The Annals of Statistics
2012, Vol. 40, No. 3, 1906–1933
DOI: 10.1214/12-AOS1026

FLEXIBLE GENERALIZED VARYING COEFFICIENT REGRESSION MODELS

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

Kangwon National University, Universität Mannheim and Seoul National University

This paper studies a very flexible model that can be used widely to analyze the relation between a response and multiple covariates. The model is nonparametric, yet renders easy interpretation for the effects of the covariates. The model accommodates both continuous and discrete random variables for the response and covariates. It is quite flexible to cover the generalized varying coefficient models and the generalized additive models as special cases. Under a weak condition we give a general theorem that the problem of estimating the multivariate mean function is equivalent to that of estimating its univariate component functions. We discuss implications of the theorem for sieve and penalized least squares estimators, and then investigate the outcomes in full details for a kernel-type estimator. The kernel estimator is given as a solution of a system of nonlinear integral equations. We provide an iterative algorithm to solve the system of equations and discuss the theoretical properties of the estimator and the algorithm. Finally, we give simulation results.

1. Introduction. The varying coefficient regression model, proposed by Hastie and Tibshirani (1993), and studied by Yang et al. (2006), Roca-Pardinas and Sperlich (2010) and Lee, Mammen and Park (2012), is known to be a useful tool for analyzing the relation between a response and a multivariate covariate. For a response Y and covariates $\mathbf{X}$ and $\mathbf{Z}$, they assumed that the mean regression function, $m(\mathbf{x},\mathbf{z}) \equiv E(Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})$, takes the

Received January 2012; revised June 2012.

[1] Supported by the Basic Science Research Program through the NRF funded by the Korea government (MEST) (No. 2010-0021396).

[2] Supported by the NRF grant funded by the Korea government (MEST) (No. 2010-0017437).

AMS 2000 subject classifications. Primary 62G08; secondary 62G20.

Key words and phrases. Varying coefficient models, kernel smoothing, entropy, projection, Hilbert space, quasi-likelihood, integral equation, Newton–Raphson approximation.

form $m(\mathbf{x},\mathbf{z}) = x_1 f_1(z_1) + \cdots + x_d f_d(z_d)$ for some unknown univariate functions f_j. The model is simple in structure, gives easy interpretation, and yet is flexible since the dependence of the response variable on the covariates is modeled in a nonparametric way. The major hurdle in the practical application of this model is that one needs to pair up each "X-covariate" with only one "Z-covariate." Typically this is not the case in practice. In principle, each X-covariate may interact with any number of Z-covariates to explain the variation in Y. Moreover, it is often difficult to differentiate X-types from Z-types in a group of the covariates.

In this paper we are concerned with quite a more flexible setting than the usual varying coefficient model. Suppose that we are given a group of covariates, $X_1,\ldots,X_D$. Let $\mathbf{X} = (X_1,\ldots,X_D)^\top$ and d be an integer such that $d \le D$. The model we are interested in assumes that there is a link function, say g, such that the mean function $m(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ satisfies

$$g(m(\mathbf{x})) = x_1\left(\sum_{k\in I_1} f_{1k}(x_k)\right) + \cdots + x_d\left(\sum_{k\in I_d} f_{dk}(x_k)\right), \tag{1.1}$$

where the index sets $I_j \subset \{1,2,\ldots,D\}$ are known, and each I_j does not include j. The covariates that enter into one of the coefficient functions f_{jk}, that is, X_k for $k \in \mathcal{C} \equiv \bigcup_{j=1}^d I_j$ are of continuous type. For simplicity, we assume that X_k with $k \in \mathcal{C}$ are supported on the interval $[0,1]$. We allow some of the covariates X_j, for $1 \le j \le d$, to be discrete random variables. We also allow that $\mathcal{C}$ and $\{1,2,\ldots,d\}$ may have common indices. Let $\mathcal{C}_0 = \mathcal{C} \cap \{1,2,\ldots,d\}$. The case $\mathcal{C}_0 = \varnothing$, that is, $\mathcal{C} = \{d+1,\ldots,D\}$, corresponds to the situation where one can distinguish between two groups of covariates, $\{X_1,\ldots,X_d\}$ and $\{Z_1,\ldots,Z_p\} \equiv \{X_{d+1},\ldots,X_{d+p}\}$ with $D = d + p$. In this case, the model reduces to

$$g(m(\mathbf{x},\mathbf{z})) = x_1\left(\sum_{k\in I_1} f_{1k}(z_k)\right) + \cdots + x_d\left(\sum_{k\in I_d} f_{dk}(z_k)\right), \tag{1.2}$$

where $I_j \subset \{1,2,\ldots,p\}$ are index sets of Z_k. The latter model arises, for example, when one takes all X_j, for $1 \le j \le d$, to be discrete covariates. The above model reduces further to the nonparametric generalized additive model of Yu, Park and Mammen (2008) if we take $d = 1$ and $X_1 \equiv 1$.

The functions f_{jk} in the representation (1.1) are not identifiable. To see this, consider the case where $d = D = 3, I_1 = \{2,3\}, I_2 = \{3\}, I_3 = \{2\}$ so that $\mathcal{C} = \mathcal{C}_0 = \{2,3\}$. In this case, $x_1[f_{12}(x_2) + f_{13}(x_3)] + x_2 f_{23}(x_3) + x_3 f_{32}(x_2) = x_1[g_{12}(x_2) + g_{13}(x_3)] + x_2 g_{23}(x_3) + x_3 g_{32}(x_2)$, if $g_{12}(x_2) = f_{12}(x_2) + c$, $g_{13}(x_3) = f_{13}(x_2) - c$, $g_{23}(x_3) = f_{23}(x_3) + x_3$ and $g_{32}(x_2) = f_{32}(x_2) - x_2$ for some constant c. To make all f_{jk} identifiable, we use the following constraints:

$$\int f_{jk}(x_k) w_k(x_k)\, dx_k = 0, \qquad k \in \mathcal{C}, 1 \le j \le d,$$

$$\int x_k f_{jk}(x_k) w_k(x_k)\, dx_k = 0, \qquad j, k \in \mathcal{C}_0 \tag{1.3}$$

for nonnegative weight functions w_k, where $\mathcal{C}_0 = \{1, 2, \ldots, d\} \cap \mathcal{C}$. With these constraints we may rewrite model (1.1) as

$$g(m(\mathbf{x})) = \sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k + \sum_{j=1}^{d} x_j \left(\sum_{k \in I_j} f_{jk}(x_k) \right). \tag{1.4}$$

We think that our approach broadens the field of applications of varying coefficient models essentially. The link function allows us to have a discrete response. Furthermore, our model frees us from the restrictive settings of the usual varying coefficient model that one should differentiate between two types of covariates, X- and Z-type as in (1.2), and that each covariate appears in only one "nonlinear interaction term." In the case where $X_1, \ldots, X_q$, for some $q \le D - 2$, are discrete and the remaining covariates are continuous random variables, our approach allows fitting the full model, where $I_j = \{q + 1, \ldots, D\}$ for all $1 \le j \le q$ and $I_j = \{q + 1, \ldots, j - 1, j + 1, \ldots, D\}$ for $q + 1 \le j \le D$:

$$\begin{aligned} g(m(\mathbf{x})) &= x_1 \left(\sum_{k=q+1}^{D} f_{1k}(x_k) \right) + \cdots + x_q \left(\sum_{k=q+1}^{D} f_{pk}(x_k) \right) \\ &\quad + x_{q+1} \left(\sum_{k=q+2}^{D} f_{q+1,k}(x_k) \right) + \cdots + x_D \left(\sum_{k=q+1}^{D-1} f_{Dk}(x_k) \right). \end{aligned} \tag{1.5}$$

One may fit the full model in an exploratory analysis of the data, and find a parsimonious model, deleting some of the functions f_{jk} in the full model, that fits the data well.

We stress that fitting model (1.1) is not more complex than fitting other varying coefficient models such as $g(m(\mathbf{x})) = x_1 f_1(x_{d+1}) + \cdots + x_d f_d(x_{2d})$. The complexity is only in notation and theory, yet it gives full flexibility in modeling via varying coefficients. Think of the case where the true model is $g(m(\mathbf{x})) = x_1 f_{12}(x_2) + x_2 f_{21}(x_1)$. In this case, $g(m(\mathbf{x}))$ may not be well approximated by either $x_1 f_{12}(x_2)$ or $x_2 f_{21}(x_1)$ alone. Each additive term in $g(m(\mathbf{x}))$ is interpreted as a (linear) × (nonlinear) interaction. With x_2 being held fixed, modeling by $x_1 f_{12}(x_2)$ alone, for example, reflects only the linear effect of X_1, while modeling by $x_1 f_{12}(x_2) + x_2 f_{21}(x_1)$ accommodates the nonlinear effect of X_1 as well.

Xue and Yang (2006) discussed a special case of model (1.2) where one can differentiate between X-type and Z-type variables and all $I_j = \{1, 2, \ldots, p\}$.

They proved that the functions f_{jk} with the constraints (1.3) are identifiable. The essential assumption was that the smallest eigenvalue of $E(\mathbf{X}\mathbf{X}^\top|\mathbf{Z}=\mathbf{z})$ is bounded away from zero where $\mathbf{X}=(X_1,\ldots,X_d)^\top$ and $\mathbf{Z}=(Z_1,\ldots,Z_p)^\top$, although they put a stronger one; see their condition (C2). Their approach cannot be extended to our model (1.1). To see this, consider the model (1.5) and think of $E(\mathbf{X}\mathbf{X}^\top|\mathbf{X}^c=\mathbf{x}^c)$ where $\mathbf{X}=(X_1,\ldots,X_D)^\top$ and $\mathbf{X}^c=(X_{q+1},\ldots,X_D)^\top$. Certainly, the matrix is singular for all $\mathbf{x}^c$ if $D-q\geq 2$. One of our main tasks in this paper is to relax the requirement that the smallest eigenvalue of $E(\mathbf{X}\mathbf{X}^\top|\mathbf{X}^c=\mathbf{x}^c)$ is bounded away from zero; see assumption (A0) in Section 2. This weaker condition is typically satisfied. In Lemma 1 in the Appendix, we show that, under the weaker condition, the L_2-norms of m and of the function tuple $(f_{jk}:k\in I_j, 1\leq j\leq d)$ in the model (1.1) are equivalent, modulo the norming constraints (1.3). This has important implications in estimating the model (1.1). First, it implies that the functions f_{jk} are identifiable. For some types of estimators of m, it also gives directly the first-order properties of the corresponding estimators of f_{jk}. In the next section, we discuss the implications. In the subsequent sections, we focus on the smooth backfitting method of estimating the model (1.1), where we also use the main idea contained in Lemma 1 to derive its asymptotic properties. The original idea of smooth backfitting was introduced by Mammen, Linton and Nielsen (1999) for fitting additive models, and it has been further developed in other contexts; see Yu, Park and Mammen (2008) and Lee, Mammen and Park (2010, 2012), for example.

Earlier works on varying coefficient models focused on the model $m(\mathbf{x},\mathbf{z})=x_1f_1(\mathbf{z})+\cdots+x_df_d(\mathbf{z})$, where a single covariate $\mathbf{Z}$ (univariate or multivariate) enters into all coefficient functions. This model was proposed and studied by Chen and Tsay (1993), Kauermann and Tutz (1999), Fan and Zhang (1999, 2000), Cai, Fan and Li (2000), Cai, Fan and Yao (2000) and Fan, Yao and Cai (2003). Mammen and Nielsen (2003) added a link function to this model: $g(m(\mathbf{x},\mathbf{z}))=x_1f_1(\mathbf{z})+\cdots+x_df_d(\mathbf{z})$. The case where f_j are time-varying was treated by Hoover et al. (1998), Huang, Wu and Zhou (2002, 2004), Wang, Li and Huang (2008), and Noh and Park (2010). Heim et al. (2007) also considered the case where all coefficient functions are defined on a single 3D spatial space. Fitting these models is simple. A univariate or multivariate smoothing across the single variable $\mathbf{Z}$, or on a time scale, or on a multidimensional spatial space, gives directly estimators of f_j without further projection (by marginal integration or backfitting, e.g.) onto relevant function spaces. However, this suffers from the curse of dimensionality when the dimension of $\mathbf{Z}$ is high. For this reason, most works were focused on univariate Z. Some time series models related to the functional coefficient model, with X_j being unobserved common factors that depend on time, have been proposed and studied by, for example, Fengler, Härdle and Mammen (2007) and Park et al. (2009).

2. Equivalence in entropies of function classes. In this section we will show that the nonparametric components, $f_{jk}: k \in I_j, 1 \le j \le d$, of our model (1.4) with constraints (1.3) can be estimated with a one-dimensional nonparametric rate. This means that our model avoids the curse of dimensionality. It is easy to check that the function $m(\mathbf{x})$ can be estimated with a one-dimensional rate. This follows by application of results from empirical process theory; see below. We will use that the L_2-norms of m and of the tuples $(\boldsymbol{\alpha}; f_{jk}: 1 \le j \le d, k \in I_j)$ are equivalent; see Lemma 1 in the proofs section. Here, $\boldsymbol{\alpha}$ denotes the vector with elements $\alpha_j: 1 \le j \le d$, $\alpha_{jk}: j < k, j, k \in \mathcal{C}_0$. Our next result uses this fact to show that the rate for the estimation of m carries over to the estimation of $(\boldsymbol{\alpha}; f_{jk}: 1 \le j \le d, k \in I_j)$.

