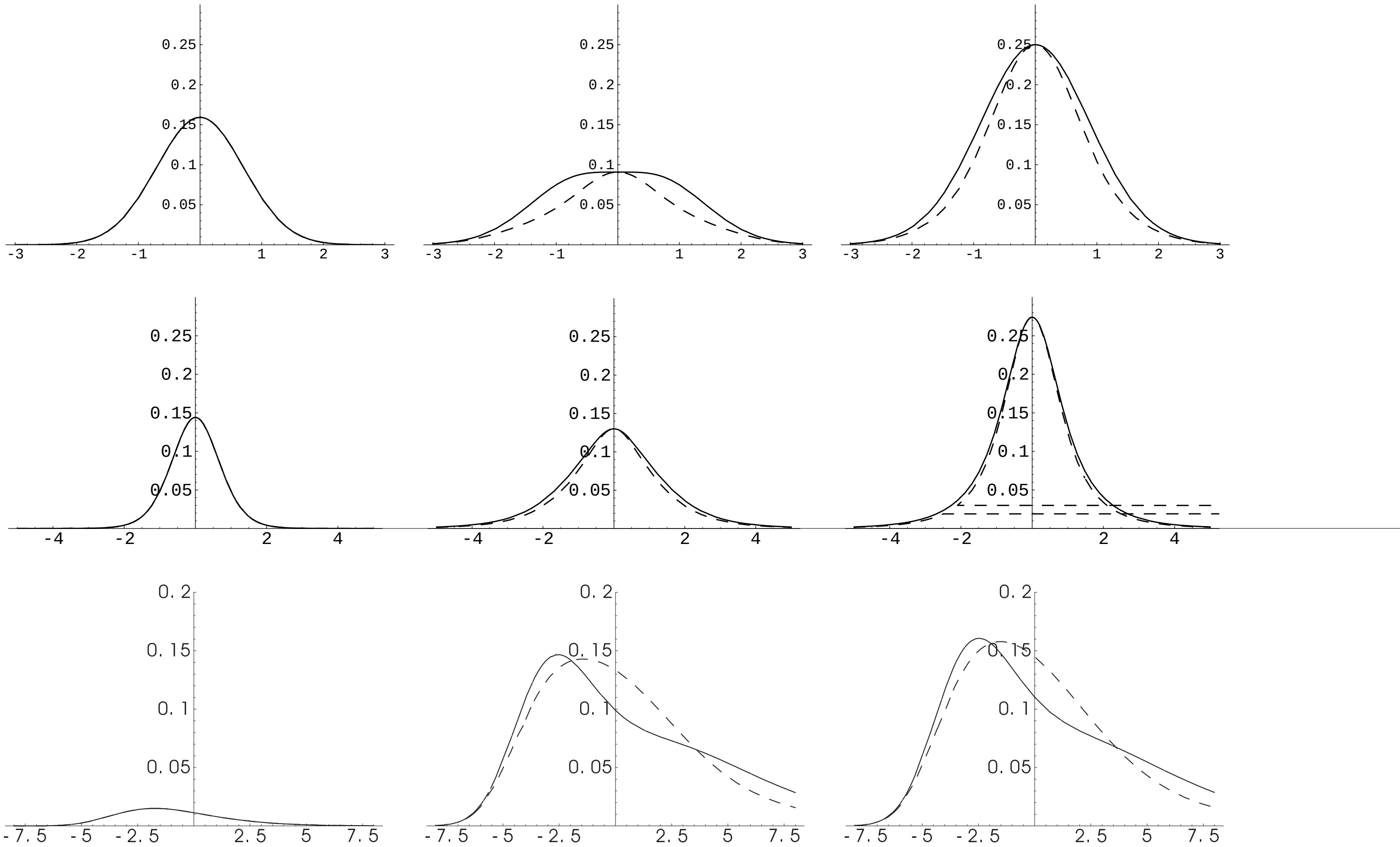

Figure 2: The figure shows curves for squared bias, variance and mean squared error for Example 5.2, i. e. $g(\varepsilon) = \varepsilon^2 - \sigma^2$. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. In the first row the errors are standard normally distributed. In the second row the errors are student's t-distributed with five degrees of freedom. The third row displays results for the double exponential distribution. The two bias curves are identical in all three examples.

