The Annals of Statistics
2010, Vol. 38, No. 5, 2857–2883
DOI: 10.1214/10-AOS808

BACKFITTING AND SMOOTH BACKFITTING FOR ADDITIVE QUANTILE MODELS

By Young Kyung Lee[1], Enno Mammen[2] and Byeong U. Park[3]

Kangwon National University, University of Mannheim and Seoul National University

In this paper, we study the ordinary backfitting and smooth backfitting as methods of fitting additive quantile models. We show that these backfitting quantile estimators are asymptotically equivalent to the corresponding backfitting estimators of the additive components in a specially-designed additive mean regression model. This implies that the theoretical properties of the backfitting quantile estimators are not unlike those of backfitting mean regression estimators. We also assess the finite sample properties of the two backfitting quantile estimators.

1. Introduction. Nonparametric additive models are powerful techniques for high-dimensional data. They enable us to avoid the curse of dimensionality and estimate the unknown functions in high-dimensional settings at the same accuracy as in univariate cases. In the mean regression setting, there have been many proposals for fitting additive models. These include the ordinary backfitting procedure of Buja, Hastie and Tibshirani (1989), whose theoretical properties were studied later by Opsomer and Ruppert (1997) and Opsomer (2000), the marginal integration technique of Linton and Nielsen (1995), and the smooth backfitting of Mammen, Linton and Nielsen (1999), Mammen and Park (2006) and Yu, Park and Mammen (2008). It is widely accepted that the marginal integration method still suffers from the curse

Received October 2009; revised January 2010.
[1]Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0058380).
[2]Supported by DFG-project MA 1026/9-2 of the Deutsche Forschungsgemeinschaft.
[3]Supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. 2010-0017437).
AMS 2000 subject classifications. Primary 62G08; secondary 62G20.
Key words and phrases. Backfitting, nonparametric regression, quantile estimation, additive models.

of dimensionality since it does not produce rate-optimal estimates unless smoothness of the regression function increases with the number of additive components. On the contrary, the ordinary backfitting and smooth backfitting are known to achieve the univariate optimal rate of convergence under certain regularity conditions.

In this paper, we are concerned with nonparametric estimation of additive conditional quantile functions. Conditional quantile estimation is also a very useful tool for exploring the structure of the conditional distribution of a response given a predictor. A collection of conditional quantiles, when graphed, give a picture of the entire conditional distribution. It can be used directly to construct conditional prediction intervals. Also, it may be a basis for verifying the presence of conditional heteroscedasticity; see Furno (2004), for example. Various other applications of conditional quantile estimation may be found in Yu, Lu and Stander (2003). In the nonadditive setting, there have been many proposals for this problem, which include the work by Jones and Hall (1990), Chaudhuri (1991), Yu and Jones (1998) and Lee, Lee and Park (2006). There have been also some proposals for additive quantile regression. Fan and Gijbels (1996) provided a direct extension of the ordinary backfitting method to quantile regression, but without discussing its statistical properties. Lu and Yu (2004) gave a heuristic discussion of the asymptotic limit of a backfitting local linear quantile estimator. Horowitz and Lee (2005) studied an extension of the two-stage procedure of Horowitz and Mammen (2004) to quantile regression. Their estimator is a one-step kernel smoothing iteration of an orthogonal series estimator.

The main theme of this paper is to discuss the statistical properties of the ordinary and smooth backfitting methods in additive quantile regression. The methods are difficult to analyze since there exists no explicit definition for the ordinary backfitting estimator and, for both estimators, the objective functions defining the estimators are not differentiable. We borrow empirical process techniques to tackle the problem. In particular, we devise a theoretical mean regression model by using a Bahadur representation for the sample quantiles. We show that the least squares ordinary and smooth backfitting estimators in this theoretical mean regression model are asymptotically equivalent to the corresponding quantile estimators in the original model. This makes the theoretical properties of the two backfitting quantile estimators well understood from the existing theory for the corresponding least squares backfitting mean regression estimators. The theory was confirmed by a simulation study. Also, it was observed in the simulation study that the smooth backfitting estimator outperformed the ordinary backfitting estimator in additive quantile regression.

The paper is organized as follows. In the next section, the ordinary and smooth backfitting methods for additive quantile regression are introduced

and their theoretical properties are provided. In Section 3, some computational aspects of the smooth backfitting method are discussed. The simulation results for the finite sample properties of the two backfitting methods are presented in Section 4. Technical details are given in Section 5.

2. Main results. It is assumed for one-dimensional response variables $Y^1, \ldots, Y^n$ that

$$Y^i = m_0 + m_1(X_1^i) + \cdots + m_d(X_d^i) + \varepsilon^i, \qquad 1 \le i \le n. \tag{2.1}$$

Here, ε^i are error variables, $m_1, \ldots, m_d$ are unknown functions from $\mathbb{R}$ to $\mathbb{R}$ satisfying $\int m_j(x_j) w_j(x_j)\, dx_j = 0$ for some weight functions w_j, m_0 is an unknown constant, and $X^i = (X_1^i, \ldots, X_d^i)$ are random design points in $\mathbb{R}^d$. Throughout the paper, we assume that (X^i, ε^i) are i.i.d. and that X_j^i takes its values in a bounded interval I_j. Furthermore, it is assumed that the conditional α-quantile of ε^i given X^i equals zero. This model excludes interesting auto-regression models, but it simplifies our asymptotic analysis. We expect that our results can be extended to dependent observations under mixing conditions.

The ordinary backfitting estimator is based on an iterative algorithm. The estimate of m_j is updated by the following equation:

$$\hat{m}_j^{\mathrm{BF}}(x_j) = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \tau_\alpha \left(Y^i - \theta - \hat{m}_0^{\mathrm{BF}} - \sum_{\ell=1, \neq j}^d \hat{m}_\ell^{\mathrm{BF}}(X_\ell^i) \right) \times K_{j,h_j}(x_j, X_j^i). \tag{2.2}$$

Here, τ_α is the so called "check function" defined by $\tau_\alpha(u) = u\{\alpha - I(u < 0)\}$, and $K_{j,g}$ are kernel functions with bandwidth g; see the assumptions below. To simplify the mathematical argumentation, the minimization in (2.2) runs over a compact set Θ. It is assumed that all values of the function m_j lie in the interior of Θ. As in the case of mean regression, the ordinary backfitting estimator is not defined as a solution of a global minimization problem.

The smooth backfitting estimator is also based on an iterative algorithm. The estimate of m_j is updated by the following integral equation:

$$\hat{m}_j^{\mathrm{SBF}}(x_j) = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \int \tau_\alpha \left(Y^i - \theta - \hat{m}_0^{\mathrm{SBF}} - \sum_{\ell=1, \neq j}^d \hat{m}_\ell^{\mathrm{SBF}}(x_\ell) \right) \times \prod_{\ell=1, \neq j} K_{\ell,h_\ell}(x_\ell, X_\ell^i)\, dx_\ell \cdot K_{j,h_j}(x_j, X_j^i), \tag{2.3}$$

where the integration is over the support of $(X_1^i, \ldots, X_{j-1}^i, X_{j+1}^i, \ldots, X_d^i)$. This is an iterative scheme for obtaining $\hat{m}_j^{\mathrm{SBF}}, j = 0, 1, \ldots, d$, which mini-

mize

$$\sum_{i=1}^{n} \int \tau_\alpha \left(Y^i - \hat{m}_0^{\mathrm{SBF}} - \sum_{j=1}^{d} \hat{m}_j^{\mathrm{SBF}}(x_j) \right) \times K_{1,h_1}(x_1, X_1^i) \cdots K_{d,h_d}(x_d, X_d^i)\, dx_1 \cdots dx_d, \tag{2.4}$$

where the integration is over the support of X^i. The minimizations or iterations are done under the constraints

$$\int_{I_j} \hat{m}_j^l(x_j) w_j(x_j)\, dx_j = 0, \qquad j = 1, \ldots, d \text{ and } l = \mathrm{BF}, \mathrm{SBF} \tag{2.5}$$

for some weight functions w_j. One may take unknown weight functions such as the marginal densities of X_j and use consistent estimators of them as the weight functions w_j in the integrals (2.5). But this would lead to more complicated bias calculation.

We compare our model (2.1) with the following theoretical model. For $i = 1, \ldots, n$, let $Z^1, \ldots, Z^n$ be one-dimensional variables such that

$$Z^i = m_0 + m_1(X_1^i) + \cdots + m_d(X_d^i) + \eta^i. \tag{2.6}$$

Here, the constant m_0, the functions $m_1, \ldots, m_d$ and the covariates $X_1^i, \ldots, X_d^i$ are those in (2.1). The error variables η^i are defined by

$$\eta^i = -\frac{I(\varepsilon^i \le 0) - \alpha}{f_{\varepsilon|X}(0|X^i)},$$

where $f_{\varepsilon|X}$ is the conditional density of ε given X. This definition is motivated from the Bahadur representation of sample quantiles [Bahadur (1966)]. For an independent sample of $\varepsilon^1, \ldots, \varepsilon^n$ with densities f_i and α-quantiles being equal to 0, the Bahadur expansion states that the αth sample quantile $\hat{\theta}_\alpha$ of $\varepsilon^1, \ldots, \varepsilon^n$ is asymptotically equivalent to the weighted average

$$\frac{\sum_{i=1}^n f_i(0)\eta^i}{\sum_{i=1}^n f_i(0)},$$

where $\eta^i = -\{I(\varepsilon^i \le 0) - \alpha\} f_i(0)^{-1}$. Thus, the estimator $\hat{\theta}_\alpha$ is asymptotically equivalent to the minimizer of

$$\theta \to \sum_{i=1}^{n} f_i(0)(\eta^i - \theta)^2.$$

This consideration suggests that the ordinary and smooth backfitting estimators defined at (2.2) and (2.3), respectively, may be approximated well by the corresponding weighted local least squares estimators in the model (2.6). Note that the model (2.6) is an additive model with errors η^i having

conditional mean zero given the covariates X^i. Thus, the weighted ordinary backfitting estimators $\hat{m}_j^{*,\mathrm{BF}}$ in this model are defined by the following iterations:

$$\begin{aligned}\hat{m}_j^{*,\mathrm{BF}}(x_j) &= \operatorname*{arg\,min}_{\theta\in\Theta} \sum_{i=1}^{n} \left\{ Z^i - \theta - \hat{m}_0^{*,\mathrm{BF}} - \sum_{\ell=1,\neq j}^{d} \hat{m}_\ell^{*,\mathrm{BF}}(X_\ell^i) \right\}^2 \\ &\qquad \times f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) \\ &= \sum_{i=1}^{n} \left\{ Z^i - \hat{m}_0^{*,\mathrm{BF}} - \sum_{\ell=1,\neq j}^{d} \hat{m}_\ell^{*,\mathrm{BF}}(X_\ell^i) \right\} f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) \\ &\qquad \times \left\{ \sum_{i=1}^{n} f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) \right\}^{-1}.\end{aligned} \tag{2.7}$$

Also, the weighted smooth backfitting estimators $\hat{m}_j^{*,\mathrm{SBF}}$ in the model (2.6) are defined by

$$\begin{aligned}\hat{m}_j^{*,\mathrm{SBF}}(x_j) &= \tilde{m}_j^{*,\mathrm{SBF}}(x_j) - \hat{m}_0^{*,\mathrm{SBF}} \\ &\qquad - \sum_{\ell=1,\neq j}^{d} \int \hat{m}_\ell^{*,\mathrm{SBF}}(x_\ell) \frac{\hat{f}^w_{X_j,X_\ell}(x_j, x_\ell)}{\hat{f}^w_{X_j}(x_j)}\, dx_\ell,\end{aligned} \tag{2.8}$$

where

$$\tilde{m}_j^{*,\mathrm{SBF}}(x_j) = n^{-1} \sum_{i=1}^{n} Z^i f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) \hat{f}^w_{X_j}(x_j)^{-1},$$

$$\hat{f}^w_{X_j}(x_j) = n^{-1} \sum_{i=1}^{n} f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i),$$

$$\hat{f}^w_{X_j,X_\ell}(x_j, x_\ell) = n^{-1} \sum_{i=1}^{n} f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) K_{\ell,h_\ell}(x_\ell, X_\ell^i)$$

are weighted modifications of the marginal Nadaraya–Watson estimator and the kernel estimators of the one- and two-dimensional marginal densities of X, respectively. The latter two are in fact kernel estimators of

$$f^w_{X_j}(x_j) = \int f_{\varepsilon|X}(0|x) f_X(x)\, dx_{-j} = f_{\varepsilon,X_j}(0, x_j),$$

$$f^w_{X_j,X_\ell}(x_j, x_\ell) = \int f_{\varepsilon|X}(0|x) f_X(x)\, dx_{-(j,\ell)} = f_{\varepsilon,X_j,X_\ell}(0, x_j, x_\ell),$$

respectively, where $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)^\top$ and $x_{-(j,\ell)}$ is a vector that has elements x_l with $1 \le l \le d$ and $l \neq j, \ell$.