In the description of our method and in our theory we will also make use of a different representation of the model (1.4). In this representation of the model, we collect those coefficients that are functions of the same continuous covariate and put them together as an additive component. Suppose that, among $X_1, \ldots, X_d$ in model (1.4), there are r $(0 \le r \le d)$ variables whose indices do not enter into $\mathcal{C}$. Without loss of generality, we denote them by $X_1, \ldots, X_r$. Let $p = D - r \ge 2$ be the number of indices in $\mathcal{C}$. Thus $\mathcal{C} = \{r+1, \ldots, r+p\}$ and $\mathcal{C}_0 = \{r+1, \ldots, d\}$. Define

$$\tilde{\mathbf{X}}_k = \{X_j : r + k \in I_j, 1 \le j \le d\}, \qquad 1 \le k \le p. \tag{2.1}$$

The vector $\tilde{\mathbf{X}}_k$ is the collection of all X_j, for $1 \le j \le d$, that have interactions with X_{r+k} in the form of $X_j f_{j,r+k}(X_{r+k})$. Thus, $\tilde{\mathbf{X}}_k$ does not include X_{r+k} as its element. Let d_k denote the number of the index sets I_j that contain $r+k$. Thus, $\tilde{\mathbf{X}}_k$ is of d_k-dimension. Likewise, for a given vector $\mathbf{x}$, we denote the above rearrangements of $\mathbf{x}$ by $\tilde{\mathbf{x}}_k, 1 \le k \le p$. Also, define $\mathbf{f}_k = \{f_{j,r+k} : r + k \in I_j, 1 \le j \le d\}$ for $1 \le k \le p$. Then model (1.4) can be represented as

$$g(m(\mathbf{x})) = \sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k + \tilde{\mathbf{x}}_1^\top \mathbf{f}_1(x_{r+1}) + \cdots + \tilde{\mathbf{x}}_p^\top \mathbf{f}_p(x_{r+p}). \tag{2.2}$$

To give an example of the above representation, consider the case where $d = D = 3$, $I_1 = \{2, 3\}$, $I_2 = \{3\}$, $I_3 = \{2\}$ so that $\mathcal{C} = \mathcal{C}_0 = \{2, 3\}$. In this case, $r = 1$, $\tilde{\mathbf{x}}_1 = (x_1, x_3)^\top$, $\tilde{\mathbf{x}}_2 = (x_1, x_2)^\top$, $\mathbf{f}_1 = (f_{12}, f_{32})^\top$, $\mathbf{f}_2 = (f_{13}, f_{23})^\top$, and thus

$$x_1[f_{12}(x_2) + f_{13}(x_3)] + x_2 f_{23}(x_3) + x_3 f_{32}(x_2) = \tilde{\mathbf{x}}_1^\top \mathbf{f}_1(x_2) + \tilde{\mathbf{x}}_2^\top \mathbf{f}_2(x_3).$$

Suppose now that one has an estimator $\hat{m}$ of m with

$$g(\hat{m}(\mathbf{x})) = \sum_{j=1}^{d} \hat{\alpha}_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \hat{\alpha}_{jk} x_j x_k + \sum_{j=1}^{d} x_j \biggl(\sum_{k \in I_j} \hat{f}_{jk}(x_k) \biggr), \tag{2.3}$$

where the estimated functions $\hat{f}_{jk}$ satisfy the constraints in (1.3). We make the following assumption:

(A0) It holds that the product measure $\prod_{j=1}^{D} P_{X_j}$ has a density w.r.t. the distribution $P_{\mathbf{X}}$ of $\mathbf{X}$ that is bounded away from zero and infinity on the support of $P_{\mathbf{X}}$. Here, P_{X_j} is the marginal distribution of X_j. The marginal distributions are absolutely continuous w.r.t. Lebesgue measure or they are discrete measures with a finite support. Furthermore, the weight functions w_j for $j \in \mathcal{C}$ in the constraints in (1.3) are chosen so that w_j/p_{X_j} is bounded away from zero and infinity on the support of P_{X_j}. Here, p_{X_j} is the density of X_j. The smallest eigenvalues of the matrices $E[\tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^\top | X_{r+k} = z_k]$ for $1 \le k \le p$ are bounded away from zero for z_k in the support of $p_{X_{r+k}}$.

The condition on $E[\tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^\top | X_{r+k} = z_k]$ in (A0) is typically satisfied. For example, consider the model (1.5), where $X_1, \ldots, X_q$ are discrete random variables whose indices do not enter into $\mathcal{C}$. Thus, $r = q$ and $p = D - q$. According to configuration (2.1), we get $\tilde{\mathbf{X}}_k = (X_j : 1 \le j \le D, j \ne q+k)^\top$ which is the covariate vector $\mathbf{X}$ without X_{q+k}. In this case, $E[\tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^\top | X_{r+k} = z_k]$ is positive definite if the support of the conditional distribution $P_{\tilde{\mathbf{X}}_k | X_{q+k} = z_k}$ contains $(D-1)$ linearly independent vectors.

We get the following theorem for the rate of convergence of the components of $\hat{m}$.

Theorem 1. *Suppose that assumption* (A0) *applies, and that an estimator $\hat{m}$ of m with (2.3) and (1.3) satisfies that, for a null sequence κ_n,*

$$\int [g(\hat{m}(\mathbf{x})) - g(m(\mathbf{x}))]^2 P_{\mathbf{X}}(d\mathbf{x}) = O_p(\kappa_n^2). \tag{2.4}$$

Then it holds that

$$\int [\hat{f}_{jk}(x_k) - f_{jk}(x_k)]^2 p_{X_k}(x_k)\, dx_k = O_p(\kappa_n^2)$$

for all $k \in I_j, 1 \le j \le d$.

It is easy to construct estimators that fulfill (2.4). We will discuss this for the case where i.i.d. observations $(\mathbf{X}^i, Y^i)$ are made on the random vector $(\mathbf{X}, Y)$ with $\mathbf{X}^i \equiv (X_1^i, \ldots, X_D^i)^\top$ of D-dimension. Examples for estimators that fulfill (2.4) are sieve estimators or penalized least squares estimators. If one makes entropy conditions on some function classes $\mathcal{F}_{jk}$, then Theorem 1 can be used to show that the entropy conditions carry over to the class $\mathcal{M} = \{m : g(m(\mathbf{x}))$ has the structure (1.4) for some $\boldsymbol{\alpha} \in \mathcal{A}$ and $f_{jk} \in \mathcal{F}_{jk}, k \in I_j, 1 \le j \le d\}$ for some compact set $\mathcal{A}$. Using empirical process methods one can then show that sieve estimators or penalized least squares estimators fulfill (2.4). Below we outline this for the case where $\mathcal{F}_{jk}$ are the classes of l-times differentiable functions for some $l \ge 2$.

The penalized least squares estimator $(\hat{\boldsymbol{\alpha}}^{\mathrm{PEN}}; \hat{f}_{jk}^{\mathrm{PEN}} : 1 \le j \le d, k \in I_j)$ minimizes

$$n^{-1}\sum_{i=1}^{n}\Biggl\{Y_i - g^{-1}\Biggl(\sum_{j=1}^{d}\alpha_j X_j^i + \sum_{\substack{j<k\\ j,k\in\mathcal{C}_0}}\alpha_{jk}X_j^i X_k^i + \sum_{j=1}^{d} X_j^i \sum_{k\in I_j} f_{jk}(X_k^i)\Biggr)\Biggr\}^2 + \lambda_n^2 J(\mathbf{f}), \tag{2.5}$$

where we put $J(\mathbf{f}) = \sum_{k\in I_1}\int D_z^l f_{1k}(z)^2\,dz + \cdots + \sum_{k\in I_d}\int D_z^l f_{dk}(z)^2\,dz$. Here, the functions $\hat{f}_{jk}^{\mathrm{PEN}}$ are chosen so that (1.3) holds, and for a function g, $D_z^l g$ denotes its lth order derivative. We get the following result for the rate of convergence of the penalized least squares estimators $\hat{f}_{jk}^{\mathrm{PEN}}$.

Corollary 1. *Suppose that all assumptions of Theorem 1 hold, that the link function g has an absolutely bounded derivative, and that the estimators $\hat{f}_{jk}^{\mathrm{PEN}}$ of f_{jk} are defined by (2.5). Suppose that, for $k \in I_j, 1 \le j \le d$, the functions f_{jk} have square integrable derivatives of order l. Furthermore, assume that the (conditional) distribution of $\varepsilon_i = Y_i - m(\mathbf{X}^i), 1 \le i \le n$, has subexponential tails. That is, there are constants $t_\varepsilon, c_\varepsilon > 0$ such that*

$$\sup_{1\le i\le n} E[\exp(t|\varepsilon_i|)|\mathbf{X}^1, \ldots, \mathbf{X}^n] < c_\varepsilon$$

almost surely for $|t| \le t_\varepsilon$. Choose λ_n such that $\lambda_n^{-1} = O_p(n^{l/(2l+1)})$ and $\lambda_n = O_p(n^{-l/(2l+1)})$. Then it holds that

$$\int [\hat{f}_{jk}^{\mathrm{PEN}}(z) - f_{jk}(z)]^2 p_{X_k}(z)\,dz = O_p(n^{-2l/(2l+1)}),$$

$$\int D_z^l \hat{f}_{jk}^{\mathrm{PEN}}(z)^2\,dz = O_p(1)$$

for all $k \in I_j, 1 \le j \le d$.

We now discuss sieve estimation of m. We will do this for spline sieves. Denote by $\mathcal{G}_{n,c}$ the space of all spline functions of order l with knot points $0, L_n^{-1}, 2L_n^{-1}, \ldots, 1$ and with lth derivative absolutely bounded by c. The spline sieve estimator $(\hat{\boldsymbol{\alpha}}^{\mathrm{SIEVE}}; \hat{f}_{jk}^{\mathrm{SIEVE}} : 1 \le j \le d, k \in I_j)$ minimizes

$$n^{-1}\sum_{i=1}^{n}\Biggl\{Y_i - g^{-1}\Biggl(\sum_{j=1}^{d}\alpha_j X_j^i + \sum_{\substack{j<k\\ j,k\in\mathcal{C}_0}}\alpha_{jk}X_j^i X_k^i + \sum_{j=1}^{d} X_j^i \sum_{k\in I_j} f_{jk}(X_k^i)\Biggr)\Biggr\}^2 \tag{2.6}$$

over all functions f_{jk} in $\mathcal{G}_{n,c}$. Again, the functions $\hat{f}_{jk}^{\mathrm{SIEVE}}$ are chosen so that (1.3) holds. We get the following result for the rate of convergence of the sieve estimators $\hat{f}_{jk}^{\mathrm{SIEVE}}$.

COROLLARY 2. *Suppose that all assumptions of Theorem 1 hold, that the link function g has an absolutely bounded derivative and that the estimators $\hat{f}_{jk}^{\mathrm{SIEVE}}$ of f_{jk} are defined by (2.6). Suppose that, for $k \in I_j, 1 \le j \le d$, the functions f_{jk} have derivatives of order l that are absolutely bounded by c. Furthermore, assume that $E|\varepsilon|^\gamma < \infty$ holds for some $\gamma > 2 + l^{-1}$. Choose L_n such that $L_n^{-1} = O(n^{-1/(2l+1)})$ and $L_n = O(n^{1/(2l+1)})$. Then it holds that*

$$\int [\hat{f}_{jk}^{\mathrm{SIEVE}}(z) - f_{jk}(z)]^2 p_{X_k}(z)\, dz = O_p(n^{-2l/(2l+1)})$$

for all $k \in I_j, 1 \le j \le d$.

Both results, Corollaries 1 and 2, can be generalized to quasi-likelihood estimation. Then, the estimators are defined as in (2.5) or (2.6), respectively, but with the squared error $(y - g^{-1}(u))^2$ replaced by $Q(g^{-1}(u), y)$. For the definition of Q, see the next section. One can show that the results of Corollaries 1 and 2 still hold under the additional assumptions (A1) and (A2) of Mammen and van de Geer (1997). This can be proved along the lines of arguments of the latter paper. In the subsequent sections, we discuss kernel estimation of the model (1.1).