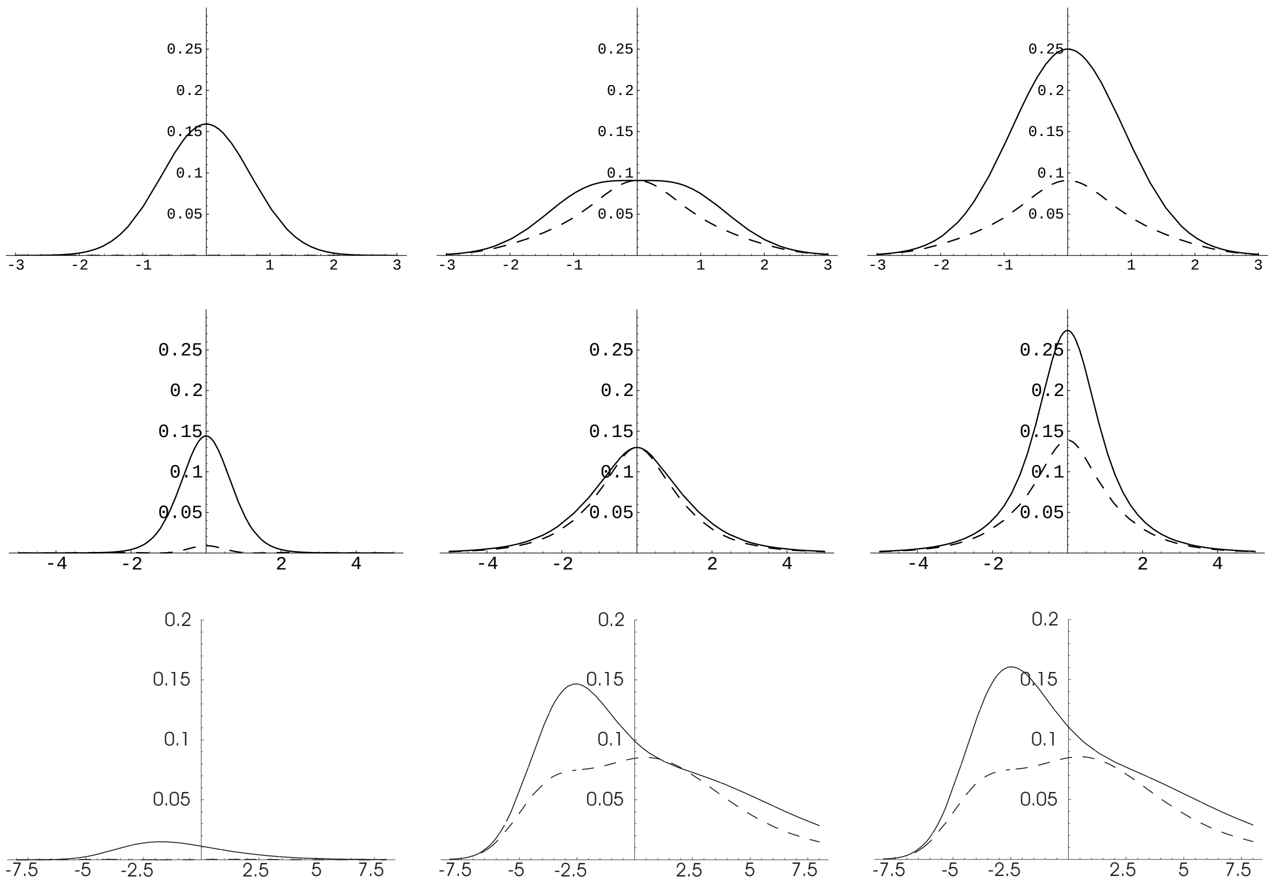

Figure 3: The figure shows curves for squared bias, variance and mean squared error for Example 5.3, i. e. $g(\varepsilon) = (\varepsilon, \varepsilon^2 - \sigma^2)^T$. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. In the first row the errors are standard normally distributed and the dashed bias curve is zero. In the second row results for student's t-distributed errors with five degrees of freedom are displayed, whereas in the third row the errors are double exponentially distributed.