Our first result (Proposition 2.1) shows that each application of the updating equations (2.7) and (2.8) in the theoretical model (2.6), respectively, lead to asymptotically equivalent results with those at (2.2) and (2.3) in the original model (2.1). In the next step, we will apply Proposition 2.1 for iterative applications of the backfitting updates. We will show that the asymptotic equivalence remains to hold for iterative applications of the backfitting procedures as long as the number of iterations is small enough. By extending the results for backfitting and smooth backfitting estimators in mean regression, we will use this fact to get our main result (Theorem 2.2). The latter states an asymptotic normality result for the ordinary and smooth backfitting quantile estimators in additive models. Its proof is based on an argument that carries an asymptotic normality result in mean regression over to quantile regression.

We now introduce assumptions that guarantee asymptotic equivalence between the mean and the quantile backfitting estimators after one cycle of update. Further assumptions that are needed for iterative updates will be given after Proposition 2.1. For simplicity, we state Proposition 2.1 and its conditions only for the updates of the first additive component. In abuse of notation, we denote the estimators of the components $m_j, 2 \le j \le d$, at the preceding iteration step, by $\hat{m}_2^l, \dots, \hat{m}_d^l$, where l stands for BF, SBF, $*$,BF or $*$,SBF. The updates of the first component that are obtained by plugging these estimators into (2.2), (2.3), (2.7) and (2.8), respectively, are denoted by $\hat{m}_1^{\mathrm{BF}}, \hat{m}_1^{\mathrm{SBF}}, \hat{m}_1^{*,\mathrm{BF}}$ and $\hat{m}_1^{*,\mathrm{SBF}}$. Thus, for simplicity of notation, we use the same kind of symbol for the updates ($j = 1$) and for the inputs of the backfitting algorithms ($2 \le j \le d$).

We make the following assumptions:

(A1) The d-dimensional vector X^i has compact support $I = I_1 \times \cdots \times I_d$ for bounded intervals $I_j = [a_j, b_j]$ and its density f_X is continuous and strictly positive on I.

(A2) There exist constants $C_K, C_S > 0$ such that for all $x_j \in I_j$, $1 \le j \le d$, the kernels $K_{j,g}(x_j, \cdot)$ are positive, bounded by $C_K g^{-1}$, have bounded support $\subset [x - C_S g, x + C_S g]$, and are Lipschitz continuous with Lipschitz constant bounded by $C_K g^{-2}$. The weight functions w_j are bounded functions with $w_j(x_j) \ge 0$ for $x_j \in I_j$ and $\int w_j(x_j)\, dx_j > 0$.

(A3) The conditional density $f_{\varepsilon|X}(0|x)$ of ε given $X = x$ is bounded away from zero and infinity for $x \in I$. Furthermore, it satisfies the following uniform Lipschitz condition:

$$|f_{\varepsilon|X}(e|x) - f_{\varepsilon|X}(0|x)| \le C_1 |e|$$

for $x \in I$ and for e in a neighborhood of 0 with a constant $C_1 > 0$ that does not depend on x.

(A4) The bandwidths $h_1, \dots, h_d$ are of order $n^{-1/5}$.

Assumptions (A1)–(A4) are standard smoothing assumptions. In particular, (A2) is fulfilled for convolution kernels with an appropriate boundary correction.

For the properties of the updated estimators, the estimators at the preceding iteration step need to fulfill certain regularity conditions. We will proceed with the following assumptions that are stated for some constants $0 < \rho \leq 1$, $\Delta_1, \Delta_2, \Delta_3 > 0$ and $0 \leq \xi \leq (1+\rho)\Delta_1$.

(A5) For $j = 2, \ldots, d$, it holds for $l = \text{BF}$ and $l = \text{SBF}$ that

$$\sup_{a_j + C_S h_j \leq x_j \leq b_j - C_S h_j} |\hat{m}_j^l(x_j) - m_j(x_j)| = O_P(n^{-(4+4\rho)/(10+15\rho)-\Delta_1}),$$

$$\sup_{a_j \leq x_j \leq b_j} |\hat{m}_j^l(x_j) - m_j(x_j)| = O_P(n^{-[(4+4\rho)/(10+15\rho)-\Delta_1]/2}).$$

(A6) There exist random functions $g_2, \ldots, g_d$ with derivatives that fulfill the Lipschitz condition

$$|g_j'(x_j) - g_j'(x_j^*)| \leq C|x_j - x_j^*|^\rho n^\xi$$

for $j = 2, \ldots, d$ and $x_j, x_j^* \in I_j$. Furthermore, these functions satisfy

$$\sup_{a_j \leq x_j \leq b_j} |\hat{m}_j^l(x_j) - g_j(x_j)| = O_P(n^{-2/5-\Delta_2})$$

for $l = \text{BF}$ and $l = \text{SBF}$.

(A7) For $j = 2, \ldots, d$, it holds for $l = \text{BF}$ and $l = \text{SBF}$ that

$$\sup_{a_j + C_S h_j \leq x_j \leq b_j - C_S h_j} |\hat{m}_j^l(x_j) - \hat{m}_j^{*,l}(x_j)| = O_P(n^{-2/5-\Delta_3}),$$

$$\sup_{a_j \leq x_j \leq b_j} |\hat{m}_j^l(x_j) - \hat{m}_j^{*,l}(x_j)| = O_P(n^{-1/5-\Delta_3}).$$

We briefly comment on the assumptions (A5)–(A7). A more detailed discussion is given after Theorem 2.2. Assumption (A5) requires suboptimal rates for the preceding estimators that are plugged in for the update of the first component. Assumption (A6) states that the class of possible realizations of the preceding estimators is not too rich. We assume that the preceding estimators are in a neighborhood of the class of functions with Lipschitz continuous derivatives. Other classes could be used but for a Lipschitz class it is relatively easy to check if a function belongs to it. Note that we do not assume that the quantile estimator itself has a smooth derivative. In general, such an assumption does not hold because quantile kernel estimators are not smooth. Assumption (A7) is very natural. It states that the estimators that are plugged into the updating equation of the quantile model and of the mean regression model differ only by second order terms.

Without this assumption, it cannot be expected that the updated estimators differ also only by second order terms. We will see below that this assumption is automatically fulfilled if we apply Proposition 2.1 for an analysis of iterative applications of the backfitting algorithms. In the assumptions (A5) and (A7), if one replaces the interior region $[a_j + C_S h_j, b_j - C_S h_j]$ by the whole range $[a_j, b_j]$ and if one uses boundary corrected kernels, then one can also replace in Proposition 2.1 the suprema over the interior region by those over the whole range, and the estimators achieve the rate $n^{-2/5}$ at the boundary, too.

PROPOSITION 2.1. *Under the assumptions* (A1)–(A7)*, it holds for the updated estimators with* $l = \mathrm{BF}$ *and with* $l = \mathrm{SBF}$ *that for some* $\delta > 0$

$$\sup_{a_1 + C_S h_1 \le x_1 \le b_1 - C_S h_1} |\hat{m}_1^l(x_1) - \hat{m}_1^{*,l}(x_1)| = O_P(n^{-2/5-\delta}),$$

$$\sup_{a_1 \le x_1 \le b_1} |\hat{m}_1^l(x_1) - \hat{m}_1^{*,l}(x_1)| = O_P(n^{-1/5-\delta}).$$

The additional factor $n^{-\delta}$ allows an iterative application of the proposition. This has an important implication. We recall that the backfitting algorithms for mean regression have a geometric rate of convergence. In particular, in the case of smooth backfitting, only square integrability for the initial estimator is required for the algorithm to achieve the geometric rate of convergence, see Theorem 1 of Mammen, Linton and Nielsen (1999). Suppose one chooses square integrable functions, say $\hat{m}_2^{\mathrm{BF},[0]}, \ldots, \hat{m}_d^{\mathrm{BF},[0]}$ as the starting value in the algorithm for the backfitting quantile estimator and that one runs a cycle of backfitting iterations (2.2) for $j = 1, \ldots, d$. Then we get updates $\hat{m}_2^{\mathrm{BF},[l]}, \ldots, \hat{m}_d^{\mathrm{BF},[l]}$ with $l = 1$ and after further cycles with $l > 1$. (Note that by construction of the backfitting estimator we do not need a pilot version of $m_1^{\mathrm{BF},[0]}$.) Then, one can think of running the backfitting mean regression algorithm (2.7) with the same initial estimators $\hat{m}_2^{\mathrm{BF},[0]}, \ldots, \hat{m}_d^{\mathrm{BF},[0]}$ in parallel with the backfitting quantile regression algorithm (2.2). This results in updates $\hat{m}_2^{*,\mathrm{BF},[l]}, \ldots, \hat{m}_d^{*,\mathrm{BF},[l]}$ for $l \ge 1$. In the proof of our next theorem, we will see that after l cycles of the two parallel iterations, the difference $\hat{m}_j^{\mathrm{BF},[l]} - \hat{m}_j^{*,\mathrm{BF},[l]}$ is of order $O_P(n^{-2/5-\delta})$ in the interior, and of order $O_P(n^{-1/5-\delta})$ at the boundaries. This holds as long as $l \le C_{\mathrm{iter}} \log n$ with C_{iter} small enough. On the other hand, we will show that $\hat{m}_j^{*,\mathrm{BF},[C_{\mathrm{iter}} \log n]}$ is asymptotically equivalent to the limit of the backfitting algorithm $\hat{m}_j^{*,\mathrm{BF},[\infty]}$, if C_{iter} is large enough. If the pilot estimators $\hat{m}_2^{\mathrm{BF},[0]}, \ldots, \hat{m}_d^{\mathrm{BF},[0]}$ are accurate enough, then the constant C_{iter} can be chosen such that both requirements are fulfilled. This will allow us to

get the asymptotic limit distribution of $\hat{m}_j^{*,\mathrm{BF},[C_{\mathrm{iter}}\log n]}$, and thus that of $\hat{m}_j^{\mathrm{BF},[C_{\mathrm{iter}}\log n]}$.