3. Estimation based on kernel smoothing.

We will introduce a kernel estimator based on backfitting and develop a complete asymptotic theory for this estimator. Again, we will do this for the case where i.i.d. observations $(\mathbf{X}^i, Y^i)$ are made on the random vector $(\mathbf{X}, Y)$. Model (1.1) can be rewritten as (1.4) with constraints (1.3). It can be shown that the parameters α_j and α_{jk} in (1.4) can be estimated at a faster rate than the nonparametric functions f_{jk}. For this reason we neglect the parametric parts for simplicity of presentation. Thus we consider model (1.1) with the constraints (1.3). In this setting, our alternative representation (2.2) becomes

$$g(m(\mathbf{x})) = \tilde{\mathbf{x}}_1^\top \mathbf{f}_1(x_{r+1}) + \cdots + \tilde{\mathbf{x}}_p^\top \mathbf{f}_p(x_{r+p}). \tag{3.1}$$

Let Q be the quasi-likelihood function such that $\partial Q(\mu, y)/\partial\mu = (y - \mu)/V(\mu)$, where V is a function for modeling the conditional variance $v(\mathbf{x}) \equiv \mathrm{var}(Y|\mathbf{X} = \mathbf{x})$ by $v(\mathbf{x}) = V(m(\mathbf{x}))$. The quasi-likelihood for the mean regression function m is then given by $\sum_{i=1}^n Q(m(\mathbf{X}^i), Y^i)$, and taking into account the structure of the model (3.1) the quasi-likelihood for $\mathbf{f}_k$ is

$$\sum_{i=1}^n Q(g^{-1}(\tilde{\mathbf{X}}_1^{i\top} \mathbf{f}_1(X_{r+1}^i) + \cdots + \tilde{\mathbf{X}}_p^{i\top} \mathbf{f}_p(X_{r+p}^i)), Y^i). \tag{3.2}$$

We take the smooth backfitting approach [Mammen, Linton and Nielsen (1999), Yu, Park and Mammen (2008), Lee, Mammen and Park (2010, 2012)]. We maximize the integrated kernel-weighted quasi-likelihood

$$L_Q(\boldsymbol{\eta}) \equiv \int n^{-1} \sum_{i=1}^{n} Q(g^{-1}(\boldsymbol{\eta}_1(z_1)^\top \tilde{\mathbf{X}}_1^i + \cdots + \boldsymbol{\eta}_p(z_p)^\top \tilde{\mathbf{X}}_p^i), Y^i) \times K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z} \tag{3.3}$$

over the tuple of functions $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p)$, each $\boldsymbol{\eta}_k$ being a vector of univariate functions that satisfy the constraints of (1.3), where $\mathbf{X}^{c,i} = (X_{r+1}^i, \ldots, X_{r+p}^i)^\top$ and $\mathbf{z} = (z_1, \ldots, z_p)^\top$. Here and throughout the paper, we label the elements of a tuple $\boldsymbol{\eta}$ in such a way that

$$\begin{aligned} \boldsymbol{\eta}(\mathbf{z}) &= (\boldsymbol{\eta}_1(z_1)^\top, \ldots, \boldsymbol{\eta}_p(z_p)^\top)^\top, \\ \boldsymbol{\eta}_k &= \{\eta_{j,r+k} : r + k \in I_j, 1 \le j \le d\} \end{aligned} \tag{3.4}$$

with r introduced at the beginning of Section 2, and with this representation the constraints of (1.3) on the elements of $\boldsymbol{\eta}$ are

$$\begin{aligned} &\int \eta_{jl}(u) w_l(u)\, du = 0, \qquad r+1 \le l \le r+p, 1 \le j \le d; \\ &\int u\eta_{jl}(u) w_l(u)\, du = 0, \qquad r+1 \le j, l \le d. \end{aligned} \tag{3.5}$$

We take a kernel such that

$$\int K_{h_j}(u, v)\, dv = 1 \qquad \text{for all values of } u. \tag{3.6}$$

This kernel can be constructed from the standard kernel of the form $K_{h_j}(u - v) \equiv h_j^{-1} K((u - v)/h_j)$, where K is a symmetric nonnegative function, in such a way that $K_{h_j}(u, v) = K_{h_j}(u - v)/\int K_{h_j}(u - w)\, dw$. It was used in Mammen, Linton and Nielsen (1999), Yu, Park and Mammen (2008) and Lee, Mammen and Park (2010, 2012), and will be used in our technical arguments.

The smooth backfitting estimators of $\mathbf{f}_k$ in our model (3.1) are $\hat{\mathbf{f}}_k$ which maximize L_Q at (3.3). In Section 4, we detail an iterative procedure to get the estimators. Here, we provide their theoretical properties. We begin with some notational definitions. We let $p_{\mathbf{X}^c}$ denote the marginal density of $\mathbf{X}^c = (X_{r+1}, \ldots, X_{r+p})^\top$. Also, we let p_j and p_{jk} be the marginal densities of X_{r+j} and (X_{r+j}, X_{r+k}), respectively. Define $Q_r(u, y) = \partial^r Q(g^{-1}(u), y)/\partial u^r$ and $\mathbf{W}_j(\mathbf{z}; \boldsymbol{\eta}) = (\mathbf{W}_{j1}(\mathbf{z}; \boldsymbol{\eta}), \ldots, \mathbf{W}_{jp}(\mathbf{z}; \boldsymbol{\eta}))$, where

$$\mathbf{W}_{jk}(\mathbf{z}; \boldsymbol{\eta}) = -E[Q_2(\tilde{\mathbf{X}}^\top \boldsymbol{\eta}(\mathbf{X}^c), m(\mathbf{X})) \tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_k^\top | \mathbf{X}^c = \mathbf{z}] p_{\mathbf{X}^c}(\mathbf{z})$$

for $1 \le j, k \le p$, and $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1^\top, \ldots, \tilde{\mathbf{X}}_p^\top)^\top$. Let

$$\mathbf{W}(\mathbf{z};\boldsymbol{\eta}) = (\mathbf{W}_1(\mathbf{z};\boldsymbol{\eta})^\top, \ldots, \mathbf{W}_p(\mathbf{z};\boldsymbol{\eta})^\top)^\top = \mathbf{W}(\mathbf{z};\boldsymbol{\eta})^\top.$$

Throughout this paper we write $\mathbf{W}(\mathbf{z}) = \mathbf{W}(\mathbf{z};\mathbf{f})$, where $\mathbf{f}$ is the true tuple of the coefficient functions. With slight abuse of notation, we also write

$$\begin{aligned}
\mathbf{W}_{jk}(z_j, z_k;\boldsymbol{\eta}) &= -E[Q_2(\tilde{\mathbf{X}}^\top \boldsymbol{\eta}(\mathbf{X}^c), m(\mathbf{X}))\tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_k^\top | X_{r+j} = z_j, X_{r+k} = z_k] \\
&\quad \times p_{jk}(z_j, z_k), \\
\mathbf{W}_{jj}(z_j;\boldsymbol{\eta}) &= -E[Q_2(\tilde{\mathbf{X}}^\top \boldsymbol{\eta}(\mathbf{X}^c), m(\mathbf{X}))\tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_k^\top | X_{r+j} = z_j] p_j(z_j).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbf{W}_{jj}(z_j;\boldsymbol{\eta}) &= \int \mathbf{W}_{jj}(\mathbf{z};\boldsymbol{\eta})\, d\mathbf{z}_{-j}, \mathbf{W}_{jk}(z_j, z_k;\boldsymbol{\eta}) = \int \mathbf{W}_{jk}(\mathbf{z};\boldsymbol{\eta})\, d\mathbf{z}_{-(j,k)}, \\
\mathbf{W}_{jj}(z_j) &= \int \mathbf{W}_{jj}(\mathbf{z})\, d\mathbf{z}_{-j}, \mathbf{W}_{jk}(z_j, z_k) = \int \mathbf{W}_{jk}(\mathbf{z})\, d\mathbf{z}_{-(j,k)}.
\end{aligned}$$

Here and throughout the paper, $\mathbf{z}_{-j}$ for a given vector $\mathbf{z}$ denotes the vector without its jth entry, and $\mathbf{z}_{-(j,k)}$ without its jth and kth entries. Due to the conditions (A0) and (A1), $\mathbf{W}_{jj}(z_j)$ is positive definite for all $z_j \in [0,1]$. However, $\mathbf{W}(\mathbf{z})$ may not be positive definite.

The relevant space of functions we are dealing with is the one that consists of tuples $\boldsymbol{\eta}$ of univariate functions with the representation at (3.4) such that its elements satisfy the constraints (3.5). For any two functions $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$ of this type, define $\langle \boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)} \rangle_\# = \int \boldsymbol{\eta}^{(1)\top}(\mathbf{z})\mathbf{W}(\mathbf{z})\boldsymbol{\eta}^{(2)}(\mathbf{z})\, d\mathbf{z}$ whenever it exists. We denote by $\mathcal{H}(\mathbf{W})$ the resulting space of tuples $\boldsymbol{\eta}$. The space is equipped with the inner product $\langle \cdot, \cdot \rangle_\#$. Let $\| \cdot \|_\#$ be its induced norm, that is, $\|\boldsymbol{\eta}\|_\#^2 = \int \boldsymbol{\eta}(\mathbf{z})^\top \mathbf{W}(\mathbf{z})\boldsymbol{\eta}(\mathbf{z})\, d\mathbf{z}$.

Define $p_{j,\mathbf{X}}^{(1)}(\mathbf{x}) = \partial p_{\mathbf{X}}(\mathbf{x})/\partial x_{r+j}$. Likewise, define $m_j^{(1)}(\mathbf{x}) = \partial m(\mathbf{x})/\partial x_{r+j}$ and $m_j^{(2)}(\mathbf{x}) = \partial^2 m(\mathbf{x})/\partial x_{r+j}^2$. Let $\boldsymbol{\Delta}_{jk}$ denote a d_j-vector such that its ℓth element $\Delta_{jk,\ell} = 1$ if the ℓth element of $\tilde{\mathbf{x}}_j$ in our rearrangement (2.1) equals x_{r+k} and $\Delta_{jk,\ell} = 0$ otherwise. Define a $(d_1 + \cdots + d_p)$-vector $\boldsymbol{\Delta}_k$ by $\boldsymbol{\Delta}_k^\top = (\boldsymbol{\Delta}_{1k}^\top, \ldots, \boldsymbol{\Delta}_{pk}^\top)$. In the assumptions given in the Appendix, we assume $n^{1/5}h_j \to c_j$ as $n \to \infty$ for some constants $0 < c_j < \infty$. For such constants, define

$$\tilde{\boldsymbol{\beta}}_j(z_j) = \mathbf{W}_{jj}(z_j)^{-1} p_j(z_j) \sum_{k=1}^{p} c_k^2 E(\mathbf{b}_{jk}(\mathbf{X}) | X_{r+j} = z_j) \int t^2 K(t)\, dt,$$

where $\mathbf{b}_{jk}(\mathbf{X})$ are d_j-vectors given by

$$\mathbf{b}_{jk}(\mathbf{X}) = \left(m_k^{(1)}(\mathbf{X}) - \frac{\boldsymbol{\Delta}_k^\top \mathbf{f}(\mathbf{X}^c)}{g'(m(\mathbf{X}))} \right)$$

$$
\begin{aligned}
&\times \Biggl[\frac{\tilde{\mathbf{X}}_j}{V(m(\mathbf{X}))g'(m(\mathbf{X}))}\frac{p^{(1)}_{k,\mathbf{X}}(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})} \\
&\qquad - \tilde{\mathbf{X}}_j \boldsymbol{\Delta}_k^\top \mathbf{f}(\mathbf{X}^c)\biggl(\frac{V'(m(\mathbf{X}))}{V(m(\mathbf{X}))^2 g'(m(\mathbf{X}))^2} + \frac{g''(m(\mathbf{X}))}{V(m(\mathbf{X}))g'(m(\mathbf{X}))^3}\biggr) \\
&\qquad\qquad + \frac{\boldsymbol{\Delta}_{jk}}{V(m(\mathbf{X}))g'(m(\mathbf{X}))}\Biggr] \\
&+ \frac{1}{2}\frac{\tilde{\mathbf{X}}_j}{V(m(\mathbf{X}))g'(m(\mathbf{X}))}\biggl(m_k^{(2)}(\mathbf{X}) + \frac{g''(m(\mathbf{X}))(\boldsymbol{\Delta}_k^\top \mathbf{f}(\mathbf{X}^c))^2}{g'(m(\mathbf{X}))^3}\biggr).
\end{aligned}
$$

Let $\boldsymbol{\beta}_*(\mathbf{z}) = (\boldsymbol{\beta}_{*j}(z_j) : 1 \le j \le p)$ be a solution of

$$
\boldsymbol{\beta}_{*j}(z_j) = \tilde{\boldsymbol{\beta}}_j(z_j) - \sum_{k \neq j} \int [\mathbf{W}_{jj}(z_j)]^{-1} \mathbf{W}_{jk}(z_j, z_k)\boldsymbol{\beta}_{*k}(z_k)\, dz_k, \qquad 1 \le j \le p, \tag{3.7}
$$

and put $\boldsymbol{\beta}_j(z_j)$ to be the normalized versions of $\boldsymbol{\beta}_{*j}(z_j)$ so that they satisfy the constraints (3.5). Below in a theorem we show that $\boldsymbol{\beta}(\mathbf{z}) \equiv (\boldsymbol{\beta}_1(z_1)^\top, \ldots, \boldsymbol{\beta}_p(z_p)^\top)^\top$ is the asymptotic bias of the smooth backfitting estimator $\hat{\mathbf{f}}(\mathbf{z}) = (\hat{\mathbf{f}}_1(z_1)^\top, \ldots, \hat{\mathbf{f}}_p(z_p)^\top)^\top$.