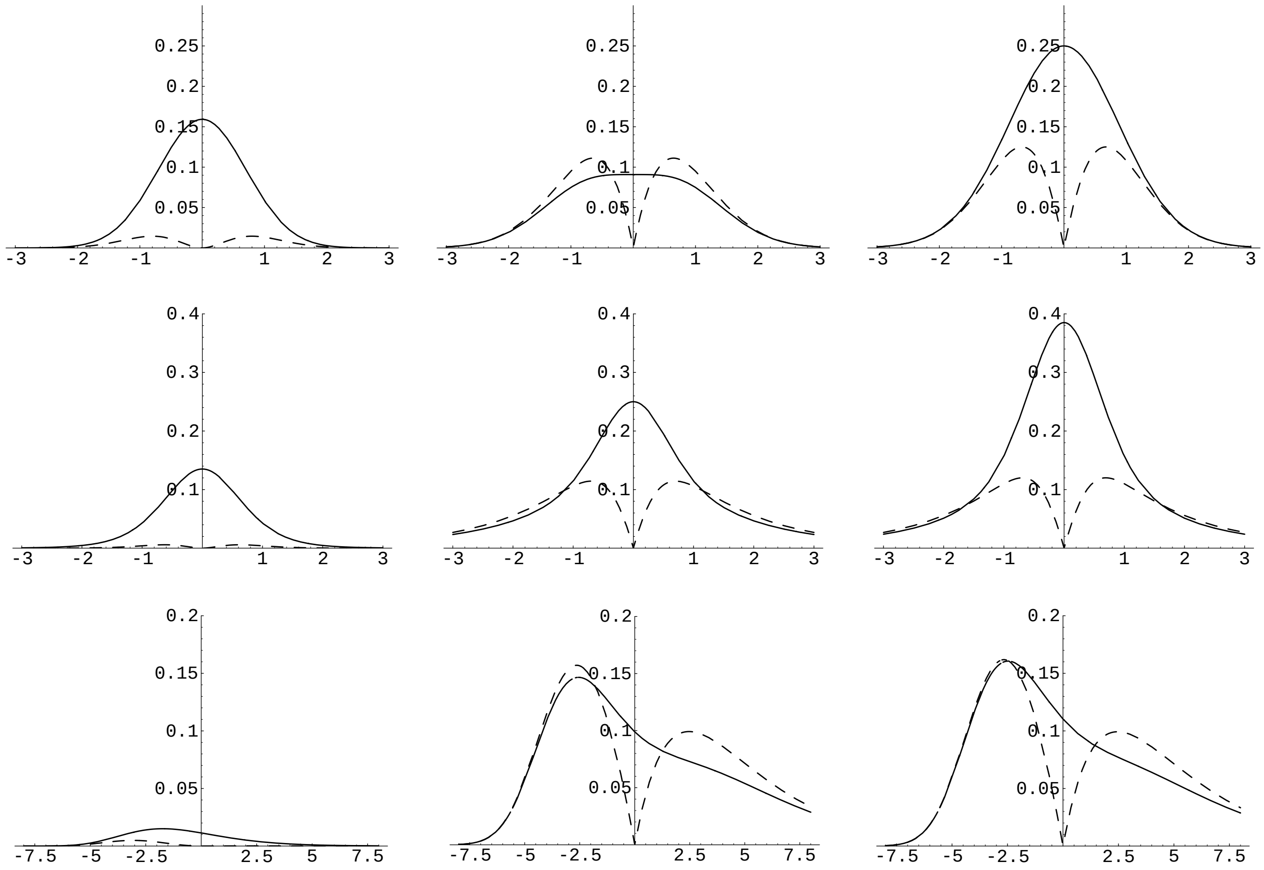

Figure 4: The figure shows curves for squared bias, variance and mean squared error for Example 5.4, i.e. $g(\varepsilon) = I\{\varepsilon \leq 0\} - b$. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. The first row corresponds to standard normally distributed errors ($b = 0.5$), the second row to student's t-distributed errors with three degrees of freedom ($b = 0.5$). In the third row we have double exponentially distributed errors ($b = 0.570371$).