Similar findings also hold for the smooth backfitting estimator. We denote the starting values by $\hat{m}_2^{\mathrm{SBF},[0]},\dots,\hat{m}_d^{\mathrm{SBF},[0]}$ and the updates by $\hat{m}_2^{\mathrm{SBF},[l]},\dots,\hat{m}_d^{\mathrm{SBF},[l]}$ or $\hat{m}_2^{*,\mathrm{SBF},[l]},\dots,\hat{m}_d^{*,\mathrm{SBF},[l]}$, respectively.

The following theorem summarizes our discussion. For the theorem, we need the following additional assumptions:

(A8) There exist constants $c_K, C_D > 0$, $C_S' \geq 0$ such that for $a_j + C_S' h_j \leq x_j, u_j \leq b_j - C_S' h_j$ it holds that $K_{j,h_j}(x_j,u_j) = h_j^{-1} K[h_j^{-1}(x_j - u_j)]$ for a function K with $\int K(v)\,dv = 1$ and $\int vK(v)\,dv = 0$. For all $x_j, u_j \in I_j$, $1 \leq j \leq d$, the kernels $K_{j,g}(x_j,u_j)$ have a second derivative w.r.t. x_j that is bounded by $C_D g^{-3}$ and they fulfill $\int K_{j,g}(x_j,v_j)\,dv_j \geq c_K$ and $\int K_{j,g}(v_j,u_j)\,dv_j = 1$.

(A9) The function $f^w_{X_k|X_j}(x_k|x_j) \equiv f^w_{X_j,X_k}(x_j,x_k)/f^w_{X_j}(x_j)$ has a second derivative w.r.t. x_j that is bounded over $x_j \in I_j$, $x_k \in I_k$, $1 \leq j,k \leq d$, $k \neq j$.

The last condition in (A8) implies that the one-dimensional kernel density estimators integrate to one and that they are equal to the corresponding marginalization of higher-dimensional product-kernel density estimators. This assumption simplifies bias calculation of the backfitting estimators.

THEOREM 2.2. *Assume that* (A1)–(A4), (A8) *and* (A9) *hold, and that* (A5) *and* (A6) *are satisfied by* $\hat{m}_j^{\mathrm{BF}} = \hat{m}_j^{\mathrm{BF},[0]}$ *and* $\hat{m}_j^{\mathrm{SBF}} = \hat{m}_j^{\mathrm{SBF},[0]}$ *($j = 2,\dots,d$) with* $\xi, \Delta_2, \Delta_3, \frac{2}{5} - \frac{1+\rho}{2+3\rho}\frac{4}{5} - \Delta_1 > 0$ *small enough. Then, we get for* $\hat{m}_j^{l,\mathrm{iter}} = \hat{m}_j^{l,[C_{\mathrm{iter}}\log n]}$ *with an appropriate choice of* $C_{\mathrm{iter}} = C_{\mathrm{iter},l}$ *($l = \mathrm{BF}$ and $l = \mathrm{SBF}$) that for* $a_j < x_j < b_j$

$$\begin{aligned}&\sqrt{nh_j}[\hat{m}_j^{l,\mathrm{iter}}(x_j) - m_j(x_j) - h_j^2\beta_j(x_j)]\\ &\quad\to N\left(0, \frac{\alpha(1-\alpha)}{f_{\varepsilon,X_j}(0,x_j)^2} f_{X_j}(x_j)\int K^2(u)\,du\right)\end{aligned}$$

in distribution, where $\beta_j(x_j) = \beta_j^*(x_j) - \int \beta_j^*(u_j)w_j(u_j)\,du_j$, $\beta_j^*(x_j) = h_j^{-2} \times m_j'(x_j)\int(u_j - x_j)K_{j,h_j}(x_j,u_j)\,du_j + \mu_{2,K}\frac{1}{2}m_j''(x_j) + \mu_{2,K}\beta_j^{**}(x_j)$, $\mu_{2,K} = \int v^2K(v)\,dv$ *and* $(\beta_1^{**},\dots,\beta_d^{**})$ *is a tuple of functions that minimizes*

$$\int\left[\sum_{j=1}^d\left(m_j'(x_j)\frac{\partial f_{\varepsilon,X}(0,x)/\partial x_j}{f_{\varepsilon,X}(0,x)} - \beta_j^{**}(x_j)\right)\right]^2 f_{\varepsilon,X}(0,x)\,dx.$$

Note that the first term in the definition of β_j^* is of order $n^{1/5}$ at the boundary but vanishes in the interior of I_j. Because of the norming with the weight function w_j, the bias function β_j is shifted from β_j^* by $\int \beta_j^*(u_j) w_j(u_j)\, du_j$. One can estimate the bias and the variance terms because they only require two-dimensional objects if one calculates them with the backfitting algorithms.

We now come back to discussion of the assumptions (A5)–(A7). Assumption (A5) allows that the starting estimators have a suboptimal rate. In particular, it requires that the starting estimators are consistent. For example, one could use here orthogonal series estimators, smoothing splines or sieve estimators. In the simulations, we got good results by using constant functions as starting values, that is, functions that are not consistent. For backfitting mean regression, it is known that every starting value works. Because of the nonlinearity of quantile regression, we do not expect that such a result can be proved for quantile regression. In our result, we did not specify the required rate for the pilot estimator. But, if one does this, we conjecture that one can get the statement of Theorem 2.2 with pilot estimators that have much slower rates. For such a theorem, one has to prove a modification of Proposition 2.1 with the following statement: for the estimators at the preceding stage of the backfitting algorithms, less accurate error bounds would suffice to get that the difference between the backfitting estimators $\hat{m}_1$ and $\hat{m}_1^*$ at the current stage of the algorithm is of higher order than the accuracy of the preceding estimators. This would allow one to weaken the assumptions on the rate of the starting estimators.

Assumption (A7) is not required for Theorem 2.2. This is because running the iterative algorithms (2.7) and (2.8) is only imaginary and in the proof we choose to use the same starting values as in the real iterative algorithms (2.2) and (2.3), respectively. Thus, (A7) is automatically satisfied at the beginning of the iterations. Proposition 2.1 tells us that the updated estimators also fulfill (A7). This holds with the same rate but with multiplicative factors. For this reason, after L backfitting cycles the difference between the mean regression and the quantile regression estimators is not of order $(C \times L)n^{-2/5-\delta}$, but of order $C^L n^{-2/5-\delta}$, for some $\delta > 0$, $C > 1$. For a number of iterations, $C_{\text{iter}} \log n$ such that $C_{\text{iter}} \log C < \delta$ this is of order $o(n^{-2/5})$.

Compared with the results for mean regression backfitting estimators, our results for quantile estimation are weaker in two aspects. First, we need initial estimators that are consistent, whereas in mean regression one can start with arbitrary initial values. This restriction comes from the nonlinearity of the quantile functional. Second, we put restrictions on the number of iteration steps. It must be of logarithmic order with a factor that is not too small and not too large. When letting run the two parallel backfitting procedures

for mean and quantile regression, we were not able to control in the proof the difference between the two outcomes if the number of iterations is too large. We conjecture that both restrictions are necessary only for technical reasons in our approach for the proof. In our simulation, we started with nonconsistent pilot estimators and we let the algorithms run until the outcomes were stabilized. According to our experience in the simulation, there seemed practically no advantage in limiting the number of iterations and there was also no problem when starting the algorithm with initial estimators that were far away from the corresponding underlying regression functions.

A natural extension of our results is to study local polynomial quantile estimators. This can be done along the lines of this paper by putting smoothness restrictions also on the higher order terms of the local polynomial fit. This can be done relatively easily for local polynomial smooth backfitting. For local polynomial ordinary backfitting, it would require also essentially new theoretical results for mean regression. We do not follow this line in this paper.

3. Numerical implementation. In practical implementations of the smooth backfitting method, one may approximate the integral at (2.3) by Monte Carlo integration. This can be done in several ways. In one version, one generates $(U_2^j, \dots, U_d^j)$ for $1 \le j \le M$ from a $(d-1)$-variate uniform distribution on $I_2 \times \cdots \times I_d$. Then an approximation of $\hat{m}_1^{\mathrm{SBF}}(x_1)$ may be obtained by

$$\hat{m}_1^{\mathrm{SBF}}(x_1) \approx \operatorname*{arg\,min}_{\theta \in \Theta} \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_\alpha(Y_i - \theta - \hat{m}_0^{\mathrm{SBF}} - \hat{m}_2^{\mathrm{SBF}}(U_2^j) - \cdots - \hat{m}_d^{\mathrm{SBF}}(U_d^j)) \times K_{1,h_1}(x_1, X_1^i) K_{2,h_2}(U_2^j, X_2^i) \cdots K_{d,h_d}(U_d^j, X_d^i).$$

In practical implementation, the values U_k^j can be chosen from a finite grid of equidistant points. Then the algorithm has to update the function values of the additive components on this grid.

In another version, one generates independent $U_{\ell,i,j}$ for $\ell = 2, \dots, d$, $i = 1, \dots, n$, $j = 1, \dots, J$, where $U_{\ell,i,j}$ has density $K_{\ell,h_l}(\cdot, X_\ell^i)$. Again, in practical implementation, the values of these random variables can be chosen from a finite grid of equidistant points. Then the smooth backfitting estimator at x_1 is calculated by

$$\hat{m}_1^{\mathrm{SBF}}(x_1) \approx \operatorname*{arg\,min}_{\theta \in \Theta} \sum_{i=1}^{n} \sum_{j=1}^{J} \tau_\alpha(Y_i - \theta - \hat{m}_0^{\mathrm{SBF}} - \hat{m}_2^{\mathrm{SBF}}(U_{2,i,j}) - \cdots - \hat{m}_d^{\mathrm{SBF}}(U_{d,i,j})) K_{1,h_1}(x_1, X_1^i).$$

This means that the smooth backfitting estimator can be calculated by an algorithm that is designed for the ordinary backfitting with sample $(Y_i, X_1^i,$

$U_{2,i,j}, \ldots, U_{d,i,j})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. In this case, the speed of the algorithm for the smooth backfitting behaves as that for the ordinary backfitting with sample size Jn.

In the last algorithm, the values $U_{\ell,i,j}$ could be replaced by deterministic choices such that for fixed i and ℓ the probability density $K_{\ell,h_\ell}(\cdot, X_\ell^i)$ put equal mass between neighbored points of $U_{\ell,i,j}$, that is,

$$\int_{-\infty}^{U_{\ell,i,j}} K_{\ell,h_\ell}(x_\ell, X_\ell^i)\, dx_\ell = j/(J+1), \qquad j = 1, \ldots, J.$$

Suppose that $K_{\ell,h_\ell}(\cdot, z)$ is symmetric about z. Then the algorithm calculates the ordinary backfitting estimates when $J = 1$, since in that case $U_{\ell,i,1} = X_\ell^i$. It also approximates the smooth backfitting estimates as $J \to \infty$. Thus, there exists a broad band of compromises between the ordinary backfitting and the smooth backfitting for intermediate choices of J.