For the special case where the matrix $\mathbf{W}(\mathbf{z})$ is invertible, one has an interpretation of $\boldsymbol{\beta}(\mathbf{z})$ as the projection of the asymptotic bias of the full-dimensional local quasi-likelihood estimator that maximizes

$$
n^{-1}\sum_{i=1}^n Q(g^{-1}(\boldsymbol{\eta}_1(\mathbf{z})^\top \tilde{\mathbf{X}}_1^i + \cdots + \boldsymbol{\eta}_p(\mathbf{z})^\top \tilde{\mathbf{X}}_p^i), Y^i) K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z}).
$$

One can check that its asymptotic bias, $\boldsymbol{\beta}_{\mathrm{mlt}}(\mathbf{z})$, is given by

$$
\boldsymbol{\beta}_{\mathrm{mlt}}(\mathbf{z}) = \mathbf{W}(\mathbf{z})^{-1} p_{\mathbf{X}^c}(\mathbf{z}) \sum_{k=1}^p c_k^2 E(\mathbf{b}_k(\mathbf{X}) | \mathbf{X}^c = \mathbf{z}) \int t^2 K(t)\, dt,
$$

where $\mathbf{b}_k(\mathbf{X})^\top = (\mathbf{b}_{1k}(\mathbf{X})^\top, \ldots, \mathbf{b}_{pk}(\mathbf{X})^\top)$.

The tuple $\boldsymbol{\beta}_{\mathrm{mlt}}$ does not belong to $\mathcal{H}(\mathbf{W})$. The asymptotic bias $\boldsymbol{\beta}(\mathbf{z})$ of the smooth backfitting estimator $\hat{\mathbf{f}}(\mathbf{z})$ is identical to the projection of $\boldsymbol{\beta}_{\mathrm{mlt}}$ onto $\mathcal{H}(\mathbf{W})$, that is, $\boldsymbol{\beta} = \operatorname{argmin}_{\boldsymbol{\eta} \in \mathcal{H}(\mathbf{W})} \|\boldsymbol{\beta}_{\mathrm{mlt}} - \boldsymbol{\eta}\|_\#$. This projection interpretation of $\boldsymbol{\beta}(\mathbf{z})$ is not available if $\mathbf{W}(\mathbf{z})$ is not invertible. In general, one has to define $\boldsymbol{\beta}(\mathbf{z})$ through the integral equations (3.7).

For the asymptotic variance of $\hat{\mathbf{f}}$, we define

$$
\boldsymbol{\Sigma}_j(z_j) = \frac{1}{c_j p_j(z_j)} \int K^2(t)\, dt \biggl[E\biggl(\frac{\tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_j^\top}{V(m(\mathbf{X}))g'(m(\mathbf{X}))^2}\bigg| X_{r+j} = z_j\biggr)\biggr]^{-1}
$$

$$\times E\left(\frac{v(\mathbf{X})\tilde{\mathbf{X}}_j\tilde{\mathbf{X}}_j^\top}{V(m(\mathbf{X}))^2 g'(m(\mathbf{X}))^2}\Big| X_{r+j}=z_j\right)$$

$$\times \left[E\left(\frac{\tilde{\mathbf{X}}_j\tilde{\mathbf{X}}_j^\top}{V(m(\mathbf{X}))g'(m(\mathbf{X}))^2}\Big| X_{r+j}=z_j\right)\right]^{-1},$$

where $v(\mathbf{x})$ is the conditional variance of Y given $\mathbf{X}=\mathbf{x}$.

Theorem 2. *Under* (A0) *in Section 2 and those* (A1)–(A5) *in the Appendix, there exists a unique maximizer $\hat{\mathbf{f}}$ of the integrated kernel-weighted quasi-likelihood (3.3) with probability tending to one. The maximizer $\hat{\mathbf{f}}$ satisfies*

$$\int |\hat{\mathbf{f}}(\mathbf{z})-\mathbf{f}(\mathbf{z})|^2 p_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}=O_p(n^{-4/5}),$$

$$\sup_{z_j\in[2h_j,1-2h_j]}|\hat{\mathbf{f}}_j(z_j)-\mathbf{f}_j(z_j)|=O_p(n^{-2/5}\sqrt{\log n}),$$

where $|\cdot|$ denote the Euclidean norm.

Theorem 3. *Assume that* (A0) *in Section 2 and those* (A1)–(A5) *in the Appendix hold. Then, for all $\mathbf{z}$ in the interior of the support of $p_{\mathbf{X}^c}$, it follows that $n^{2/5}(\hat{\mathbf{f}}_j(z_j)-\mathbf{f}_j(z_j))$ are jointly asymptotically normal with mean $(\boldsymbol{\beta}_1(z_1)^\top,\dots,\boldsymbol{\beta}_p(z_p)^\top)^\top$ and variance* $\operatorname{diag}(\boldsymbol{\Sigma}_j(z_j))$.

In the special case where $m(\mathbf{x})=x_1f_1(x_{p+1})+\cdots+x_pf_d(x_{2p})$ for $\mathbf{x}=(x_1,\dots,x_{2p})^\top$, thus the link g is the identity function and $Q(\mu,y)=-(y-\mu)^2/2$, the asymptotic bias and variance of $\hat{\mathbf{f}}_j(z_j)$ stated in Theorem 3 coincide with those in Theorem 2 of Lee, Mammen and Park (2012). This can be seen by noting that $\mathbf{W}(\mathbf{z})$ is invertible in this case, and that $V=g'=1$, $\tilde{\mathbf{X}}=(X_1,\dots,X_p)^\top$, $\mathbf{X}^c=(X_{p+1},\dots,X_{2p})^\top$ and $\mathbf{W}(\mathbf{z})^{-1}E(\tilde{\mathbf{X}}X_j|\mathbf{X}^c=\mathbf{z})p_{\mathbf{X}^c}(\mathbf{z})=\mathbf{1}_j$ where $\mathbf{1}_j$ is the p-dimensional unit vector with the jth entry being equal to one.

Theorem 3 can be also viewed as an extension of Theorem 2 of Yu, Park and Mammen (2008). In the latter work, smooth backfitting for the additive model $g(m(\mathbf{z}))=f_1(z_1)+\cdots+f_p(z_p)$ for a link g was considered. As we mentioned earlier, model (3.1) reduces to the above model by taking $d_k\equiv 1$, $\tilde{\mathbf{X}}_k=X_k\equiv 1$ for $1\le k\le p$ and $r=p$. In this case, $\mathbf{W}(\mathbf{z})$ is not invertible so that the projection interpretation of $\boldsymbol{\beta}(\mathbf{z})$ is not valid. If one replaces $m(\mathbf{x})$ in the formula of $\tilde{\boldsymbol{\beta}}_j(z_j)$ by $g^{-1}(f(\mathbf{x}^c))\equiv g^{-1}(f_1(x_{r+1})+\cdots+f_p(x_{r+p}))$, $\mathbf{W}_{jk}$ by the corresponding quantities for the latter model, which are

$$W_{jk}(z_j,z_k)=\int V(g^{-1}(f(\mathbf{z})))^{-1}g'(g^{-1}(f(\mathbf{z})))^{-2}p_{\mathbf{X}^c}(\mathbf{z})\,d\mathbf{z}_{-(j,k)},$$

$$W_{jj}(z_j)=\int V(g^{-1}(f(\mathbf{z})))^{-1}g'(g^{-1}(f(\mathbf{z})))^{-2}p_{\mathbf{X}^c}(\mathbf{z})\,d\mathbf{z}_{-j},$$

then one can verify that the solution of the system of integral equations in (3.7) concides with the asymptotic bias in Yu, Park and Mammen (2008). The asymptotic variance $\mathbf{\Sigma}_j(z_j)$ given above also reduces to the one in Yu, Park and Mammen (2008).

REMARK 1. In the case where the link g in model (3.1) is the identity (or a linear) function and when the covariates X_j are independent, one may apply marginally a kernel smoothing method to estimating each coefficient function. To see this, suppose that all $\tilde{\mathbf{X}}_j$ contains 1 as their first entry and any entry of $\tilde{\mathbf{X}}_j$ does not equal to any of X_{r+k}, $k \neq j$. Then, $E(Y|\tilde{\mathbf{X}}_j, X_{r+j}) = \tilde{\mathbf{X}}_j^\top[\mathbf{f}_j(X_{r+j}) + \mathbf{c}_j]$ for some constant vector $\mathbf{c}_j$. This means that $\mathbf{f}_j$ minimizes $E[Y - \tilde{\mathbf{X}}_j^\top \boldsymbol{\eta}_j(X_{r+1})]^2$ over $\boldsymbol{\eta}_j$ subject to a normalization. Thus, the marginal smoothing that minimizes $n^{-1}\sum_{i=1}^n (Y^i - \boldsymbol{\eta}_j^\top \tilde{\mathbf{X}}_j^i)^2 K_{h_j}(X_{r+j}^i, z_j)$ for each j and each point z_j gives a consistent estimator of $\mathbf{f}_j(z_j)$. This marginal smoothing approach is not valid, even with independent covariates, in case the link function g is nonlinear. In the latter case, one needs to use a projection method such as the smooth backfitting defined above, or a marginal integration technique, to obtain appropriate estimators.

4. Implementation.

In this section, we discuss how to find $\hat{\mathbf{f}}$ maximizing L_Q at (3.3). Our method of finding $\hat{\mathbf{f}}$ is based on an iteration scheme. By considering the Fréchet differentials of L_Q, we see that

$$\int n^{-1} \sum_{i=1}^n Q_1(\tilde{\mathbf{X}}^{i\top} \hat{\mathbf{f}}(\mathbf{z}), Y^i) \tilde{\mathbf{X}}_j^i K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j} = \mathbf{0}_j, \qquad 1 \le j \le p, z_j \in [0,1], \tag{4.1}$$

where $\mathbf{0}_j$ is the zero vector of dimension d_j. The system of equations is nonlinear. We take the Newton–Raphson approach to find a solution by iteration. For a vector of functions $(\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_p^\top)^\top$ where $\boldsymbol{\eta}_j(\mathbf{z}) = \boldsymbol{\eta}_j(z_j)$, define

$$\hat{\mathbf{F}}_j(\boldsymbol{\eta})(z_j) = \int n^{-1} \sum_{i=1}^n Q_1(\tilde{\mathbf{X}}^{i\top} \boldsymbol{\eta}(\mathbf{z}), Y^i) \tilde{\mathbf{X}}_j^i K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j}. \tag{4.2}$$

The system of equations in (4.1) is then expressed as $\hat{\mathbf{F}}_j(\hat{\mathbf{f}}) = \mathbf{0}_j, 1 \le j \le p$. Our algorithm runs an *outer* iteration which is based on a Newton–Raphson approximation of the system of equations. Each outer-step solves a linearized system of equations to update the approximation of $\hat{\mathbf{f}}$, which requires an additional iteration, called *inner* iteration.

To describe the algorithm, suppose that we are at the sth outer-step to update $\hat{\mathbf{f}}^{[s-1]}$ in the previous outer-step. Considering the Fréchet differentials of $\hat{\mathbf{F}}_j$ at $\hat{\mathbf{f}}^{[s-1]}$, we get the following approximation: for $1 \le j \le p$,

$$\begin{aligned}\hat{\mathbf{F}}_j(\hat{\mathbf{f}})(z_j) &\simeq \hat{\mathbf{F}}_j(\hat{\mathbf{f}}^{[s-1]})(z_j)\\ &\quad + \sum_{k=1}^{p} \int n^{-1} \sum_{i=1}^{n} Q_2(\tilde{\mathbf{X}}^{i\top} \hat{\mathbf{f}}^{[s-1]}(\mathbf{z}), Y^i)\\ &\qquad \times \tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_k^{i\top} [\hat{\mathbf{f}}_k(z_k) - \hat{\mathbf{f}}_k^{[s-1]}(z_k)] K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j}.\end{aligned} \tag{4.3}$$

This gives an updating equation for $\hat{\mathbf{f}}^{[s]}$. Define, for $1 \le j \le p$ and for $1 \le j \ne k \le p$,

$$\hat{\mathbf{W}}_{jk}^{[s]}(z_j, z_k) = -\int n^{-1} \sum_{i=1}^{n} Q_2(\tilde{\mathbf{X}}^{i\top} \hat{\mathbf{f}}^{[s]}(\mathbf{z}), Y^i) \tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_k^{i\top} K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-(j,k)},$$

$$\hat{\mathbf{W}}_{jj}^{[s]}(z_j) = -\int n^{-1} \sum_{i=1}^{n} Q_2(\tilde{\mathbf{X}}^{i\top} \hat{\mathbf{f}}^{[s]}(\mathbf{z}), Y^i) \tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_j^{i\top} K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j}.$$

Define $\tilde{\boldsymbol{\delta}}_j^{[s]}(z_j) = [\hat{\mathbf{W}}_{jj}^{[s]}(z_j)]^{-1} \hat{\mathbf{F}}_j(\hat{\mathbf{f}}^{[s]})(z_j)$. Also, let $\hat{\boldsymbol{\delta}}_j^{[s]}(z_j) = \hat{\mathbf{f}}_j^{[s]}(z_j) - \hat{\mathbf{f}}_j^{[s-1]}(z_j)$. We get from (4.3) the following linearized system of updating equations:

$$\hat{\boldsymbol{\delta}}_j^{[s]}(z_j) = \tilde{\boldsymbol{\delta}}_j^{[s-1]}(z_j) - \sum_{k \ne j} \int [\hat{\mathbf{W}}_{jj}^{[s-1]}(z_j)]^{-1} \hat{\mathbf{W}}_{jk}^{[s-1]}(z_j, z_k) \hat{\boldsymbol{\delta}}_k^{[s]}(z_k)\, dz_k, \qquad 1 \le j \le p. \tag{4.4}$$

Solving the system of equations (4.4) for $\hat{\boldsymbol{\delta}}_j^{[s]}$ and then updating $\hat{\mathbf{f}}_j^{[s-1]}$ by $\hat{\mathbf{f}}_j^{[s]} = \hat{\mathbf{f}}_j^{[s-1]} + \hat{\boldsymbol{\delta}}_j^{[s]}$ constitutes the sth step in the outer iteration. Below in Theorem 4, we show that the outer iteration converges to a solution that satisfies (4.1).