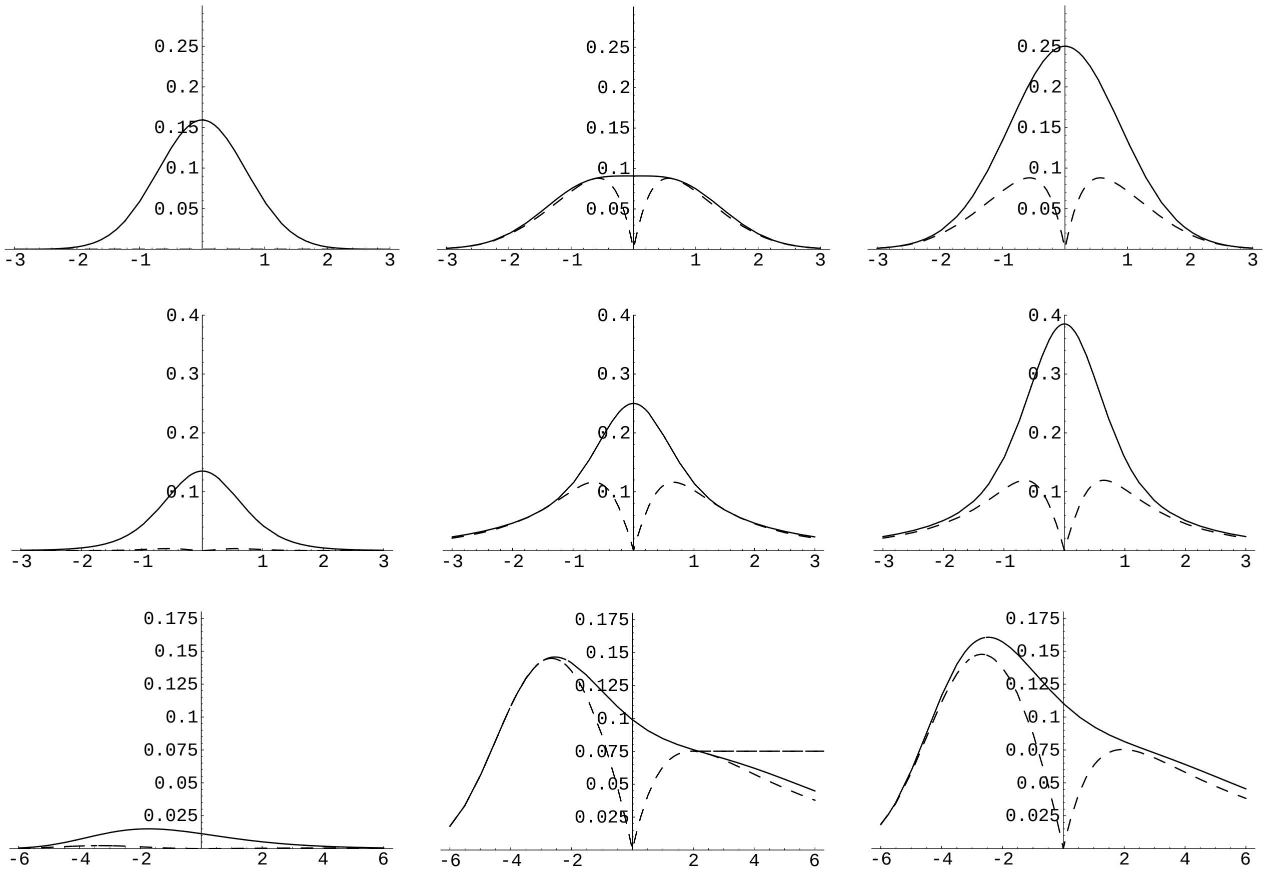

Figure 5: The figure shows curves for squared bias, variance and mean squared error for Example 5.5, i.e. $g(\varepsilon) = (\varepsilon, I\{\varepsilon \leq 0\} - b)^T$. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. The error distribution in the first row is standard normal ($b = 0.5$) and the dashed bias curve vanishes. In the second row the error distribution is student's t with three degrees of freedom ($b = 0.5$), whereas in the third row the errors are double exponentially distributed ($b = 0.570371$).