4. Simulation study. In this section, we illustrate the asymptotic equivalence asserted in Proposition 2.1. We compared the numerical properties of the ordinary backfitting (BF) and the smooth backfitting (SBF) estimators defined at (2.2) and (2.3) with their theoretical mean regression versions defined at (2.7) and (2.8), respectively.

In the simulation, we considered the following model:

$$Y^i = f_1(X_1^i) + f_2(X_2^i) + f_3(X_3^i) + \{\sigma_1(X_1^i) + \sigma_2(X_2^i) + \sigma_3(X_3^i)\}U^i,$$

where U^i are i.i.d. $N(0,1)$, $f_1(x_1) = x_1^3$, $f_2(x_2) = \sin(\pi x_2)$, $f_3(x_3) = 2 \times \exp(-16x_3^2)$, $\sigma_1(x_1) = \cos(x_1)$, $\sigma_2(x_2) = \exp(x_2)$ and $\sigma_3(x_3) = \exp(x_3)$. With this model, the centered version of the jth additive component of the α-quantile function equals

$$m_j(x_j; \alpha) = c_j + f_j(x_j) + \sigma_j(x_j)\Phi^{-1}(\alpha),$$

where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution and c_j is the constant that makes $Em_j(X_j^1; \alpha) = 0$. We considered two different cases for the distribution of X^i. One was the case where the components of X^i were independent. In this case, X^i were generated from $N_3(0, J)$ truncated outside $[-1, 1]^3$, where J denotes the identity matrix of dimension $d = 3$. This means the density of X^i was $f_X(x) = \varphi(x)I(x \in [-1, 1]^3)/\int_{[-1,1]^3} \varphi(z)\, dz$, where φ denotes the density function of $N_3(0, J)$. The second was the case where the components of X^i were correlated. In this case, $X^i \sim N_3(0, V)$ truncated outside $[-1, 1]^3$, where $V \equiv (v_{ij})$ has $v_{ii} = 1$ and $v_{ij} = 0.9$ for $i \neq j$. Because of the truncation, the actual correlation equals 0.644. The sample sizes were $n = 200$ and $n = 500$. These relatively large sample sizes were considered to let the asymptotic results in Section 2 be well in effect.

TABLE 1
Mean integrated squared errors of the estimators

Sample size	Distribution of X	Method	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
$n = 200$	Uncorrel.	BF	0.09345	0.07457	0.08770
		BF*	0.09585	0.07512	0.08208
		SBF	0.08818	0.07039	0.08209
		SBF*	0.09436	0.07455	0.07937
	Correl.	BF	0.09043	0.07165	0.08382
		BF*	0.09864	0.07539	0.08276
		SBF	0.08555	0.06712	0.07937
		SBF*	0.09136	0.07140	0.08412
$n = 500$	Uncorrel.	BF	0.05240	0.04020	0.04881
		BF*	0.04959	0.04121	0.04729
		SBF	0.04905	0.03827	0.04557
		SBF*	0.05045	0.04178	0.04896
	Correl.	BF	0.05463	0.04182	0.05094
		BF*	0.05137	0.04305	0.05312
		SBF	0.05186	0.03983	0.04743
		SBF*	0.05496	0.04221	0.05296

Note: BF* denotes the theoretical mean regression ordinary backfitting estimator, and SBF* denotes the theoretical mean regression smooth backfitting estimator.

Implementation of the ordinary and smooth backfitting methods requires optimization involving the nonsmooth function τ_α. For this, we used R function `rq()` in the library `quantreg`. For the smooth backfitting, we discretized the integrals on a fine grid in $[-1,1]^3$. We used

$$K_{j,g}(x,u) = \left[\int K\left(\frac{x-u}{g}\right)dx\right]^{-1} K\left(\frac{x-u}{g}\right), \tag{4.1}$$

where K is Epanechinikov kernel given by $K(u) = (3/4)(1-u^2)I_{[-1,1]}(u)$. For the bandwidths, we took $h_1 = h_2 = h_3 = h$ for simplicity. Normalization was done in each iteration so that $\int \hat{m}_j(x_j)\hat{f}_{X_j}(x_j)\,dx_j = 0$. Note that we used estimates of f_{X_j} in the normalization, instead of fixed weight functions which we considered in our theoretical development for simplicity. Using a different weight function changes the estimator only by an additive constant. To get the density estimates $\hat{f}_{X_j}$, we used the same kernel K and the bandwidth h that we employed for quantile estimation. We chose the initial estimates in the iterative algorithms (2.2), (2.3), (2.7) and (2.8) to be zero. It was found that the algorithms converged with this initial choice in all cases.

Table 1 show Monte Carlo estimates, based on 200 pseudo-samples, of the mean integrated squared errors,

$$\mathrm{MISE} = E\int\{\bar{m}_1(x_1)+\bar{m}_2(x_2)+\bar{m}_3(x_3)-m_1(x_1)-m_2(x_2)-m_3(x_3)\}^2 f_X(x)\,dx,$$

where f_X is the density function of X^i, and $\bar{m}_j$ represents $\hat{m}_j^{\mathrm{BF}}$, $\hat{m}_j^{\mathrm{SBF}}$, $\hat{m}_j^{*,\mathrm{BF}}$ or $\hat{m}_j^{*,\mathrm{SBF}}$. For each estimator, its MISE was estimated by $\overline{\mathrm{ISE}} = \sum_{r=1}^{200}\mathrm{ISE}_r/200$, where ISE_r is the value of the integrated squared error

$$\int\{\bar{m}_1(x_1)+\bar{m}_2(x_2)+\bar{m}_3(x_3)-m_1(x_1)-m_2(x_2)-m_3(x_3)\}^2 f_X(x)\,dx$$

for the rth sample. We computed the estimates of the additive regression function with bandwidths on a grid in $[0.1, 1.5]$. The values for $\hat{m}^{\mathrm{BF}}$ and $\hat{m}^{*,\mathrm{BF}}$ reported in the table are for the bandwidths that gave optimal performance of $\hat{m}^{\mathrm{BF}}$, and likewise those for $\hat{m}^{\mathrm{SBF}}$ and $\hat{m}^{*,\mathrm{SBF}}$ are for the bandwidths that gave optimal performance of $\hat{m}^{\mathrm{SBF}}$. In most cases, the estimated MISE was minimized around $h = 0.5$ when $n = 200$, and around $h = 0.4$ when $n = 500$. This is roughly consistent with the theory that the size of the optimal bandwidth equals $n^{-1/5}$ for univariate smoothing, according to which the ratio of the optimal bandwidths for $n = 200$ and $n = 500$ equals $(500/200)^{1/5} \approx 1.20$.

To compare $\hat{m}^{\mathrm{BF}}$ and $\hat{m}^{\mathrm{SBF}}$ with their theoretical mean regression counterparts $\hat{m}^{*,\mathrm{BF}}$ and $\hat{m}^{*,\mathrm{SBF}}$, we find that the two corresponding MISE values are very close, and that in most cases the differences get smaller as n increases. This supports our theory presented in Section 2. In the table, we also find that the size of the estimated MISE for $n = 500$ is nearly half of the corresponding value for $n = 200$. This supports the fact that the ordinary and smooth backfitting estimators enjoy the univariate rate of convergence $n^{-4/5}$ in MISE, since $(500/200)^{4/5} \approx 2.08$.

According to Table 1, the MISE values of the estimators at $\alpha = 0.5$ are always smaller than those at $\alpha = 0.2$ and $\alpha = 0.8$. Note that, in Theorem 2.2, $f_{X_j}^w(x_j)$ is nothing else than the joint density of (ε, X_j) at the point $(0, x_j)$. Under our simulation model, the conditional density can be expressed as

$$f_{X_j}^w(x_j) = \int \frac{1}{\sigma_1(x_1)+\sigma_2(x_2)+\sigma_3(x_3)}\phi\biggl(\frac{\Phi^{-1}(\alpha)}{\sigma_1(x_1)+\sigma_2(x_2)+\sigma_3(x_3)}\biggr) \times f_X(x)\,dx_{-j}$$

for $j = 1, 2$ and 3, where ϕ denotes the density of the standard normal distribution. According to Theorem 2.2, this implies that the theoretical value of the integrated variance increases as α gets away from 0.5. This explains why we have larger MISE values for α away from 0.5. Similar numerical

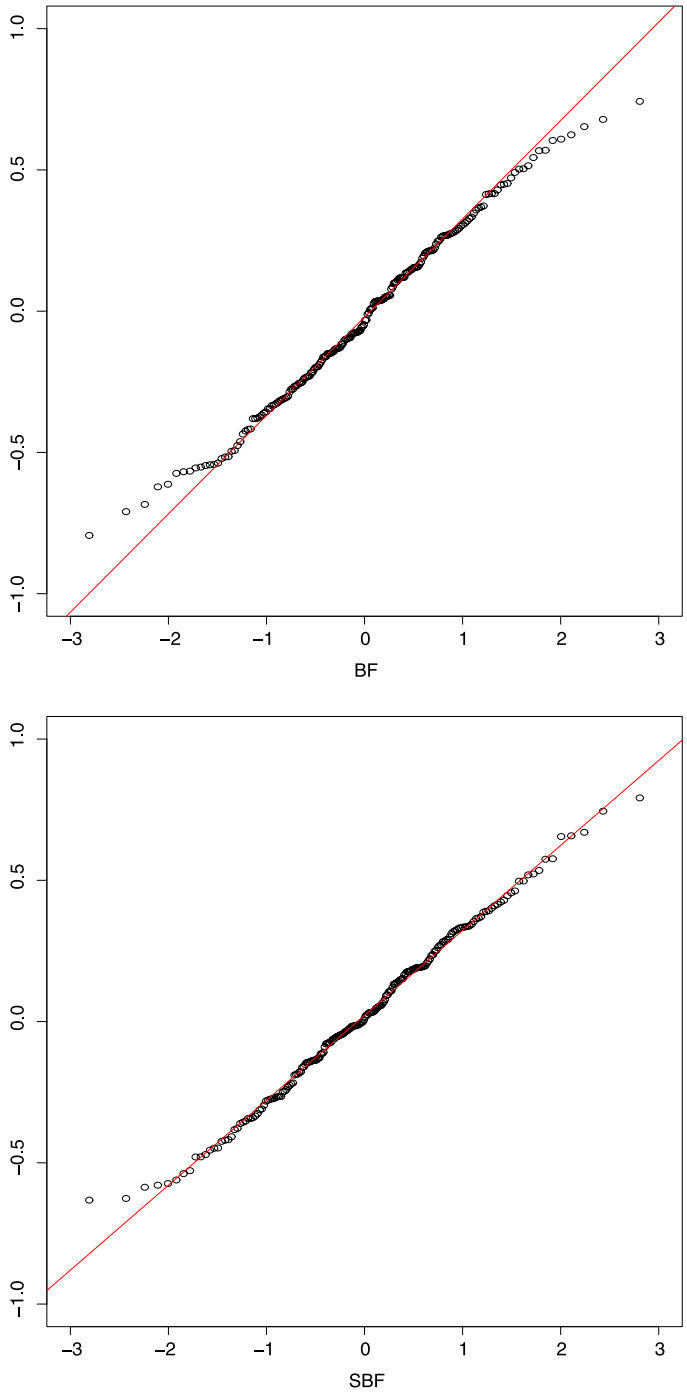

FIG. 1. *Normal Q–Q plots for* $\hat{m}_2^{\mathrm{BF}}(x)$ *and* $\hat{m}_2^{\mathrm{SBF}}(x)$ *based on 200 values computed from pseudo-samples in the case where* $x = 0$, $\alpha = 0.5$, $n = 200$ *and the components of* X^i *were correlated. The theoretical quantiles are on the horizontal axis and the sample quantiles are on the vertical axis.*

evidences were also observed by Yu and Jones (1998) and Lee, Lee and Park (2006).