The system of equations (4.4) cannot be solved since to get $\hat{\boldsymbol{\delta}}_j^{[s]}$ one requires knowledge of the other $\hat{\boldsymbol{\delta}}_k^{[s]}, k \ne j$. To solve (4.4) we need an *inner* iteration. Suppose that we are at the ℓth inner-step of the sth outer-step to update $\hat{\boldsymbol{\delta}}_j^{[s,\ell-1]}, 1 \le j \le p$, in the previous inner iteration step. We apply (4.4):

$$\begin{aligned}\hat{\boldsymbol{\delta}}_j^{[s,\ell]}(z_j) = \tilde{\boldsymbol{\delta}}_j^{[s-1]}(z_j) - \sum_{k \ne j} \int [\hat{\mathbf{W}}_{jj}^{[s-1]}(z_j)]^{-1} \hat{\mathbf{W}}_{jk}^{[s-1]}(z_j, z_k)&\\ \times \hat{\boldsymbol{\delta}}_k^{[s,\ell-1]}(z_k)\, dz_k, \qquad 1 \le j \le p.&\end{aligned} \tag{4.5}$$

Existence of a unique solution of (4.4) and the convergence of the inner iteration to the solution are demonstrated below in Theorem 4. For the starting values $\hat{\boldsymbol{\delta}}_j^{[s,0]}$ in the inner iteration of the sth outer-step, one may use the limit of the inner iteration in the previous outer-step $\hat{\boldsymbol{\delta}}_j^{[s-1,\infty]}$.

For a convergence criterion of the outer iteration, one may check whether the values of the left-hand side of (4.1) are sufficiently small, or use the difference between the two updates $\hat{\mathbf{f}}_k^{[s-1]}$ and $\hat{\mathbf{f}}_k^{[s]}$:

$$\int |\hat{\mathbf{f}}_k^{[s]}(z_k) - \hat{\mathbf{f}}_k^{[s-1]}(z_k)|^2 \, dz_k. \tag{4.6}$$

In the latter case, one should use the normalized versions of the updates. Recall the configuration of $\boldsymbol{\eta}_k$ in (3.4). The normalized version of a given set of tuples $\boldsymbol{\eta}_{*k}$ may be obtained by the following formulas. Let the weight functions w_l be normalized so that $\int w_l(u)\,du = 1$. Then

$$\begin{aligned}
\eta_{jl}(u) &= \eta_{jl*}(u) \\
&\quad - \int \eta_{*jl}(u) w_l(u)\,du, \qquad 1 \le j \le r \text{ or } d+1 \le l \le r+p, \qquad (4.7)\\
\eta_{jl}(u) &= \eta_{jl*}(u) - a_{jl} - b_{jl}u, \qquad r+1 \le j, l \le d,
\end{aligned}$$

where

$$\begin{aligned}
a_{jl} &= \int \eta_{*jl}(u) w_l(u)\,du - b_{jl} \int u w_l(u)\,du, \\
b_{jl} &= \left[\int \left(u - \int t w_l(t)\,dt \right)^2 w_l(u)\,du \right]^{-1} \\
&\quad \times \int \left(u - \int t w_l(t)\,dt \right) \eta_{*jl}(u) w_l(u)\,du.
\end{aligned}$$

One should also use the normalized $\hat{\boldsymbol{\delta}}_{jl}$ for the convergence of the inner iteration.

THEOREM 4. *Assume that* (A0) *in Section 2 and* (A1)–(A5) *in the Appendix hold. Then there exist constants* $0 < C_1, \tau < \infty$ *and* $0 < \gamma < 1$ *such that, if the initial choice* $\hat{\mathbf{f}}^{[0]}$ *satisfies*

$$\int |\hat{\mathbf{f}}^{[0]}(\mathbf{z}) - \hat{\mathbf{f}}(\mathbf{z})|^2 p_{\mathbf{X}^c}(\mathbf{z})\,d\mathbf{z} \le \tau^2 \tag{4.8}$$

with probability tending to one, then $\int |\hat{\mathbf{f}}^{[s]}(\mathbf{z}) - \hat{\mathbf{f}}(\mathbf{z})|^2 p_{\mathbf{X}^c}(\mathbf{z})\,d\mathbf{z} \le C_1 4^{-(s-1)} \gamma^{2^s - 1}$ *with probability tending to one. Also, for each outer-step there exists a solution of the system of equations (4.4) that is unique, and the inner iteration converges at a geometric rate. Furthermore, if the initial choice* $\hat{\mathbf{f}}^{[0]}$ *satisfies (4.8) with probability tending to one, then there exist some constants* $0 < C_2 < \infty$ *and* $0 < \rho < 1$ *such that, with probability tending to one,* $\int |\hat{\boldsymbol{\delta}}^{[s,\ell]}(\mathbf{z}) - \hat{\boldsymbol{\delta}}^{[s,\infty]}(\mathbf{z})|^2 p_{\mathbf{X}^c}(\mathbf{z})\,d\mathbf{z} \le C_2 \rho^{2\ell}$ *for sufficiently large* s, *where* $\hat{\boldsymbol{\delta}}^{[s,\infty]}$ *is a solution of the system of equations in (4.4).*

The theorem shows that the number of iterations that is needed for a desired accuracy of the calculation of the backfitting estimator does not depend on the sample size. If the desired accuracy is of order n^{-c} for some constant c, then a logarithmic number of iterations suffices. Thus the complexity of the algorithm only increases very moderately for increasing sample size. We have no good bound on the required accuracy of the starting values, that is, on the choice of τ. In our practical experience the algorithm was very robust against poor choices of the starting value. In fact, in the simulation study we chose $\hat{\mathbf{f}}^{[0]} = \mathbf{0}$ and it worked quite well. A more deliberate choice is a version of the marginal integration estimator studied by Yang et al. (2006), or a spline estimator that we discussed in Section 2. These are consistent so that they satisfy the condition (4.8), but they cost additional numerical computation.

As for the choices of the bandwidths h_j, one may estimate the optimal bandwidths $h_j^{\mathrm{opt}} = c_j^* n^{-1/5}$, where $\mathbf{c}^* = (c_1^*, \ldots, c_p^*)$ is defined by

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \sum_{j=1}^{p} \int [|\boldsymbol{\beta}_j(z_j, \mathbf{c})|^2 + \operatorname{trace}(\boldsymbol{\Sigma}_j(z_j, c_j))] p_{Z_j}(z_j)\, dz_j. \tag{4.9}$$

Here, we write $\boldsymbol{\beta}_j(z_j, \mathbf{c})$ and $\boldsymbol{\Sigma}_j(z_j, c_j)$, instead of $\boldsymbol{\beta}_j(z_j)$ and $\boldsymbol{\Sigma}_j(z_j)$ as defined in Section 3, to stress their dependence on the vector of the bandwidth constants $\mathbf{c} = (c_1, \ldots, c_p)$. To describe a simple plug-in method, get parametric estimates of f_{jk} by maximizing (3.2) over the class of pth order polynomials $f_{jk}(x) = a_{jk}^{(0)} + a_{jk}^{(1)} x + \cdots + a_{jk}^{(p)} x^p$, and obtain a kernel estimate of the density $p_{\mathbf{X}}$. Then one can estimate $\boldsymbol{\beta}_j$ by plugging these estimates into the formulas of $\tilde{\boldsymbol{\beta}}_j, \mathbf{W}_{jj}, \mathbf{W}_{jk}$ $(j \neq k)$ and solving the system of equations in (3.7) by iteration. One can also estimate $\boldsymbol{\Sigma}_j$. Put these estimates of $\boldsymbol{\beta}_j$ and $\boldsymbol{\Sigma}_j$ into the right-hand side of (4.9) to get an estimate of $\mathbf{c}^*$. A similar idea was adopted by Lee, Mammen and Park (2012) for the classical varying coefficient model. An alternative way is to develop a method similar to the penalized least squares bandwidth selector proposed by Mammen and Park (2005). This would need higher-order stochastic expansions for the quasi-likelihood of the smooth backfitting estimators. Finally, we want to mention that $\boldsymbol{\beta}_j$ depends on the whole vector $\mathbf{c}$ contrary to $\boldsymbol{\Sigma}_j$, the latter only involving c_j. This is not the case with the local linear smooth backfitting where both depend on c_j only, see the next section. Thus, a grid search for $\mathbf{c}^*$ at (4.9) may be computationally expensive for large p. In this case, one may apply an iteration scheme which, in each iteration step, updates c_j for $1 \leq j \leq p$ one by one with the other $c_k, k \neq j$, being held fixed at the values obtained in the previous step.

5. Extension to higher-order local smoothing.

In the previous two sections we considered smooth backfitting based on Nadaraya–Watson smoothing. Here, we discuss its extension to local polynomial smoothing. We focus on the local linear case. The extension to the general case is immediate, but needs more involved notation. For a function η of interest, the basic idea of local linear smoothing is to approximate $\eta(u)$ for u near a point z by $\eta(z) + \eta'(z)(u - x)$, where η' is the first derivative of η. Thus, we maximize

$$\int n^{-1} \sum_{i=1}^{n} Q(g^{-1}(\tilde{\mathbf{X}}_1^{i\top} \boldsymbol{\eta}_1(z_1, X_{r+1}^i) + \cdots + \tilde{\mathbf{X}}_p^{i\top} \boldsymbol{\eta}_p(z_p, X_{r+p}^i)), Y^i) \\ \times K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z},$$

where $\boldsymbol{\eta}_j(z_j, X_{r+j}^i) = \boldsymbol{\eta}_{j0}(z_j) + \boldsymbol{\eta}_{j1}(z_j) h_j^{-1}(X_{r+j}^i - z_j)$. The maximizers, denoted by $\hat{\mathbf{f}}_{j0}$ and $\hat{\mathbf{f}}_{j1}$ which correspond to $\boldsymbol{\eta}_{j0}$ and $\boldsymbol{\eta}_{j1}$, respectively, are the estimators of the true $\mathbf{f}_j$ and $h_j \mathbf{f}_j'$, where $\mathbf{f}_j'$ is the vector of the first derivatives of the entries in $\mathbf{f}_j$. Again, $\hat{\mathbf{f}}_{j0}$ should be normalized according to (4.7).