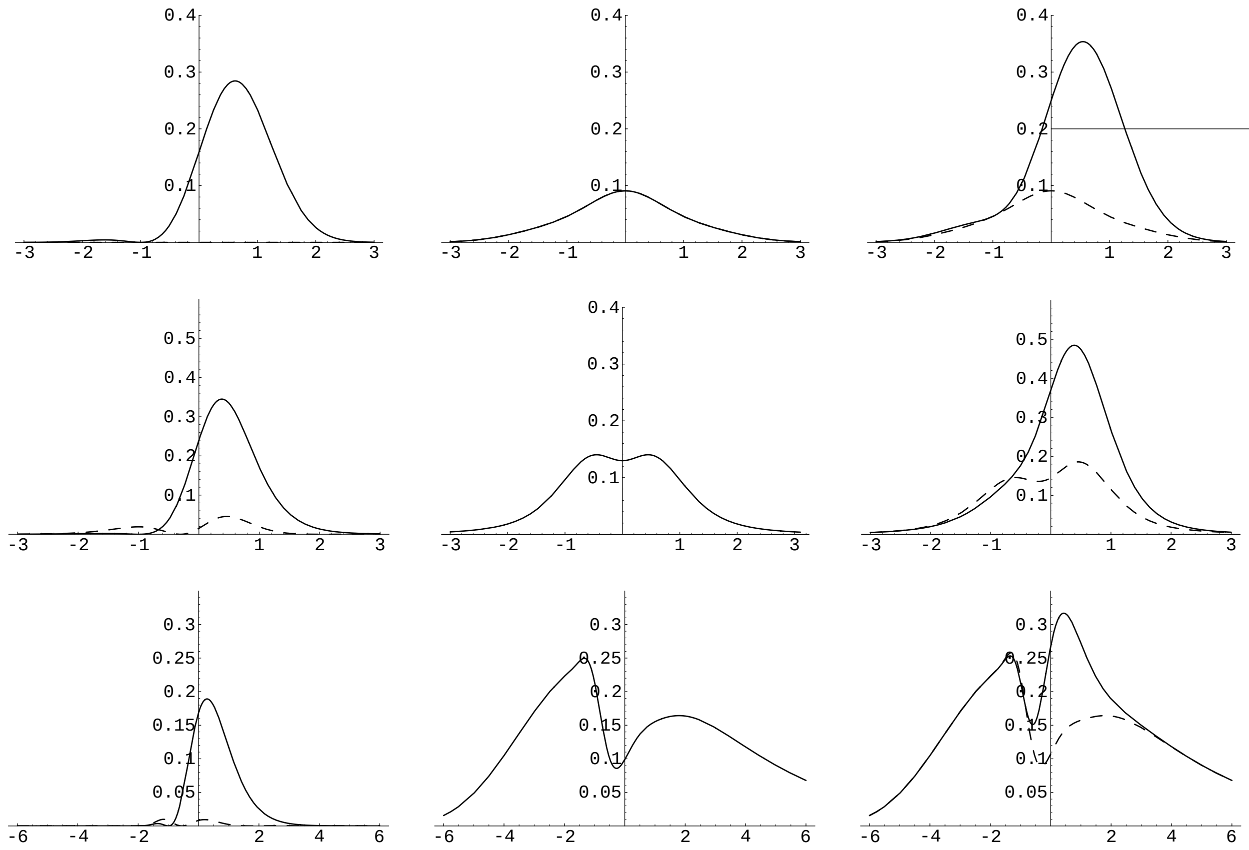

Figure 6: The figure shows curves for squared bias, variance and mean squared error for Example 5.6, i.e. $g(\varepsilon) = (\varepsilon, \varepsilon^2 - 1)^T$ in the heteroscedastic setting. The solid lines correspond to $\hat{F}_n$, the dashed lines to the new estimator $\overline{F}_n$. The error distribution in the first row is standard normal, in the second row standardized t-distribution with five degrees of freedom and the standardized double exponential distribution in the last row. The two variance curves coincide for all considered distributions. In the normal case the dashed bias curve vanishes.