Figure 1 illustrates the asymptotic normality of $\hat{m}_j^{\mathrm{BF}}$ and $\hat{m}_j^{\mathrm{SBF}}$. It depicts the normal Q–Q plots of the 200 values of $\hat{m}_2^{\mathrm{BF}}(x)$ and $\hat{m}_2^{\mathrm{SBF}}(x)$ at $x = 0$ when $\alpha = 0.5$ and $n = 200$. The figure is for the case where the components of X^i are correlated. Although it exhibits slight departures from normality at tails, the figure suggests that the distributions of the estimators get close to normal even for moderate sample sizes. We obtained other Q–Q plots that corresponded to other components j, other points x or other quantile levels α, and also repeated them in other simulation models. They looked not much different from the case we report here.

Figure 2 illustrates how the four curve estimates $\hat{m}_j^{\mathrm{BF}}$, $\hat{m}_j^{*,\mathrm{BF}}$, $\hat{m}_j^{\mathrm{SBF}}$ and $\hat{m}_j^{*,\mathrm{SBF}}$ computed from a single typical sample look like. In the top two panels, the long-dashed and dotted curves, respectively, represent $\hat{m}_j^{\mathrm{BF}}$ and $\hat{m}_j^{*,\mathrm{BF}}$ computed from a sample for which the value of the integrated squared error

$$\int \{\hat{m}_j^{\mathrm{BF}}(x_j) - m_j(x_j)\}^2 \, dx_j$$

was the median of those values obtained from the 200 pseudo-samples. Similarly, the bottom two panels depict $\hat{m}_j^{\mathrm{SBF}}$ and $\hat{m}_j^{*,\mathrm{SBF}}$ computed from a sample that gave the median performance in terms of the integrated squared error

$$\int \{\hat{m}_j^{\mathrm{SBF}}(x_j) - m_j(x_j)\}^2 \, dx_j.$$

In the figure the solid curves represent the true functions. In comparison of the pairs, m_j^{BF} versus $\hat{m}_j^{*,\mathrm{BF}}$ and $\hat{m}_j^{\mathrm{SBF}}$ versus $\hat{m}_j^{*,\mathrm{SBF}}$, we find that the two corresponding curves move together relatively closer than with the true function, although there are some places where they are more distant in the case of the backfitting estimator for $\alpha = 0.2$ (top left panel). The figure is for the estimates of the second component function when $n = 500$ and the components of X^i were correlated. Those for other cases gave similar lesson, so that are not included here.

One may be also interested in comparing the two backfitting quantile estimators $\hat{m}^{\mathrm{BF}}$ and $\hat{m}^{\mathrm{SBF}}$ in terms of MISE. For this, we computed the standard errors of the differences between the estimated values of MISE of the respective estimators. In Table 2, we provide the average differences $\overline{\mathrm{DIFF}}$ and their standard errors calculated by the formula

$$\mathrm{S.E.} = \sqrt{\sum_{r=1}^{200} (\mathrm{DIFF}_r - \overline{\mathrm{DIFF}})^2 / (199 \times 200)},$$

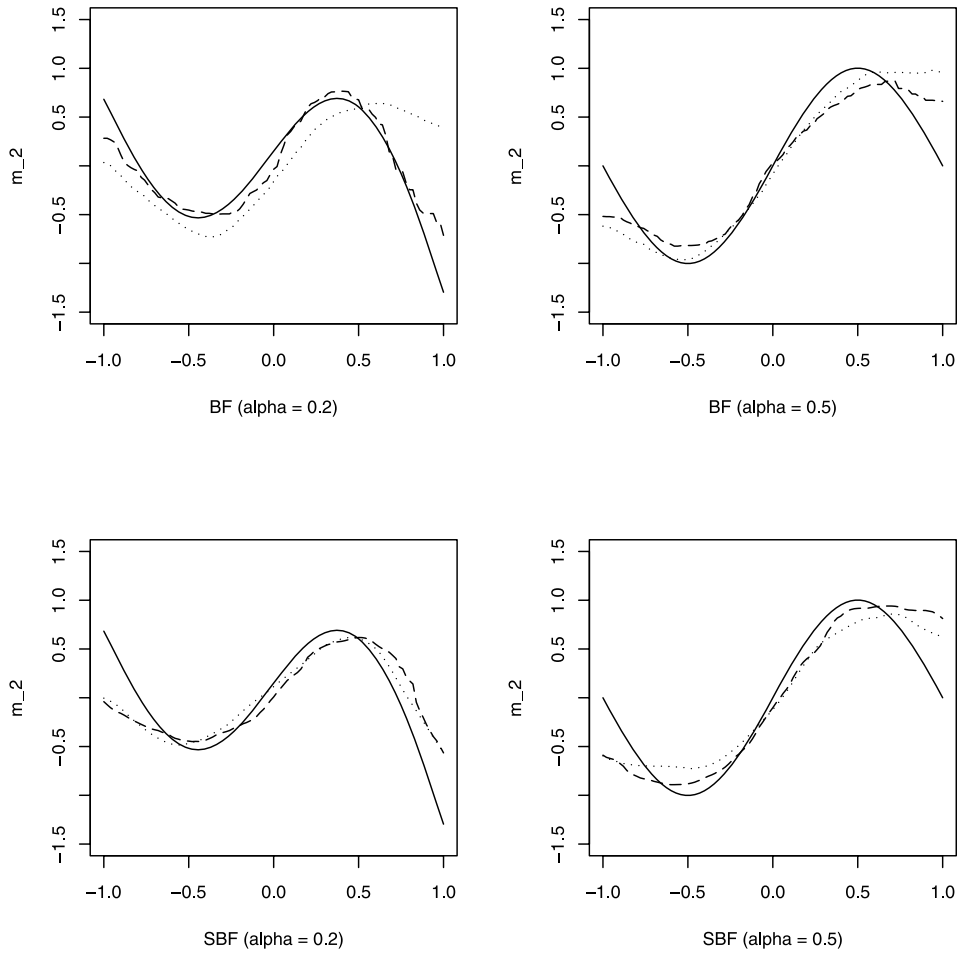

FIG. 2. *Estimates of a component function computed from a sample that gave the median performance in terms of the integrated squared error of* $\hat{m}_j^{\mathrm{BF}}$ *or* $\hat{m}_j^{\mathrm{SBF}}$*, when* $n = 500$ *and the covariates were correlated. Long-dashed and dotted curves in the top two panels are* $\hat{m}_j^{\mathrm{BF}}$ *and* $\hat{m}_j^{*,\mathrm{BF}}$*, respectively, and those in the bottom two panels are* $\hat{m}_j^{\mathrm{SBF}}$ *and* $\hat{m}_j^{*,\mathrm{SBF}}$*. Left two panels are for the case* $\alpha = 0.2$ *and the right are for* $\alpha = 0.5$*. Solid curves represent the true component functions.*

where $\overline{\mathrm{DIFF}}$ denotes the average of DIFF_r over 200 pseudo-samples, and

$$\mathrm{DIFF}_r = (\text{ISE of } \hat{m}^{\mathrm{BF}} \text{ for the } r\text{th sample}) - (\text{ISE of } \hat{m}^{\mathrm{SBF}} \text{ for the } r\text{th sample}).$$

Comparing the two backfitting quantile estimators, we find that the smooth backfitting estimators have smaller values of the estimated MISE in all cases than the ordinary backfitting estimators. In particular, all the differences are statistically significant, exceeding two standard errors. Although not reported in the paper, we also compared the two backfitting quantile estima-

TABLE 2
Differences in mean integrated squared errors of BF and SBF estimators

Sample size	Distribution of X	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
$n = 200$	Uncorrel.	0.00527	0.00418	0.00561
		(0.00099)	(0.00063)	(0.00087)
	Correl.	0.00488	0.00453	0.00445
		(0.00098)	(0.00068)	(0.00096)
$n = 500$	Uncorrel.	0.00335	0.00193	0.00324
		(0.00045)	(0.00028)	(0.00038)
	Correl.	0.00277	0.00199	0.00351
		(0.00042)	(0.00034)	(0.00042)

Note: the numbers are averages of (ISE of $\hat{m}^{\mathrm{BF}}$) − (ISE of $\hat{m}^{\mathrm{SBF}}$) over 200 pseudo-samples, and their standard errors are given in the parentheses.

tors with their oracle versions. An oracle estimator of an additive component is the one obtained by using true functions for the other components. We found that in all cases the two backfitting quantile estimators had similar performance as their oracle versions.

5. Proofs.

5.1. *Proof of Proposition 2.1.* We only give the proof for the ordinary backfitting etimator. The proof will be given for $a_1 + C_S n^{-1/5} \le x_1 \le b_1 - C_S n^{-1/5}$. The proofs for the smooth backfitting estimator and for boundary points follow by similar arguments. For simplicity of notation, we also assume that $d = 2$.

The basic asymptotic argument for a treatment of parametric and nonparametric quantile estimators is a Bahadur expansion. It states that the quantile estimator is asymptotically equivalent to a linear statistic, that is, to a sum of independent variables. This expansion would directly carry over to our case if the pilot functions (input) of the backfitting algorithms would be nonrandom. Because this is not the case, we have to generalize the Bahadur approach. We have to show that the Bahadur expansion holds uniformly over a class of pilot functions. Furthermore, we have to verify that the pilot estimators lie in this function class with probability tending to one. The latter is guaranteed by the assumptions (A5) and (A6). The uniform expansion is the main step of our proof.

Define

$$V_i(\theta, \mu_2, x_1)$$
$$= K_{1,h_1}(x_1, X_1^i)[\tau_\alpha(Y^i - \theta - \mu_2(X_2^i)) - \tau_\alpha(\varepsilon^i + m_1(X_1^i) - m_1(x_1))$$

$$- (\theta - m_1(x_1) + \mu_2(X_2^i) - m_2(X_2^i)) \times (I(\varepsilon^i + m_1(X_1^i) - m_1(x_1) < 0) - \alpha)].$$

Let $J_1 \equiv J_1(x_1)$ and $J_2 \equiv J_2(x_1)$ be index sets defined by

$$J_1 = \{i : |X_1^i - x_1| \le Ch_1, a_2 + C_S n^{-1/5} \le X_2^i \le b_2 - C_S n^{-1/5}\},$$
$$J_2 = \{i : |X_1^i - x_1| \le Ch_1, a_2 \le X_2^i < a_2 + C_S n^{-1/5} \text{ or } b_2 - C_S n^{-1/5} < X_2^i \le b_2\}.$$

Put

$$\begin{aligned} &D(\theta, \mu_2, x_1) \\ &\quad = \sum_{i=1}^{n} [V_i(\theta, \mu_2, x_1) - E^{\mathcal{X}} V_i(\theta, \mu_2, x_1)] \\ &\quad = \sum_{i \in J_1} [V_i(\theta, \mu_2, x_1) - E^{\mathcal{X}} V_i(\theta, \mu_2, x_1)] \\ &\qquad + \sum_{i \in J_2} [V_i(\theta, \mu_2, x_1) - E^{\mathcal{X}} V_i(\theta, \mu_2, x_1)] \\ &\quad \equiv D_1(\theta, \mu_2, x_1) + D_2(\theta, \mu_2, x_1), \end{aligned}$$

where $E^{\mathcal{X}}$ is the conditional expectation given $\mathcal{X} = \{X^1, \ldots, X^n\}$. Let M_1 and M_2 denote the numbers of elements of J_1 and J_2, respectively. These are random variables. Since h_1 is of order $n^{-1/5}$ and the density f_X is strictly positive on its support, M_1 is of order $n \times n^{-1/5} = n^{4/5}$ and M_2 is of order $n \times n^{-1/5} \times n^{-1/5} = n^{3/5}$. Thus, there exist constants $C_1 > 0$ and $C_2 > 0$ such that $C_1 n^{4/5} \le M_1 \le 2C_1 n^{4/5}$ and $C_2 n^{3/5} \le M_2 \le 2C_2 n^{3/5}$ with probability tending to one.