To describe the algorithms, write $\hat{\mathbf{f}}_j = (\hat{\mathbf{f}}_{j0}^\top, \hat{\mathbf{f}}_{j1}^\top)^\top$. They satisfy $\hat{\mathbf{F}}_j(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_p) = \mathbf{0}_j$, $1 \le j \le p$, where $\mathbf{0}_j$ denotes now the zero vector of dimension $2d_j$,

$$\hat{\mathbf{F}}_j(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p) = \int n^{-1} \sum_{i=1}^{n} Q_1\left(\sum_{j=1}^{p} \tilde{\mathbf{X}}_j^{i\top} \boldsymbol{\eta}_j(z_j, X_{r+j}^i), Y^i\right) \\ \times \mathbf{a}(X_{r+j}^i, z_j) \otimes \tilde{\mathbf{X}}_j^i K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j}$$

$\boldsymbol{\eta}_j = (\boldsymbol{\eta}_{j0}^\top, \boldsymbol{\eta}_{j1}^\top)^\top$ and $\mathbf{a}(X_{r+j}^i, z_j) = (1, h_j^{-1}(X_{r+j}^i - z_j))^\top$. The expressions for the updating equations at (4.4) and (4.5) are unchanged if, writing $\mathbf{A}_{jk}^i = \mathbf{a}(X_{r+j}^i, z_j)\mathbf{a}(X_{r+j}^i, z_k)^\top$, we redefine $\hat{\mathbf{W}}_{jk}^{[s]}$, $1 \le j, k \le p$, by

$$\hat{\mathbf{W}}_{jk}^{[s]}(z_j, z_k) = -\int n^{-1} \sum_{i=1}^{n} Q_2\left(\sum_{j=1}^{p} \tilde{\mathbf{X}}_j^{i\top} \hat{\mathbf{f}}_j^{[s]}(z_j, X_{r+j}^i), Y^i\right) \\ \times \mathbf{A}_{jk}^i \otimes \tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_k^{i\top} K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-(j,k)},$$

$$\hat{\mathbf{W}}_{jj}^{[s]}(z_j) = -\int n^{-1} \sum_{i=1}^{n} Q_2\left(\sum_{j=1}^{p} \tilde{\mathbf{X}}_j^{i\top} \hat{\mathbf{f}}_j^{[s]}(z_j, X_{r+j}^i), Y^i\right) \\ \times \mathbf{A}_{jj}^i \otimes \tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_j^{i\top} K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z})\, d\mathbf{z}_{-j}.$$

Let $\boldsymbol{\Sigma}_j$ be defined as in Section 3, and define $\boldsymbol{\beta}_j^{LL}(z_j)$ to be the normalized versions of $c_j^2 \int t^2 K(t)\, dt \mathbf{f}_j''(z_j)/2$ obtained by (4.7), where $\mathbf{f}_j''$ is the vector of the second derivatives of the entries in $\mathbf{f}_j$.

THEOREM 5. *Under* (A0) *in Section 2 and* (A1)–(A5) *in the Appendix, Theorems 2 and 4 remain valid for the local linear smooth backfitting estimators* $(\hat{\mathbf{f}}_{j0}, \hat{\mathbf{f}}_{j1})$ *and for their algorithms, respectively. As a version of Theorem 3,* $n^{2/5}(\hat{\mathbf{f}}_{j0}(z_j) - \mathbf{f}_j(z_j))$ *are jointly asymptotically normal with mean* $(\boldsymbol{\beta}_1^{LL}(z_1)^\top, \ldots, \boldsymbol{\beta}_p^{LL}(z_p)^\top)^\top$ *and variance* $\operatorname{diag}(\boldsymbol{\Sigma}_j(z_j))$.

6. Simulation study. In the simulation study, we considered a binary response Y taking values 0 and 1, and took the following model for the mean function $m(\mathbf{x})$:

$$\begin{aligned} g(m(\mathbf{X})) &= f_{02}(X_2) + f_{03}(X_3) + X_1(f_{12}(X_2) + f_{13}(X_3)) \\ &\quad + X_3 f_{32}(X_2) + X_2 f_{23}(X_3), \end{aligned} \tag{6.1}$$

where $g(u) = \log(u/(1-u))$ is the logit link and $f_{02}(z) = z^2, f_{03}(z) = 4(z - 0.5)^2, f_{12}(z) = z, f_{13}(z) = \cos(2\pi z), f_{32}(z) = e^{2z-1}, f_{23}(z) = \sin(2\pi z)$. The covariate X_1 was a discrete random variable having Bernoulli(0.5) distribution, and X_2 and X_3 were uniform$(0, 1)$ random variables. The three covariates were independent. We chose two sample sizes $n = 500$ and 1000. The number of samples was 500. For the initial estimate, we used $\hat{\mathbf{f}}^{[0]} = \mathbf{0}$. The weight functions w_l were $w_l(z) = I_{[0,1]}(z)$ for all l. We used the Epanechnikov kernel function $K(u) = (3/4)(1 - u^2)I_{[-1,1]}(u)$ and took the theoretically optimal bandwidths as defined at (4.9), which were $h_1^{\mathrm{opt}} = 0.4328, h_2^{\mathrm{opt}} = 0.2789$ for $n = 500$, and $h_1^{\mathrm{opt}} = 0.3768, h_2^{\mathrm{opt}} = 0.2428$ for $n = 1000$, in our simulation setting.

In the simulation, we also computed the cubic spline estimates with K knots placed evenly on the interval $[0, 1]$. We used the power basis for cubic splines: $s_0(z) = 1, s_1(z) = z, s_2(z) = z^2, s_3(z) = z^3, s_{3+k}(z) = (z - \xi_k)_+^3$, where ξ_k are the knot points. If one applies directly the power basis to the model (6.1), one may suffer from "near singularity" of the resulting design matrix. This is because the functions f_{jk} without satisfying our constraints are not identifiable. Taking into consideration the constraints, we adjusted the power basis so that s_1 is orthogonal to s_0, and s_j for $2 \le j \le K + 3$, are orthogonal to s_0 and s_1. The dimension of the power basis for the cubic spline approximation of the model (6.1) equals $6K + 19$. The number of knots taken was $K = 1$ which gave the best performance. The performance of the spline estimators got worse quickly as K increased.

Table 1 shows the results based on 500 datasets. For each component function f_{jk}, the table provides the integrated mean squared error (IMSE), $\int E[\hat{f}_{jk}(z) - f_{jk}(z)]^2\,dz$. The main lesson is that the spline estimators have much larger variances than the smooth backfitting estimators, while the former have smaller biases. Overall, the smooth backfitting method works quite well. Comparing the values of IMSEs for the two sample sizes, the results for the smooth backfitting method reflect the asymptotic effects fairly well. Note that the theoretical reduction of IMSE from $n = 500$ to $n = 1000$ equals $(0.5)^{4/5} \simeq 0.574$. In the simulation we also found the iterative algorithm of the smooth backfitting method in Section 4 converged very fast. The outer loop typically converged in five iterations with the criterion value 10^{-4} for the normalized difference (4.6), and that the inner loop converged in three iterations.

TABLE 1
Integrated mean squared errors (IMSE), integrated squared biases (ISB) and integrated variance (IV) of the two methods, cubic spline (SPL) and smooth backfitting (SBF), for the model (6.1)

			f_{02}	f_{12}	f_{32}	f_{03}	f_{13}	f_{23}
SPL	$n = 500$	IMSE	0.2607	0.2145	0.5034	0.2631	0.2274	0.5857
		ISB	0.0013	0.0004	0.0011	0.0006	0.0044	0.0159
		IV	0.2594	0.2141	0.5022	0.2624	0.2230	0.5699
	$n = 1000$	IMSE	0.1106	0.0817	0.2122	0.1074	0.0938	0.2453
		ISB	0.0001	0.0004	0.0001	0.0001	0.0006	0.0104
		IV	0.1105	0.0813	0.2121	0.1073	0.0932	0.2349
SBF	$n = 500$	IMSE	0.0315	0.0399	0.0274	0.1071	0.1073	0.1685
		ISB	0.0035	0.0112	0.0128	0.0090	0.0543	0.0808
		IV	0.0280	0.0288	0.0147	0.0981	0.0531	0.0877
	$n = 1000$	IMSE	0.0214	0.0210	0.0254	0.0526	0.0702	0.1103
		ISB	0.0021	0.0057	0.0107	0.0062	0.0384	0.0544
		IV	0.0193	0.0153	0.0147	0.0464	0.0318	0.0559

We also investigated how the additional terms in the modeling (1.1) affected the estimation precision when the true model was given by

$$g(m(\mathbf{x},\mathbf{z})) = x_1 f_1(z_1) + \cdots + x_d f_d(z_d)$$

for a set of covariates $(X_1,\ldots,X_d;Z_1,\ldots,Z_d)$. In the latter model, each covariate appears in only one nonlinear interaction term. For this, we estimated the following model:

$$\begin{aligned} g(m(\mathbf{X},\mathbf{Z})) &= f_{01}(Z_1) + f_{02}(Z_2) + X_1(f_{11}(Z_1) + f_{12}(Z_2)) \\ &\quad + X_2(f_{21}(Z_1) + f_{22}(Z_2)), \end{aligned} \tag{6.2}$$

where $f_{01}(z) = f_{02}(z) = 0, f_{11}(z) = \cos(2\pi z), f_{12}(z) = 0, f_{21}(z) = 0, f_{22}(z) = \sin(2\pi z)$, and the link g was the same as in the first example. The covariate X_1 was a discrete random variable having Bernoulli(0.5) distribution, X_2 was the standard normal random variable and Z_1 and Z_2 were uniform$(0,1)$ random variables. The four covariates were independent. The theoretically optimal bandwidths as defined at (4.9) were $h_1^{\mathrm{opt}} = 0.2405, h_2^{\mathrm{opt}} = 0.2469$ for $n = 500$ and $h_1^{\mathrm{opt}} = 0.2093, h_2^{\mathrm{opt}} = 0.2149$ for $n = 1000$, and we used these in the simulation.

The main purpose of this additional simulation is to compare our estimators based on the working model (6.2) with the "oracle" estimators which use the knowledge that $f_{01}(z) = f_{02}(z) = f_{12}(z) = f_{21}(z) = 0$. The system of updating equations for the oracle estimators in our setting is given by (4.4)

TABLE 2
Comparison of the smooth backfitting estimators under the extended model (6.2) and the corresponding oracle estimators

			$\boldsymbol{f_{11}}$	$\boldsymbol{f_{22}}$
Oracle SBF	$n = 500$	IMSE	0.0680	0.0639
		ISB	0.0285	0.0421
		IV	0.0395	0.0218
	$n = 1000$	IMSE	0.0400	0.0433
		ISB	0.0183	0.0309
		IV	0.0216	0.0124
SBF based on (6.2)	$n = 500$	IMSE	0.1057	0.0638
		ISB	0.0273	0.0408
		IV	0.0784	0.0230
	$n = 1000$	IMSE	0.0627	0.0427
		ISB	0.0180	0.0299
		IV	0.0447	0.0128

with the following modifications of $\hat{W}_{jk}^{[s]}(z_j, z_k)$ and $\hat{F}_j(\boldsymbol{\eta})(z_j)$: for $j \neq k$,

$$\hat{W}_{jk}^{[s]}(z_j, z_k) = -n^{-1} \sum_{i=1}^{n} Q_2(X_1^i \hat{f}_{11}^{[s]}(z_1) + X_2^i \hat{f}_{22}^{[s]}(z_2), Y^i) X_j^i X_k^i K_{\mathbf{h}}(\mathbf{Z}^i, \mathbf{z}),$$

$$\begin{aligned}\hat{W}_{jj}^{[s]}(z_j) = -\int n^{-1} \sum_{i=1}^{n} Q_2(X_1^i \hat{f}_{11}^{[s]}(z_1) + X_2^i \hat{f}_{22}^{[s]}(z_2), Y^i) \\ \times (X_j^i)^2 K_{\mathbf{h}}(\mathbf{Z}^i, \mathbf{z}) \, d\mathbf{z}_{-j},\end{aligned}$$

$$\hat{F}_j(\boldsymbol{\eta})(z_j) = \int n^{-1} \sum_{i=1}^{n} Q_1(X_1^i \eta_{11}(z_1) + X_2^i \eta_{22}(z_2), Y^i) X_j^i K_{\mathbf{h}}(\mathbf{Z}^i, \mathbf{z}) \, d\mathbf{z}_{-j},$$

where $\mathbf{Z}^i = (Z_1^i, Z_2^i)^\top$, $\mathbf{z} = (z_1, z_2)^\top$ and $\boldsymbol{\eta}(\mathbf{z}) = (\eta_{11}(z_1), \eta_{22}(z_2))^\top$. Note that all these terms are a scalar, not a matrix or a vector.

Table 2 shows the results based on 500 datasets. For each of the nonzero component functions, the table provides ISB, IV and IMSE. We see that the smooth backfitting estimators perform fairly well in comparison with their oracle versions. In particular, both have nearly the same IMSE, ISB and IV for the estimation of the second component function f_{22}. For estimating f_{11}, the smooth backfitting procedure with the extended model (6.2) gave almost the same bias as the oracle procedure, but a larger variance than the latter. This may be expected since the former has the additional component function f_{12} in the estimation. This was not the case with the estimation of f_{22}, however. The main reason is that the variances of the estimators depend highly on the design of the regressor X_2. Recall that in parametric linear regression the variance of the least squares estimator of a regression coefficient gets smaller as the corresponding regressor is more variable. In our setting, the variability of X_2 is four times as high as that of X_1. This relatively high variability of X_2 alleviated the extra sampling variability of the SBF estimator under the model (6.2).

APPENDIX: TECHNICAL DETAILS

A.1. Proof of Theorem 1. The statement of Theorem 1 follows immediately from the following lemma.