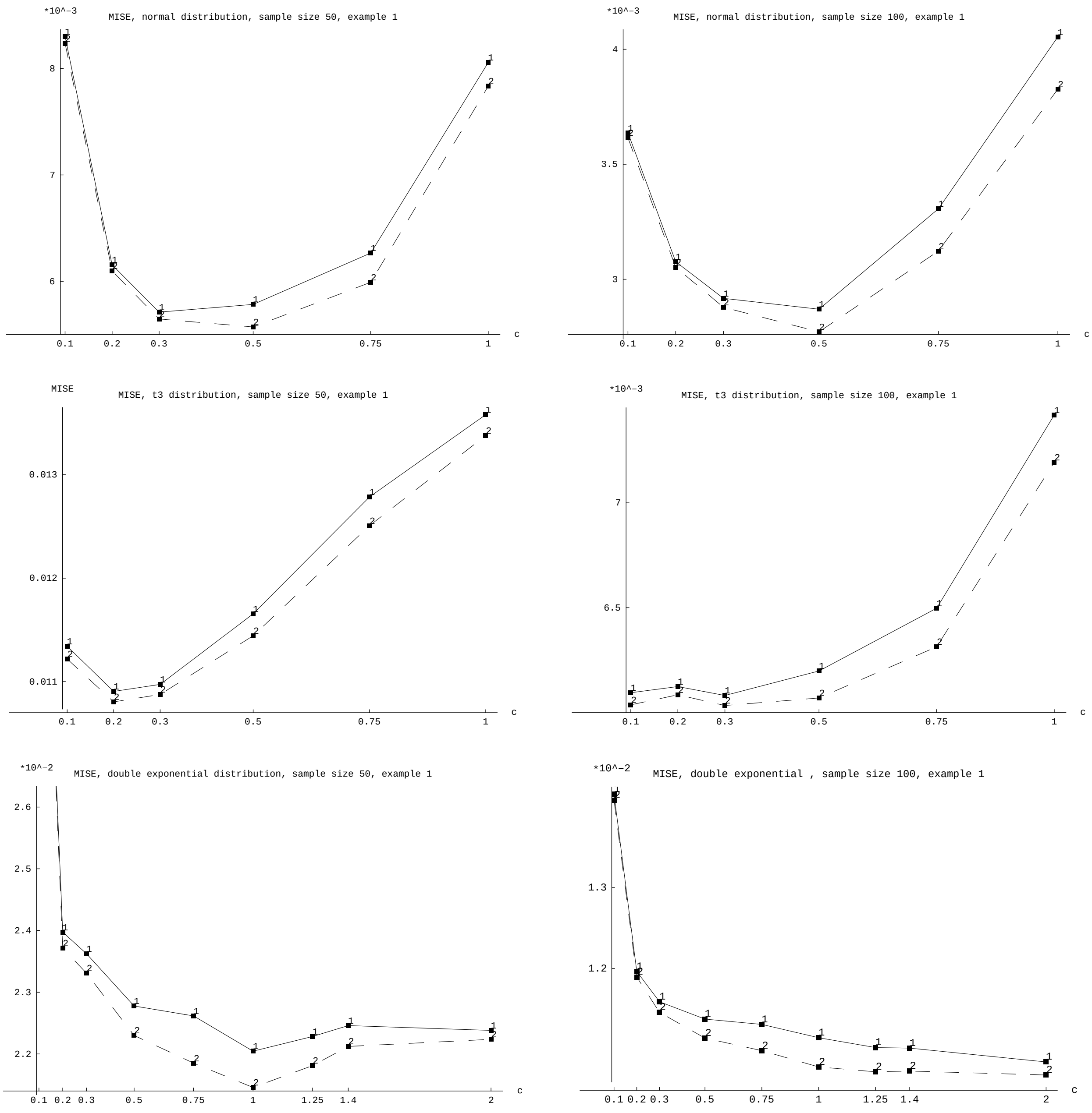

Figure 7: *Example 5.1.* The figure shows the MISE as a function of c, where for the bandwidth it holds that $h = c/n^{1/4}$. The sample size is $n = 50$ in the left panels and $n = 100$ in the right panels. The error distribution is standard normal in the first row, t_3 in the second, and double exponential in the third row. The solid curves (curves 1) correspond to $\hat{F}_n$, whereas the dashed curves (curves 2) correspond to $\overline{F}_n$.