For a fixed constant $D > 0$, we now introduce the class $\mathcal{M}_n$ of all tuples of a parameter $\theta \in \Theta$ and a function g that fulfills

$$\sup_{a_2 + C_S n^{-1/5} \le x_2 \le b_2 - C_S n^{-1/5}} |g(x_2) - m_2(x_2)| \le D n^{-(1+\rho)/(2+3\rho)4/5 - \Delta_1}$$

and whose derivative fulfills a Lipschitz condition of order ρ with Lipschitz constant C as in (A6).

For $j \ge 0$, let $\mathcal{M}_n(2^{-j})$ denote a grid of points in $\mathcal{M}_n$ such that for every $(\theta, g) \in \mathcal{M}_n$ there exists $(\theta^*, g^*) \in \mathcal{M}_n(2^{-j})$ with $|\theta^* - \theta| \le 2^{-j}$ and $\|g^* - g\|_\infty \le 2^{-j}$. Let N_j denote the number of points in the grid $\mathcal{M}_n(2^{-j})$. Note that $N_j = O\{\exp(2^{j/(1+\rho)} n^{\xi/(1+\rho)})\}$.

We apply the Bernstein inequality. For a sum of r independent random variables V_i that are absolutely bounded by a constant κ and have finite

variance bounded by σ^2, this inequality states that

$$P\left(\left|r^{-1/2}\sum_{i=1}^{r}(V_i - EV_i)\right| \geq a\right) \leq 2\exp\left(-\frac{a^2}{2a\kappa r^{-1/2} + 2\sigma^2}\right)$$
$$\leq 2\exp\left(-\frac{a}{4\kappa r^{-1/2}}\right) + 2\exp\left(-\frac{a^2}{4\sigma^2}\right).$$

We apply this inequality with a chaining argument for $D_1(\theta, \mu, x_1)$ and $D_2(\theta, \mu, x_1)$. In doing this, we take $r = M_1$ (or $r = M_2$, resp.) and $P = P^{\mathcal{X}}$ where $P^{\mathcal{X}}$ is the conditional distribution given $\mathcal{X} = \{X^1, \ldots, X^n\}$. Let J_n be chosen so that $2^{-J_n} \leq n^{-2/5-\delta} \leq 2^{-J_n+1}$ with $\delta > 0$ small enough, see below. Define $\gamma = 4(1+\rho)/[5(2+3\rho)]$ and $I_n = \{j : j \leq J_n, Dn^{-\gamma-\Delta_1} \geq 2^{-j}\}$. Furthermore, for $(\theta, \mu) \in \mathcal{M}_n(2^{-J_n})$ choose $(\theta^j, \mu^j) \in \mathcal{M}_n(2^{-j})$ with $|\theta^j - \theta| \leq 2^{-j}$ and $\|\mu^j - \mu\|_\infty \leq 2^{-j}$. For $j = J_n$, we choose $(\theta^j, \mu^j) = (\theta, \mu)$. We do not indicate the dependence of (θ^j, μ^j) on (θ, μ) in the notation. For $j \leq j_n = \min I_n$, the grid $\mathcal{M}_n(2^{-j})$ can be chosen so that it contains only one value of μ. We assume that this value is equal to $\mu^0 = m_2$. Furthermore, we choose $\theta^0 = m_1(x_1)$ and we assume w.l.o.g. that the diameter of Θ is less than one. For $j = 0$, the grid $\mathcal{M}_n(2^{-j})$ contains only one value which we choose to be (θ^0, μ^0). Then

$$P\Big(\sup_{(\theta,\mu)\in\mathcal{M}_n(2^{-J_n})} |D_1(\theta, \mu, x_1)| > n^{-4/5-2\delta}|\mathcal{X}\Big)$$
$$\leq P\Big(\sup_{(\theta,\mu)\in\mathcal{M}_n(2^{-J_n})} \Big| D_1(\theta^0, \mu^0, x_1)$$
$$+ \sum_{1\leq j<j_n} D_1(\theta^j, \mu^0, x_1) - D_1(\theta^{j-1}, \mu^0, x_1)$$
$$+ \sum_{j_n\leq j\leq J_n} D_1(\theta^j, \mu^j, x_1) - D_1(\theta^{j-1}, \mu^{j-1}, x_1)\Big|$$
$$> n^{-4/5-2\delta}|\mathcal{X}\Big).$$

Let s_j be positive numbers (depending on n) such that $\sum_{1\leq j\leq J_n} s_j \leq 1/2$. Then the right-hand side of the above inequality is bounded by

$$P(|D_1(\theta^0, \mu^0, x_1)| > 2^{-1}n^{-4/5-2\delta}|\mathcal{X})$$
$$+ \sum_{1\leq j<j_n} 2^{2j} \sup_{*} P(|D_1(\theta^j, \mu^0, x_1)$$
$$- D_1(\theta^{j-1}, \mu^0, x_1)| > s_j n^{-4/5-2\delta}|\mathcal{X}) \tag{5.1}$$

$$+\sum_{j_n\le j\le J_n} N_j N_{j-1} \sup_{**} P(|D_1(\theta^j,\mu^j,x_1) - D_1(\theta^{j-1},\mu^{j-1},x_1)| > s_j n^{-4/5-2\delta}|\mathcal{X}),$$

where $\sup_*$ and $\sup_{**}$ runs over all $(\theta^j,\mu^j)\in\mathcal{M}_n(2^{-j})$ and $(\theta^{j-1},\mu^{j-1})\in\mathcal{M}_n(2^{-j+1})$ with $|\theta^j-\theta^{j-1}|\le 2^{-j+1}$ and $\|\mu^j-\mu^{j-1}\|_\infty\le 2^{-j+1}$.

Using the Bernstein inequality with $\kappa = O(2^{-j}h_1^{-1})$, $\sigma^2 = 2^{-2j}O(n^{-\gamma-\Delta_1}\times h_1^{-2})$ and $a = M_1^{-1/2}ns_jn^{-4/5-2\delta}c$, the last sum in (5.1) can be bounded by

$$\sum_{j_n\le j\le J_n} [\exp(d_1 2^{j/(1+\rho)}n^{\xi/(1+\rho)} - d_2 s_j n n^{-4/5-2\delta}2^j h_1) + \exp(d_1 2^{j/(1+\rho)}n^{\xi/(1+\rho)} - d_2 s_j^2 M_1^{-1}n^2 n^{-8/5-4\delta}2^{2j}n^{\gamma+\Delta_1}h_1^2)] \tag{5.2}$$

for some constants $d_1,d_2>0$. Choosing $s_j = (d_3\log n)^{-1}$ with d_3 large enough, the sum at (5.2) can be bounded further by

$$\exp(-d_4 n^{d_5}) + \exp(-d_6 M_1^{-1}n^{4/5+d_7}),$$

where $d_4,\dots,d_7>0$ are some constants. Here, we used that $\delta>0$ is small enough. Using similar arguments for the first two terms in (5.1), one can bound the sum of all three terms in (5.1) by

$$\exp(-d_8 n^{d_9}),$$

where $d_8,d_9>0$ are some constants. This exponential bound entails that for $\delta>0$ small enough

$$\begin{aligned}&\sup_{\substack{(\theta,\mu_2)\in\mathcal{M}_n(2^{-J_n})\\ x_1\in I_1}} |n^{-1}D_1(\theta,\mu_2,x_1)|\\ &= \sup_{\substack{(\theta,\mu_2)\in\mathcal{M}_n(2^{-J_n})\\ x_1\in I_1}} \Big|n^{-1}\sum_{i\in J_1}\{V_i(\theta,\mu_2,x_1) - E^{\mathcal{X}}V_i(\theta,\mu_2,x_1)\}\Big| \qquad (5.3)\\ &= O_P(n^{-4/5-\delta}).\end{aligned}$$

Similarly, it can be shown that

$$\begin{aligned}&\sup_{\substack{(\theta,\mu_2)\in\mathcal{M}_n(2^{-J_n})\\ x_1\in I_1}} |n^{-1}D_2(\theta,\mu_2,x_1)|\\ &= \sup_{\substack{(\theta,\mu_2)\in\mathcal{M}_n(2^{-J_n})\\ x_1\in I_1}} \Big|n^{-1}\sum_{i\in J_2}\{V_i(\theta,\mu_2,x_1) - E^{\mathcal{X}}V_i(\theta,\mu_2,x_1)\}\Big| \qquad (5.4)\\ &= O_P(n^{-4/5-\delta}).\end{aligned}$$

We now use a Taylor expansion of $E^{\mathcal{X}}V_i(\theta,\mu_2,x_1)$ with respect to θ. Note that with $A^i=\varepsilon^i+m_1(X_1^i)-m_1(x_1)$ and $B^i=Y^i-\theta-\mu_2(X_2^i)=\varepsilon^i+m_1(X_1^i)-\theta+m_2(X_2^i)-\mu_2(X_2^i)$

$$V_i(\theta,\mu_2,x_1)=K_{1,h_1}(x_1,X_1^i)\begin{cases}0, & \text{if } A^i,B^i<0,\\ 0, & \text{if } A^i,B^i\geq 0,\\ B^i, & \text{if } A^i<0\leq B^i,\\ -B^i, & \text{if } A^i\geq 0>B^i.\end{cases}$$

For $\delta_1,\delta_2>0$ small enough, we get that uniformly for $|\theta-m_1(x_1)|\leq\delta_1$

$$\begin{aligned}&E^{\mathcal{X}}V_i(\theta,\mu_2,x_1)\\&\quad=\tfrac{1}{2}K_{1,h_1}(x_1,X_1^i)f_{\varepsilon|X}(0|X^i)\{[m_2(X_2^i)-\mu_2(X_2^i)-\theta+m_1(x_1)]^2\\&\qquad\qquad+O_P(n^{-4/5-\delta_2})+O_P(|\theta-m_1(x_1)|^3)\},\end{aligned}$$

see (A3). We now apply (5.3), (5.4) and the fact that the change of an empirical quantile cannot be larger than the largest change of an observation. We use these results to analyze the update $\hat{m}_1^{\mathrm{BF}}(x_1)$ when we plug into the iteration formula (2.2) of the backfitting estimator a choice of $\mu_2=\hat{m}_2^{\mathrm{BF}}$ that lies in $\mathcal{M}_n$. By a direct argument, it can be shown that with probability tending to one the resulting value lies in an δ_1-neighborhood of $m_1(x_1)$. Thus, using the above expansions, we get that, up to terms of order $O_P(n^{-2/5-\delta_3})$ with $\delta_3>0$ small enough, the resulting value for the update $\hat{m}_1^{\mathrm{BF}}(x_1)$ is equal to the minimum of

$$\begin{aligned}&\frac{\theta}{n}\sum_{i=1}^n K_{1,h_1}(x_1,X_1^i)[I(\varepsilon^i+m_1(X_1^i)-m_1(x_1)\leq 0)-\alpha]\\&\quad+\frac{1}{2n}\sum_{i=1}^n K_{1,h_1}(x_1,X_1^i)f_{\varepsilon|X}(0|X^i)[m_2(X_2^i)-\mu_2(X_2^i)-\theta+m_1(x_1)]^2.\end{aligned}$$