LEMMA 1. *Under assumption* (A0), *it holds that there exist constants* $0 < C_1 < C_2$ *such that for two tuples* $(\boldsymbol{\alpha}, f_{jk} : 1 \le j \le d, k \in I_j)$ *and* $(\boldsymbol{\alpha}^*, f_{jk}^* : 1 \le j \le d, k \in I_j)$ *it holds that*

$$\begin{aligned} & C_1 \int [g(m(\mathbf{x})) - g(m^*(\mathbf{x}))]^2 P_{\mathbf{X}}(d\mathbf{x}) \\ & \quad \le |\boldsymbol{\alpha} - \boldsymbol{\alpha}^*|_*^2 + \sum_{j=1}^{d} \sum_{k \in I_j} [f_{jk}(x_k) - f_{jk}^*(x_k)]^2 p_{X_k}(x_k)\, dx_k \qquad \text{(A.1)} \\ & \quad \le C_2 \int [g(m(\mathbf{x})) - g(m^*(\mathbf{x}))]^2 P_{\mathbf{X}}(d\mathbf{x}). \end{aligned}$$

Here, p_{X_k} *is the density of* X_k, *and*

$$|\boldsymbol{\alpha}|_*^2 = \int \left(\sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k \right)^2 P_{\mathbf{X}}(d\mathbf{x}),$$

$$g(m(\mathbf{x})) = \sum_{j=1}^{d} \alpha_j x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk} x_j x_k + \tilde{\mathbf{x}}_1^\top \mathbf{f}_1(x_{r+1}) + \cdots + \tilde{\mathbf{x}}_p^\top \mathbf{f}_p(x_{r+p}), \qquad \text{(A.2)}$$

$$g(m^*(\mathbf{x})) = \sum_{j=1}^{d} \alpha_j^* x_j + \sum_{\substack{j<k \\ j,k \in \mathcal{C}_0}} \alpha_{jk}^* x_j x_k + \tilde{\mathbf{x}}_1^\top \mathbf{f}_1^*(x_{r+1}) + \cdots + \tilde{\mathbf{x}}_p^\top \mathbf{f}_p^*(x_{r+p}).$$

PROOF. We only prove the second inequality of (A.1). The first one follows by direct arguments. We first observe that because of assumption (A0) it holds that for constants $c_1, c_2 > 0$,

$$\begin{aligned} & \int [g(m(\mathbf{x})) - g(m^*(\mathbf{x}))]^2 P_{\mathbf{X}}(d\mathbf{x}) \\ & \quad \ge c_1 \int [g(m(\mathbf{x})) - g(m^*(\mathbf{x}))]^2 \prod_{l=1}^{D} P_{X_l}(dx_l) \end{aligned}$$

$$\geq c_2 \int [g(m(\mathbf{x})) - g(m^*(\mathbf{x}))]^2 \prod_{l\in\mathcal{C}} w_l(x_l)\,dx_l \prod_{l\notin\mathcal{C}} P_{X_l}(dx_l).$$

Denote by $\mathbf{I}$ the right-hand side of the second inequality. Due to the constraints of (1.3) and the fact that $\tilde{\mathbf{x}}_k$ does not include x_{r+k}, those terms

$$\sum_{j=1}^{d} (\alpha_j - \alpha_j^*)x_j + \sum_{\substack{j<k \\ j,k\in\mathcal{C}_0}} (\alpha_{jk} - \alpha_{jk}^*)x_j x_k,$$

$$\tilde{\mathbf{x}}_1^\top(\mathbf{f}_1(x_{r+1}) - \mathbf{f}_1^*(x_{r+1})), \ldots, \tilde{\mathbf{x}}_p^\top(\mathbf{f}_p(x_{r+p}) - \mathbf{f}_p^*(x_{r+p}))$$

are orthogonal in $L_2(\mu)$, where μ is the product measure defined by $\mu(d\mathbf{x}) = \prod_{j\in\mathcal{C}} w_j(x_j)\,dx_j \prod_{j\notin\mathcal{C}} P_{X_j}(dx_j)$. By this and by making use of (A0) again, we get

$$\begin{aligned}
\mathbf{I} &= c_2 \int \Bigg[\sum_{j=1}^{d} (\alpha_j - \alpha_j^*)x_j + \sum_{\substack{j<k \\ j,k\in\mathcal{C}_0}} (\alpha_{jk} - \alpha_{jk}^*)x_j x_k\Bigg]^2 \prod_{l\in\mathcal{C}} w_l(x_l)\,dx_l \prod_{l\notin\mathcal{C}} P_{X_l}(dx_l) \\
&\quad + c_2 \sum_{k=1}^{p} \int [\tilde{\mathbf{x}}_k^\top(\mathbf{f}_k(x_{r+k}) - \mathbf{f}_k^*(x_{r+k}))]^2 \prod_{l\in\mathcal{C}} w_l(x_l)\,dx_l \prod_{l\notin\mathcal{C}} P_{X_l}(dx_l) \\
&\geq c_3|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*|_*^2 \\
&\quad + c_3 \sum_{k=1}^{p} \int [\tilde{\mathbf{x}}_k^\top(\mathbf{f}_k(x_{r+k}) - \mathbf{f}_k^*(x_{r+k}))]^2 P_{X_{r+k}}(dx_{r+k}) \prod_{l\in I_k^*}^{D} P_{X_l}(dx_l)
\end{aligned}$$

for some constants $c_3 > 0$ and where I_k^* denotes the set of indices of $\tilde{\mathbf{x}}_k$. The second inequality of (A.1) now follows because the smallest eigenvalues of $\int \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^\top \prod_{l\in I_k^*}^{D} P_{X_l}(dx_l)$ can be bounded from below by a positive constant times the smallest eigenvalue of $E[\tilde{\mathbf{X}}_k \tilde{\mathbf{X}}_k^\top] = \int \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^\top P_{\tilde{\mathbf{X}}_k}(d\tilde{\mathbf{x}}_k)$, where $P_{\tilde{\mathbf{X}}_k}$ denotes the distribution of $\tilde{\mathbf{X}}_k$. These eigenvalues can be bounded away from zero by assumption (A0). □

A.2. Proof of Corollaries 1 and 2. For the proof of these two corollaries, we apply Theorem 1. We have to show that (2.4) holds with $\kappa_n = n^{l/(2l+1)}$ for the penalized least squares estimator and the spline sieve estimator, respectively.

For the proof of Corollary 1, we apply Theorem 10.2 in van de Geer (2000). As discussed in van de Geer (2000) the statement of the theorem remains valid for errors with subexponential tails if the entropy bounds hold for entropies with bracketing. For the application of this theorem one needs results on the entropy with bracketing for the class of functions m that fulfill (1.4) with f_{jk} in Sobolev classes. Because g has an absolutely bounded derivative, Lemma 1 implies that the well-known entropy conditions for Sobolev classes carry over to the classes of functions m. This proves Corollary 1.

For the proof of Corollary 2 we use Theorem 1 in Chen and Shen (1998). Compare also Theorem 10.11 in van de Geer (2000). Using the above entropy bound one can easily verify Conditions A.1–A.4 in Chen and Shen (1998) with $l(\theta,(\mathbf{X},Y))=(m(\mathbf{X})-Y)^2$, $\theta=(\boldsymbol{\alpha},f_{jk};1\le j\le d,k\in I_j)$ and m as given at (A.2). Note that $l(\theta,(\mathbf{X},Y))-l(\theta_0,(\mathbf{X},Y))=(m(\mathbf{X})-m_0(\mathbf{X}))^2-2(m(\mathbf{X})-m_0(\mathbf{X}))\varepsilon$, where $\theta_0=(\boldsymbol{\alpha}_0,f_{0jk};1\le j\le d,k\in I_j)$ is the true tuple, m_0 denotes the true underlying regression function and $\varepsilon=Y-m_0(\mathbf{X})$. To check the conditions compare also the proof of Proposition 1 in the latter paper. In particular, their condition A.4 holds with $s=2l/(2l+1)$. This follows because for two functions $g_1,g_2:[0,1]\to\mathbb{R}$ with $|D_z^l g_1(z)|\le L$, $|D_z^l g_2(z)|\le L$ and $\int_0^1(g_1(z)-g_2(z))^2\,dz\le\delta^2$, it holds that $|g_1(z)-g_2(z)|\le 2(2L)^{1-c}\delta^{1-c}$ with $c=2l(2l+1)^{-1}$; see Lemma 2 in Chen and Shen (1998). The necessary conditions are simplified because we assume that the data are i.i.d.; see also Remark 1(b) in Chen and Shen (1998). To get $\varepsilon_n^{(2-s)/(\gamma-1)}B_n\ge 1$ at the end of their proof of Theorem 3 one needs that $(2-s)/(\gamma-1)<s$ which is equivalent to $\gamma>2+l^{-1}$. One can check that their proof goes through with this constraint. Thus it suffices for the i.i.d. case that $E|\varepsilon|^\gamma<\infty$ holds for some $\gamma>2+l^{-1}$.

A.3. Additional assumptions for kernel smoothing. We assume the density of $\mathbf{X}^c$ is supported on $[0,1]^p$. Thus, the integration at (3.6) is over $[0,1]$. We note that, for the normalized kernel $K_{h_j}(u_j,z_j)$ introduced in Section 3, $[2h_j,1-2h_j]$ is the interior region for z_j that does not have a boundary effect. In addition to assumption (A0) in Section 2, we collect the conditions we use for the theory in Sections 3 and 5.

(A1) The quasi-likelihood function $Q(\mu,y)$ is three times continuously differentiable with respect to μ for each y in its range, $Q_2(u,y)<0$ for $u\in\mathbb{R}$ and y in its range, the link function g is three times continuously differentiable, V is twice continuously differentiable and the conditional variance function $v(\mathbf{x})=\mathrm{var}(Y|\mathbf{X}=\mathbf{x})$ is continuous in $\mathbf{x}^c=(x_{r+1},\dots,x_{r+p})^\top$ for each $(x_1,\dots,x_r)$. The densities p_{X_j} for $r+1\le j\le r+p$ are bounded away from zero on $[0,1]$. The function V and the derivative g' are bounded away from zero. The higher-order derivatives g'' and g''' are bounded. The weight function w is continuously differentiable and fulfills $w(0)=w(1)=0$.

(A2) The partial derivatives $\partial p_{\mathbf{X}}(\mathbf{x})/\partial\mathbf{x}^c$ of the joint density function $p_{\mathbf{X}}$ exist and are continuous in $\mathbf{x}^c$ for all $(x_1,\dots,x_r)$.

(A3) The components of $\mathbf{f}_j$ are twice continuously differentiable.

(A4) $E|Y|^\alpha<\infty$ for some $\alpha>5/2$.

(A5) The kernel function K is bounded, symmetric about zero, has compact support, say $[-1,1]$, and is Lipschitz continuous. The bandwidths h_j

depend on the sample size n and satisfy $n^{1/5}h_j \to c_j$ as $n \to \infty$ for some constants $0 < c_j < \infty$.

A.4. Preliminaries for the proofs of theorems 2–4. The population versions of $\hat{\mathbf{F}}_j$ are defined by

$$\mathbf{F}_j(\boldsymbol{\eta})(z_j) = \int E[Q_1(\tilde{\mathbf{X}}^\top \boldsymbol{\eta}(\mathbf{z}), m(\mathbf{X}))\tilde{\mathbf{X}}_j | \mathbf{X}^c = \mathbf{z}] p_{\mathbf{X}^c}(\mathbf{z})\, d\mathbf{z}_{-j}.$$

For the empirical versions of $\mathbf{W}_{jk}(\mathbf{z};\boldsymbol{\eta})$, $\mathbf{W}_j(\mathbf{z};\boldsymbol{\eta})$ and $\mathbf{W}(\mathbf{z};\boldsymbol{\eta})$ introduced in Section 3, we define

$$\hat{\mathbf{W}}_{jk}(\mathbf{z};\boldsymbol{\eta}) = -n^{-1}\sum_{i=1}^n Q_2(\tilde{\mathbf{X}}^{i\top}\boldsymbol{\eta}(\mathbf{z}), Y^i)\tilde{\mathbf{X}}_j^i \tilde{\mathbf{X}}_k^{i\top} K_{\mathbf{h}}(\mathbf{X}^{c,i}, \mathbf{z}),$$

and then define $\hat{\mathbf{W}}_j(\mathbf{z};\boldsymbol{\eta})$ and $\hat{\mathbf{W}}(\mathbf{z};\boldsymbol{\eta})$ in the same way as we define $\mathbf{W}_j(\mathbf{z};\boldsymbol{\eta})$ and $\mathbf{W}(\mathbf{z};\boldsymbol{\eta})$, respectively. We write $\hat{\mathbf{W}}_{jk}(\mathbf{z}) = \hat{\mathbf{W}}_{jk}(\mathbf{z};\mathbf{f})$ in case the true $\mathbf{f}$ enters into the place of $\boldsymbol{\eta}$.