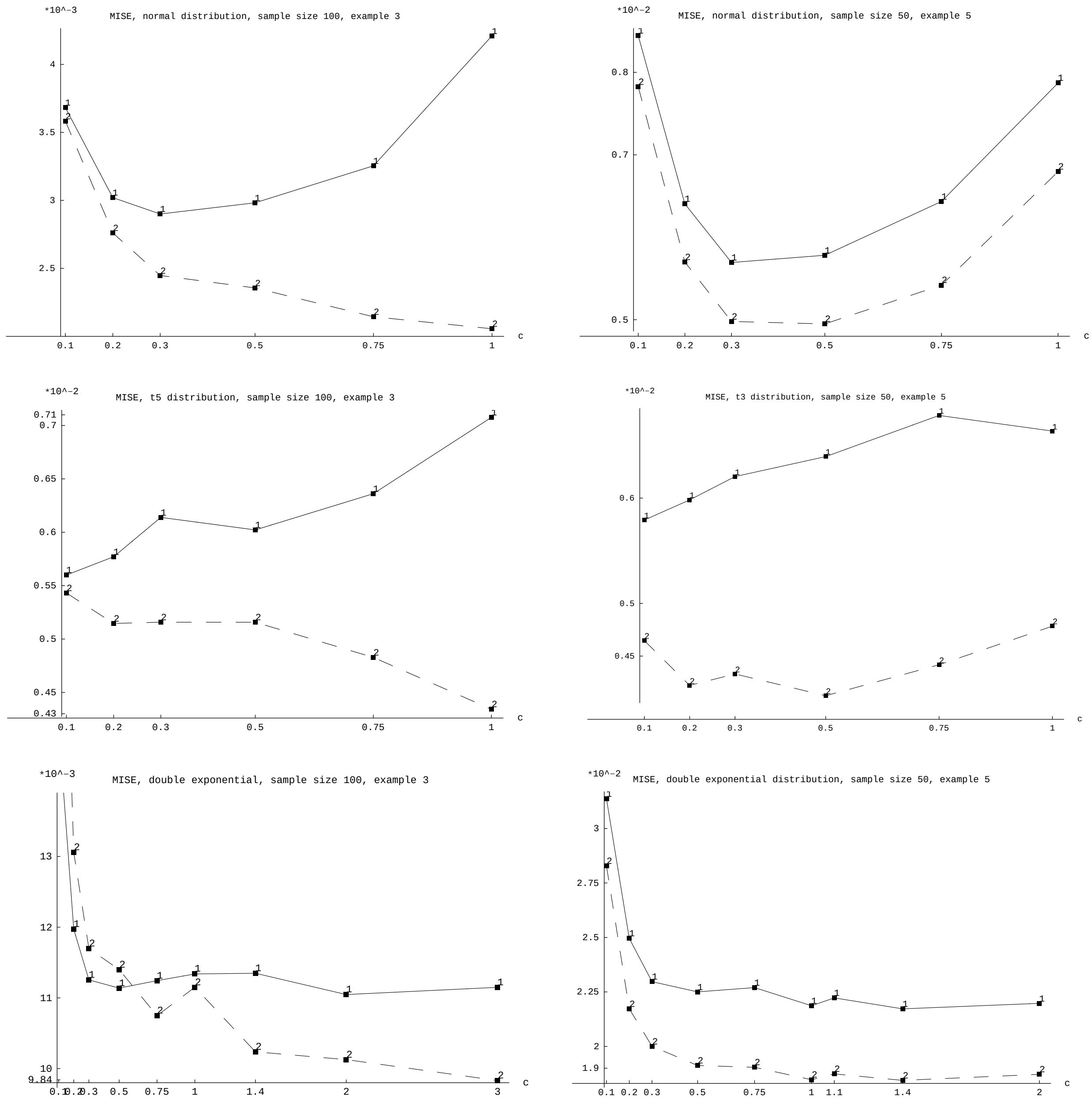

Figure 8: *Examples 5.3 (left panel) and 5.5 (right panel).* The figure shows the MISE as a function of c, where for the bandwidth it holds that $h = c/n^{1/4}$. The sample sizes are $n = 100$ in the left panel and $n = 50$ in the right panel. The error distribution is standard normal in the first row, t_5 resp. t_3 in the left and right panel of the second row, and double exponential in the third row. The solid curves (curves 1) correspond to $\hat{F}_n$, whereas the dashed curves (curves 2) correspond to $\overline{F}_n$.