The minimum of this expression is equal to

$$\begin{aligned}&m_1(x_1)-\hat{f}_{X_j}^w(x_j)^{-1}\frac{1}{n}\sum_{i=1}^n K_{1,h_1}(x_1,X_1^i)[I(\varepsilon^i+m_1(X_1^i)-m_1(x_1)\leq 0)-\alpha]\\&\quad+\hat{f}_{X_j}^w(x_j)^{-1}\frac{1}{n}\sum_{i=1}^n K_{1,h_1}(x_1,X_1^i)f_{\varepsilon|X}(0|X^i)[m_2(X_2^i)-\mu_2(X_2^i)],\end{aligned}$$

where $\hat{f}_{X_j}^w(x_j)$ has been defined after (2.8). We now use that

$$\hat{f}_{X_j}^w(x_j)^{-1}\frac{1}{n}\sum_{i=1}^n K_{1,h_1}(x_1,X_1^i)[I(\varepsilon^i+m_1(X_1^i)-m_1(x_1)\leq 0)-I(\varepsilon^i\leq 0)]$$

$$= m_1(x_1) - \hat{f}_{X_j}^{w}(x_j)^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{1,h_1}(x_1, X_1^i) f_{\varepsilon|X}(0|X^i) m_1(X_1^i)$$
$$+ O_P(n^{-2/5-\delta})$$

for $\delta > 0$ small enough. This shows that the minimum is equal to

$$\hat{f}_{X_j}^{w}(x_j)^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{1,h_1}(x_1, X_1^i) f_{\varepsilon|X}(0|X^i)[m_1(X_1^i) + m_2(X_2^i) + \eta^i]$$
$$- \hat{f}_{X_j}^{w}(x_j)^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{1,h_1}(x_1, X_1^i) f_{\varepsilon|X}(0|X^i)\mu_2(X_2^i) + O_P(n^{-2/5-\delta}).$$

This expansion holds uniformly for $x_1 \in I_1$ and $\mu_2 \in \mathcal{M}_n$.

To complete the proof, we use the fact that, if one replaces in (2.2) or (2.7) the input function $\mu_2 = \hat{m}_2^{\mathrm{BF}}$ or $\mu_2 = \hat{m}_2^{*,\mathrm{BF}}$, respectively, by another function that differs in sup-norm by an amount of order $O_P(n^{-2/5-\Delta_2})$, then the resulting estimator changes also at most by an amount of order $O_P(n^{-2/5-\Delta_2})$. In particular, if $\delta < \Delta_2$, this implies that

$$\sup_{a_1 + C_S n^{-1/5} \le x_1 \le b_1 - C_S n^{-1/5}} |\hat{m}_1^{\mathrm{BF}}(x_1) - \hat{m}_1^{*,\mathrm{BF}}(x_1)| = O_P(n^{-2/5-\delta}).$$

The other statements of Proposition 2.1 can be proved by using similar arguments.

5.2. *Proof of Theorem 2.2.* We will prove the theorem for the ordinary backfitting estimator. A proof for the smooth backfitting estimator follows along the same lines. We only give an outline of the proof. For simplicity, we assume that the condition (A6) holds with $\rho = 1$. Our basic argument runs as follows. We choose $\hat{m}_j^{*,\mathrm{BF},[0]} = \hat{m}_j^{\mathrm{BF},[0]}$. By assumption, these starting values fulfill (A5) and (A6) (with the choice $\hat{m}_j^{\mathrm{BF}} = \hat{m}_j^{*,\mathrm{BF},[0]} = \hat{m}_j^{\mathrm{BF},[0]}$). Thus, we can apply Proposition 2.1 and we get that the updates $\hat{m}_j^{*,\mathrm{BF},[1]}$ and $\hat{m}_j^{\mathrm{BF},[1]}$ fulfill (A7) (with the choices $\hat{m}_j^{*,\mathrm{BF}} = \hat{m}_j^{*,\mathrm{BF},[1]}$ and $\hat{m}_j^{\mathrm{BF}} = \hat{m}_j^{\mathrm{BF},[1]}$). We will show below that the updates $\hat{m}_j^{*,\mathrm{BF},[l]}$ of the mean regression backfitting estimator fulfill conditions (A5) and (A6) for all $l \ge 1$. With this fact, we can use an iterative argument. Suppose that we know that (A5)–(A7) hold for $\hat{m}_j^{*,\mathrm{BF},[l-1]}$ and $\hat{m}_j^{\mathrm{BF},[l-1]}$. Then with our proof below we get that $\hat{m}_j^{*,\mathrm{BF},[l]}$ fulfills (A5) and (A6). By application of Proposition 2.1, we get that (A7) holds for $\hat{m}_j^{*,\mathrm{BF},[l]}$ and $\hat{m}_j^{\mathrm{BF},[l]}$. Thus, $\hat{m}_j^{\mathrm{BF},[l]}$ lies in a neighborhood of $\hat{m}_j^{*,\mathrm{BF},[l]}$ and (A5) and (A6) also hold for $\hat{m}_j^{\mathrm{BF},[l]}$ because they are satisfied by $\hat{m}_j^{*,\mathrm{BF},[l]}$.

The bound for the distance between $\hat{m}_j^{*,\mathrm{BF},[l]}$ and $\hat{m}_j^{\mathrm{BF},[l]}$ adds up. Each application of Proposition 2.1 adds an additional term. The additional term increases with l. With a careful analysis of the arguments in the proof of Proposition 2.1, one gets that the bounds in (A5) and (A6) have to be multiplied by a factor C_*^l with a constant $C_* > 1$. If $l \le C_{\mathrm{iter}} \log n$ with $C_{\mathrm{iter}} > 0$ small enough, we get

$$\hat{m}_j^{\mathrm{BF},[l]} - \hat{m}_j^{*,\mathrm{BF},[l]} = o_P(n^{-2/5}). \tag{5.5}$$

In the second part of the proof, we will show the asymptotic normality of $\hat{m}_j^{*,\mathrm{BF},[C\log n]}$ for C large enough. The minimal sufficient value of C for this result depends on the rate of convergence of $\hat{m}_j^{*,\mathrm{BF},[0]}$ to m_j. If this rate is $n^{-2/5}$, then it can be made as small as one likes. For slower rates, one needs larger values of C. If the rate is fast enough, one can choose $C < C_{\mathrm{iter}}$. In this case, we can apply (5.5) and we get the same asymptotic normality result for $\hat{m}_j^{*,\mathrm{BF},[C_{\mathrm{iter}}\log n]}$. This will conclude the proof of Theorem 2.2.

We now prove that the updates $\hat{m}_j^{*,\mathrm{BF},[l]}$ fulfill the conditions (A5) and (A6) for all $l \ge 1$. For this purpose, we rewrite (2.7) as

$$\begin{aligned}
&\hat{m}_j^{*,\mathrm{BF},[l]}(x_j) - m_j(x_j) \\
&\quad = \tilde{m}_j^{*,A}(x_j) + \tilde{m}_j^{*,B}(x_j) + \tilde{m}_j^{*,C,[l]}(x_j) - \hat{m}_0^{*,\mathrm{BF}} \\
&\qquad - \sum_{k=1,\neq j}^{d} \int [\hat{m}_k^{*,\mathrm{BF},[l_{k,j}]}(x_k) - m_k(x_k)] f_{X_k|X_j}^{n,w}(x_k|x_j)\,dx_k,
\end{aligned} \tag{5.6}$$

where $l_{k,j} = l+1$ for $k < j$, $l_{k,j} = l$ for $k > j$, and

$$\tilde{m}_j^{*,A}(x_j) = \frac{n^{-1}\sum_{i=1}^n f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i)\eta^i}{\hat{f}_{X_j}^w(x_j)},$$

$$\tilde{m}_j^{*,B}(x_j) = \frac{n^{-1}\sum_{i=1}^n f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i)[m_j(X_j^i) - m_j(x_j)]}{\hat{f}_{X_j}^w(x_j)},$$

$$\begin{aligned}
\tilde{m}_j^{*,C,[l]}(x_j) = &- \sum_{k=1,\neq j}^{d} \Bigg(n^{-1} \sum_{i=1}^n f_{\varepsilon|X}(0|X^i) K_{j,h_j}(x_j, X_j^i) \\
&\qquad \times [\hat{m}_k^{*,\mathrm{BF},[l_{k,j}]}(X_k^i) - m_k(X_k^i)] \Bigg) (\hat{f}_{X_j}^w(x_j))^{-1} \\
&+ \sum_{k=1,\neq j}^{d} \int [\hat{m}_k^{*,\mathrm{BF},[l_{k,j}]}(x_k) - m_k(x_k)] f_{X_k|X_j}^{n,w}(x_k|x_j)\,dx_k,
\end{aligned}$$

$$f^{n,w}_{X_k|X_j}(u_k|x_j) = \frac{\int f_{\varepsilon|X}(0|u)K_{j,h_j}(x_j,u_j)f_X(u)\,du_{-k}}{\int f_{\varepsilon|X}(0|v)K_{j,h_j}(x_j,v_j)f_X(v)\,dv}.$$

The iteration (5.6) can be analyzed as the smooth backfitting algorithm in Mammen, Linton and Nielsen (1999). With $\hat{m}^{*,\mathrm{BF},[l]}_+(x) = \hat{m}^{*,\mathrm{BF},[l]}_1(x_1) + \cdots + \hat{m}^{*,\mathrm{BF},[l]}_d(x_d)$ and $m_+(x) = m_1(x_1) + \cdots + m_d(x_d)$, we can write a full cycle of iterations (5.6) as

$$\begin{aligned}\hat{m}^{*,\mathrm{BF},[l+1]}_+ - m_+ &= \tilde{m}^{*,A}_\oplus + \tilde{m}^{*,B}_\oplus + \tilde{m}^{*,C,[l]}_\oplus - \hat{m}^{*,\mathrm{BF}}_0 \\ &\quad + T_{n,+}(\hat{m}^{*,\mathrm{BF},[l]}_+ - m_+ - \mu_l) + \mu_l,\end{aligned} \tag{5.7}$$

where $\tilde{m}^{*,A}_\oplus$, $\tilde{m}^{*,B}_\oplus$ and $\tilde{m}^{*,C,[l]}_\oplus$ are some functions, $T_{n,+}$ is an operator that acts on additive mean zero functions in $L_2(f_{\varepsilon|X}(0|\cdot)f_X(\cdot))$, and $\mu_l = \int(\hat{m}^{*,\mathrm{BF},[l]}_+ - m_+)(x)f_X(x)f_{\varepsilon|X}(0|x)\,dx$. We used $\oplus$ (not $+$) as subindex in $\tilde{m}^{*,A}_\oplus$ because it is not the sum of $\tilde{m}^{*,A}_j$. The operator $T_{n,+}$ converges to an operator T_+ that is based on an iterative application of the linear transformations for the additive components g_j of an additive function g_+:

$$g_j \to -\sum_{k=1,\neq j}^{d} \int g_k(x_k) f^w_{X_k|X_j}(x_k|x_j)\,dx_k.$$

More precisely, the kernel function of $T_{n,+}$ converges to the kernel function of T_+, with respect to the sup-norm.

Arguing as in the proof of Lemma 1 in Mammen, Linton and Nielsen (1999), one can show that T_+ is a positive self-adjoint operator with operator norm strictly less than one, $\|T_+\| < 1$, and with $\|T_j m\|_\infty \le D\|m\|_2$ for a constant $D > 0$. Here, $T_j m$ is the jth additive component of $T_+ m$. This gives with constants $0 < D' < 1$ and $D'' > 0$ for n large enough

$$\|T_{n,+}\| < D'. \tag{5.8}$$

Furthermore, we have

$$\|T_{n,j}m\|_\infty \le D''\|m\|_2, \tag{5.9}$$

where $T_{n,j}m$ is the jth additive component of $T_{n,+}m$. Iterative application of (5.7) gives

$$\hat{m}^{*,\mathrm{BF},[l]}_+ - m_+ = \hat{m}^{*,A,[l]}_+ + \hat{m}^{*,B,[l]}_+ + \hat{m}^{*,C,[l]}_+ - \hat{m}^{*,\mathrm{BF}}_0 + \bar{T}^l_{n,+}(\hat{m}^{*,\mathrm{BF},[0]}_+ - m_+),$$

where $\bar{T}_{n,+}$ is an extension of $T_{n,+}$ to a nonzero mean function by putting $\bar{T}_{n,+}g = T_{n,+}(g - \mu_g) + \mu_g$ with $\mu_g = \int g(x)f_X(x)f_{\varepsilon|X}(0|x)\,dx$, and

$$\hat{m}^{*,A,[l]}_+ = \sum_{r=0}^{l-1} \bar{T}^r_{n,+}\tilde{m}^{*,A}_\oplus,$$

$$\hat{m}_+^{*,B,[l]} = \sum_{r=0}^{l-1} \bar{T}_{n,+}^r \tilde{m}_\oplus^{*,B},$$

$$\hat{m}_+^{*,C,[l]} = \sum_{r=0}^{l-1} \bar{T}_{n,+}^{l-r-1} \tilde{m}_\oplus^{*,C,[r]}.$$

Using standard bounds on $\tilde{m}_j^{*,A}$ and $\tilde{m}_j^{*,B}$, it can be verified that

$$\sup_{x_j \in I_j, l \geq 1} |\hat{m}_j^{*,A,[l]}(x_j)| = O_P(n^{-2/5}), \tag{5.10}$$

$$\sup_{x_j \in I_j, l \geq 1} |\hat{m}_j^{*,B,[l]}(x_j)| = O_P(n^{-1/5}), \tag{5.11}$$

$$\sup_{a_j + C_S h_j \leq x_j \leq b_j - C_S h_j, l \geq 1} |\hat{m}_j^{*,B,[l]}(x_j)| = O_P(n^{-2/5}), \tag{5.12}$$

where for an additive function g_+ we denote by g_j its jth additive component.

We now argue that for a constant $C_T > 0$

$$\sup_{x_j \in I_j, l \geq 1} |\bar{T}_{n,j} \bar{T}_{n,+}^{l-1} (\hat{m}_j^{*,\mathrm{BF},[0]} - m_j)(x_j)| \leq C_T \kappa_n, \tag{5.13}$$

where

$$\kappa_n = \sup_{1 \leq j \leq d} \Big[\sup_{a_j + C_S h_j \leq x_j \leq b_j - C_S h_j} |\hat{m}_j^{*,\mathrm{BF},[0]} - m_j|(x_j) + n^{-1/5} \sup_{a_j \leq x_j \leq b_j} |\hat{m}_j^{*,\mathrm{BF},[0]} - m_j|(x_j) \Big].$$

For a proof of this claim, one applies (5.8) and (5.9). Also, we argue that

$$\sup_{x_j \in I, l \geq 1} |\hat{m}_j^{C,[l]}(x_j)| = o_P(n^{-2/5}). \tag{5.14}$$

For a proof of (5.14), we note that

$$\sup_{x_j \in I_j, l \geq 1} |\tilde{m}_j^{C,[l]}(x_j)| = o_P(n^{-2/5}).$$

This follows by empirical process theory. One uses the fact that $\hat{m}_k^{*,\mathrm{BF},[l-1]} - m_k$ lies in a class of functions that have second derivatives absolutely bounded by $C_\xi n^\xi$ with $\xi > 0$ being arbitrarily small and constant C_ξ depending on ξ. This can be shown by using that the same bound applies for $\tilde{m}_j^A$ and $\tilde{m}_j^B$, and that the kernels of the operators T_+ and T_j have an absolutely bounded second derivative [see (A9)], and then applying an iterative argument.

The bounds at (5.10)–(5.14) imply that $\hat{m}_j^{*,\mathrm{BF},[l]}$ fulfills (A5) uniformly for $l \geq 1$. Using the smoothness considerations in the previous paragraph, we get that $\hat{m}_j^{*,\mathrm{BF},[l]}$ fulfills (A6) uniformly for $l \geq 1$. Thus, we get by an iterative application of Proposition 2.1 that (5.5) holds.

It remains to show the asymptotic normality result for $\hat{m}_j^{\mathrm{BF,iter}} = \hat{m}_j^{\mathrm{BF},[C_{\mathrm{iter}} \log n]}$ with C_{iter} large enough. Using the above arguments, we have for C_{iter} large enough that

$$\hat{m}_j^{*,\mathrm{BF,iter}}(x_j) - m_j(x_j) = \hat{m}_j^{A,[C_{\mathrm{iter}} \log n]}(x_j) + \hat{m}_j^{B,[C_{\mathrm{iter}} \log n]}(x_j) + o_P(n^{-2/5}).$$

We argue that

$$\sup_{l \geq 1} |\hat{m}_j^{A,[l]}(x_j) - \tilde{m}_j^A(x_j)| = o_P(n^{-2/5}), \tag{5.15}$$

$$h_j^{-2} \hat{m}_j^{B,[l]}(x_j) \to \beta_j(x_j) \qquad \text{as } l \to \infty. \tag{5.16}$$

These two claims imply that

$$\hat{m}_j^{*,\mathrm{BF}}(x_j) - m_j(x_j) = \tilde{m}_j^A(x_j) + h_j^2 \beta_j(x_j) + o_P(n^{-2/5}).$$

This expansion shows the desired asymptotic limit result by using a standard smoothing limit result for $\tilde{m}_j^A(x_j)$.

We prove (5.15) and (5.16). Claim (5.15) follows from standard smoothing theory as in Mammen, Linton and Nielsen (1999). For a proof of (5.16), we define $\beta_j^{[l]}(x_j) = \beta_j^*(x_j) - \sum_{k=1,\neq j}^d \int \beta_k^{[l_{k,j}]}(x_k) f_{X_k|X_j}^w(x_k|x_j)\,dx_k$ with $\beta_j^{[0]}(x_j) \equiv 0$. Similarly, as in (5.7), we can write a full cycle of these iterations as

$$\beta_+^{[l+1]} = \beta_\oplus^* + \bar{T}_+ \beta_+^{[l]}, \tag{5.17}$$

where $\beta_\oplus^*$ is some additive function, $\beta_+^{[l]}(x)$ is equal to $\beta_1^{[l]}(x_1) + \cdots + \beta_d^{[l]}(x_d)$ and $\bar{T}_+$ is an extension of T_+ defined by $\bar{T}_+ g = T_+(g - \mu_g) + \mu_g$ with μ_g defined as above. Note that we get $\beta_+^{[l]} = \sum_{r=0}^{l-1} \bar{T}_+^r \beta_\oplus^*$. This expansion shows that

$$\begin{aligned} \sup_{x_j \in I_j, l \geq 1} |\hat{m}_j^{*,B,[l]}(x_j) - h_j^2 \beta_j^{[l]}| &= o_P(n^{-1/5}), \\ \sup_{a_j + C_S h_j \leq x_j \leq b_j - C_S h_j, l \geq 1} |\hat{m}_j^{*,B,[l]}(x_j) - h_j^2 \beta_j^{[l]}| &= o_P(n^{-2/5}). \end{aligned} \tag{5.18}$$

Furthermore, we get that the term $\beta_+^{[l]} - \sum_{j=1}^d [h_j^{-2} m_j'(x_j) \int (u_j - x_j) K_{j,h_j}(x_j, u_j)\,du_j - \mu_{2,K} \frac{1}{2} m_j''(x_j)]$ converges to $\mu_{2,K} \beta_+^{**}$ as $l \to \infty$, where $(\beta_1^{**}, \ldots, \beta_d^{**})$ is the minimizer of

$$\int \left[\sum_{j=1}^d \left(m_j'(x_j) \frac{\partial / \partial x_j f_{\varepsilon,X}(0,x)}{f_{\varepsilon,X}(0,x)} - \beta_j^{**}(x_j) \right) \right]^2 f_{\varepsilon,X}(0,x)\,dx.$$

This follows because the updating (5.17) is given by the first-order conditions of this minimization problem. Together with (5.18), this implies (5.16).

REFERENCES

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580. MR0189095

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555. MR0994249

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their Bahadur representation. *Ann. Statist.* **19** 760–777. MR1105843

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587

Furno, M. (2004). ARCH tests and quantile regressions. *J. Stat. Comput. Simul.* **74** 277–292. MR2059314

Horowitz, J. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *J. Amer. Statist. Assoc.* **100** 1238–1249. MR2236438

Horowitz, J. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32** 2412–2443. MR2153990

Jones, M. C. and Hall, P. (1990). Mean squared error properties of kernel estimates of regression quantiles. *Statist. Probab. Lett.* **10** 283–289. MR1069903

Lee, Y. K., Lee, E. R. and Park, B. U. (2006). Conditional quantile estimation by local logistic regression. *J. Nonparametr. Stat.* **18** 357–373. MR2284188

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–101. MR1332841

Lu, Z. and Yu, K. (2004). Local linear additive quantile regression. *Scand. J. Statist.* **31** 333–346. MR2087829

Mammen, E., Linton, O. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

Mammen, E. and Park, B. U. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.* **34** 2252–2271. MR2291499

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.* **73** 166–179. MR1763322

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211. MR1429922

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93** 228–237. MR1614628

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas. *The Statistician* **52** 331–350. MR2011179

Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

Y. K. Lee
Department of Statistics
Kangwon National University
Chuncheon 200-701
Korea
E-mail: youngklee@kangwon.ac.kr

E. Mammen
Department of Economics
University of Mannheim
68131 Mannheim, L7, 3-5
Germany
E-mail: emammen@rumms.uni-mannheim.de

B. U. Park
Department of Statistics
Seoul National University
Seoul 151-747
Korea
E-mail: bupark@stats.snu.ac.kr