For a tuple $\boldsymbol{\delta} \in \mathcal{H}(\mathbf{W})$, let $\hat{\mathbf{F}}_j'(\boldsymbol{\eta})(\boldsymbol{\delta})$ denote the Fréchet differential of $\hat{\mathbf{F}}_j$ at $\boldsymbol{\eta}$ to the direction of $\boldsymbol{\delta}$. Then

$$\hat{\mathbf{F}}_j'(\boldsymbol{\eta})(\boldsymbol{\delta})(z_j) = -\int \hat{\mathbf{W}}_j(\mathbf{z};\boldsymbol{\eta})\boldsymbol{\delta}(\mathbf{z})\, d\mathbf{z}_{-j}.$$

The second term on the right-hand side of (4.3) is simply $\hat{\mathbf{F}}_j'(\hat{\mathbf{f}}^{[s-1]})(\hat{\mathbf{f}} - \hat{\mathbf{f}}^{[s-1]})(z_j)$. The population versions of $\hat{\mathbf{F}}_j'(\boldsymbol{\eta})$ are defined by $\mathbf{F}_j'(\boldsymbol{\eta})(\boldsymbol{\delta})(z_j) = -\int \mathbf{W}_j(\mathbf{z};\boldsymbol{\eta})\boldsymbol{\delta}(\mathbf{z})\, d\mathbf{z}_{-j}$. Define a linear operator $\hat{\mathbf{F}}'(\boldsymbol{\eta})$ by

$$\hat{\mathbf{F}}'(\boldsymbol{\eta})(\boldsymbol{\delta}) = ((\hat{\mathbf{F}}_1'(\boldsymbol{\eta})(\boldsymbol{\delta}))^\top, \ldots, (\hat{\mathbf{F}}_p'(\boldsymbol{\eta})(\boldsymbol{\delta}))^\top)^\top.$$

Likewise, define $\mathbf{F}'(\boldsymbol{\eta})$ from $\mathbf{F}_j'(\boldsymbol{\eta})$. In the proofs below, we use $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_p^\top)^\top$ to denote the true vector of univariate functions.

A.5. Proof of Theorem 2. In addition to $\|\cdot\|_\#$ introduced in Section 3, we consider two other norms. Let $\|\cdot\|_2$ be the $L_2(p_{\mathbf{X}^c})$-norm defined by $\|\boldsymbol{\eta}\|_2^2 = \int |\boldsymbol{\eta}(\mathbf{z})|^2 p_{\mathbf{X}^c}(\mathbf{z})\, d\mathbf{z}$. Define $\|\boldsymbol{\eta}\|_\infty = \max\{\sup_{2h_1 \le z_1 \le 1-2h_1} |\boldsymbol{\eta}_1(z_1)|, \ldots, \sup_{2h_p \le z_p \le 1-2h_p} |\boldsymbol{\eta}_p(z_p)|\}$, where $|\cdot|$ denotes the Euclidean norm. As in Section 3, we write $\mathbf{W}(\mathbf{z}) = \mathbf{W}(\mathbf{z};\mathbf{f})$, $\mathbf{W}_{jj}(z_j) = \mathbf{W}_{jj}(z_j;\mathbf{f})$, etc., for the true tuple $\mathbf{f}$. For a linear operator $\mathcal{F}$ that maps $\mathcal{H}(\mathbf{W})$ to $\mathcal{H}(\mathbf{W})$, let $\|\mathcal{F}\|_{\mathrm{op}}$ denote its operator-norm defined by $\|\mathcal{F}\|_{\mathrm{op}} = \sup_{\|\boldsymbol{\delta}\|=1} \|\mathcal{F}(\boldsymbol{\delta})\|$. Here and below, if not specified, $\|\cdot\|$ is either $\|\cdot\|_2$ or $\|\cdot\|_\infty$. We prove

$$P(\hat{\mathbf{F}}'(\mathbf{f}) \text{ is invertible and } \|\hat{\mathbf{F}}'(\mathbf{f})^{-1}\|_{\mathrm{op}} \le C_1) \to 1, \tag{A.3}$$

$$P(\|\hat{\mathbf{F}}'(\boldsymbol{\eta}) - \hat{\mathbf{F}}'(\boldsymbol{\eta}')\|_{\mathrm{op}} \le C_2\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \text{ for all } \boldsymbol{\eta}, \boldsymbol{\eta}' \in B_r(\mathbf{f})) \to 1, \tag{A.4}$$

for some constants $r, C_1, C_2 > 0$, where $B_r(\mathbf{f})$ is a ball centered on $\mathbf{f}$ with radius r. Then, the theorem follows from Newton–Kantorovich theorem [see, e.g., Deimling (1985)] since $\|\hat{\mathbf{F}}(\mathbf{f})\|_2 = O_p(n^{-2/5})$ and $\|\hat{\mathbf{F}}(\mathbf{f})\|_\infty = O_p(n^{-2/5}\sqrt{\log n})$.

By the standard techniques of kernel smoothing, one can show that, uniformly for $\mathbf{z} \in [0,1]^p$, $\hat{\mathbf{W}}_{jk}(z_j, z_k) \equiv \int \hat{\mathbf{W}}_{jk}(\mathbf{z})\, d\mathbf{z}_{-(j,k)}$ converges to $\mathbf{W}_{jk}(z_j, z_k)$ and $\hat{\mathbf{W}}_{jj}(z_j) \equiv \int \hat{\mathbf{W}}_{jj}(\mathbf{z})\, d\mathbf{z}_{-j}$ to $\mathbf{W}_{jj}(z_j)$, for $1 \le j \ne k \le p$. This gives the uniform convergence of $\hat{\mathbf{F}}'(\mathbf{f})(\boldsymbol{\delta})$ to $\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta})$ over $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\| \le R$, where $R > 0$ is an arbitrary positive real number. Thus, to prove (A.3) it suffices to show that $\mathbf{F}'(\mathbf{f})$ is invertible and has a bounded inverse. For this claim we first show that the map $\mathbf{F}'(\mathbf{f}) : \mathcal{H}(\mathbf{W}) \to \mathcal{H}(\mathbf{W})$ is one-to-one. Suppose that $\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta}) = \mathbf{0}$ for some $\boldsymbol{\delta} \in \mathcal{H}(\mathbf{W})$. We have to show that $\boldsymbol{\delta} = \mathbf{0}$. From $\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta}) = \mathbf{0}$ we get that $\boldsymbol{\delta}_j(z_j)^\top \int \mathbf{W}_j(\mathbf{z})\boldsymbol{\delta}(\mathbf{z})\, d\mathbf{z}_{-j} = 0$ for all $1 \le j \le p$. This implies

$$\begin{aligned} 0 &= \sum_{j=1}^p \int \left[\int \boldsymbol{\delta}_j(z_j)^\top \int \mathbf{W}_j(\mathbf{z})\boldsymbol{\delta}(\mathbf{z})\, d\mathbf{z}_{-j} \right] dz_j \\ &\ge \sum_{j=1}^p \sum_{k=1}^p \int (\tilde{\mathbf{x}}_j^\top \boldsymbol{\delta}_j(x_{r+j}))(\tilde{\mathbf{x}}_k^\top \boldsymbol{\delta}_k(x_{r+k})) P_{\mathbf{X}}(d\mathbf{x}) \\ &= c \int \left(\sum_{j=1}^p \tilde{\mathbf{x}}_j^\top \boldsymbol{\delta}_j(x_{r+j}) \right)^2 P_{\mathbf{X}}(d\mathbf{x}) \end{aligned}$$

for some positive constant $c > 0$. Here, for the inequality we used that $V(u)g'(u)^2$ is bounded from above for u in any compact set. Applying the arguments in the proof of Lemma 1, we get that $\boldsymbol{\delta} = \mathbf{0}$ a.s. This proves the claim that the map $\mathbf{F}'(\mathbf{f}) : \mathcal{H}(\mathbf{W}) \to \mathcal{H}(\mathbf{W})$ is one-to-one.

Next, using the fact that $\langle \mathbf{F}'(\mathbf{f})(\boldsymbol{\delta}), \boldsymbol{\eta} \rangle_\# = \langle \boldsymbol{\delta}, \mathbf{F}'(\mathbf{f})(\boldsymbol{\eta}) \rangle_\#$ for all $\boldsymbol{\delta}, \boldsymbol{\eta} \in \mathcal{H}(\mathbf{W})$, one can show that $\mathbf{F}'(\mathbf{f})$ is onto. Thus, $\mathbf{F}'(\mathbf{f})$ is invertible. To verify that $\mathbf{F}'(\mathbf{f})^{-1}$ is bounded, it suffices to prove that the bijective linear operator $\mathbf{F}'(\mathbf{f})$ is bounded, due to the bounded inverse theorem. From the assumption (A1) and the fact that support of the density of $\mathbf{X}$ is bounded, we get

$$\begin{aligned} &\|\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta})\|_\#^2 = \int |\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta})(\mathbf{z})|^2 p_{\mathbf{X}^c}(\mathbf{z})\, d\mathbf{z} \le C_3 \|\boldsymbol{\delta}\|_2^2, \\ &\|\mathbf{F}'(\mathbf{f})(\boldsymbol{\delta})\|_\infty \le C_4 \|\boldsymbol{\delta}\|_\infty \end{aligned}$$

for some constants $C_3, C_4 > 0$. This concludes that $\mathbf{F}'(\mathbf{f})$ is bounded in both of the norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$.

The claim (A.4) holds since, for any given $r > 0$, there exists a constant $C_5 > 0$ such that, with probability tending to one, $\|\hat{\mathbf{F}}'(\boldsymbol{\eta})(\boldsymbol{\delta}) - \hat{\mathbf{F}}'(\boldsymbol{\eta}')(\boldsymbol{\delta})\| \le C_5 \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \cdot \|\boldsymbol{\delta}\|$ for all $\boldsymbol{\eta}, \boldsymbol{\eta}' \in B_r(\mathbf{f})$.

A.6. Proof of Theorem 3. Let $\check{\boldsymbol{\delta}}$ denote a solution of the following equations:

$$\begin{aligned}\check{\boldsymbol{\delta}}_j(z_j) &= \tilde{\boldsymbol{\delta}}_j(z_j) \\ &\quad - \sum_{k\neq j} \int \hat{\mathbf{W}}_{jj}(z_j)^{-1}\hat{\mathbf{W}}_{jk}(z_j,z_k)\check{\boldsymbol{\delta}}_k(z_k)\,dz_k, \qquad 1\le j\le p,\end{aligned} \tag{A.5}$$

where $\tilde{\boldsymbol{\delta}}_j(z_j) = \hat{\mathbf{W}}_{jj}(z_j)^{-1}\hat{\mathbf{F}}_j(\mathbf{f})(z_j)$. We first remark that $\check{\boldsymbol{\delta}}$ exists and is unique, with probability tending to one. Define $\tilde{\mathbf{f}} = \mathbf{f} + \check{\boldsymbol{\delta}}$. We claim

$$\|\tilde{\mathbf{f}} - \mathbf{f}\|_\infty = O_p(n^{-2/5}\sqrt{\log n}). \tag{A.6}$$

Define $\tilde{\mathbf{F}}_j$ for $1\le j\le p$ by

$$\begin{aligned}\tilde{\mathbf{F}}_j(\boldsymbol{\eta})(z_j) &= \hat{\mathbf{F}}_j(\mathbf{f})(z_j) \\ &\quad + \sum_{k=1}^p \int n^{-1}\sum_{i=1}^n Q_2(\tilde{\mathbf{X}}^{i\top}\mathbf{f}(\mathbf{z}),Y^i) \\ &\qquad\qquad \times \tilde{\mathbf{X}}_j^i\tilde{\mathbf{X}}_k^{i\top}[\boldsymbol{\eta}_k(z_k)-\mathbf{f}_k(z_k)]K_{\mathbf{h}}(\mathbf{X}^{c,i},\mathbf{z})\,d\mathbf{z}_{-j}.\end{aligned}$$

Note that $\tilde{\mathbf{f}}$ is the solution of the system of equations $\tilde{\mathbf{F}}_j(\boldsymbol{\eta}) = \mathbf{0}_j, 1\le j\le p$ by the definitions of $\check{\boldsymbol{\delta}}_j$ and $\tilde{\boldsymbol{\delta}}_j$. Thus, the claim (A.6) ensures that $\hat{\mathbf{F}}_j(\tilde{\mathbf{f}})(z_j) = \tilde{\mathbf{F}}_j(\tilde{\mathbf{f}})(z_j) + o_p(n^{-2/5}) = o_p(n^{-2/5})$, uniformly for $z_j\in[2h_j,1-2h_j]$. Also, (A.6) and (A.4) give (A.3) with $\mathbf{f}$ being replaced by $\tilde{\mathbf{f}}$. This establishes $\|\hat{\mathbf{f}}-\tilde{\mathbf{f}}\|_\infty = o_p(n^{-2/5})$. The theorem follows since

$$n^{2/5}\check{\boldsymbol{\delta}}(\mathbf{z}) \xrightarrow{d} N((\boldsymbol{\beta}_1(z_1)^\top,\dots,\boldsymbol{\beta}_p(z_p)^\top)^\top, \operatorname{diag}(\boldsymbol{\Sigma}_j(z_j))),$$

the latter being proved similarly as in the proof of Theorem 2 of Lee, Mammen and Park (2012).

It remains to prove (A.6). The Fréchet differential of $\tilde{\mathbf{F}}_j$ at $\boldsymbol{\eta}$ to the direction $\boldsymbol{\delta}$, which we denote by $\tilde{\mathbf{F}}'_j(\boldsymbol{\eta})(\boldsymbol{\delta})$, does not depend on $\boldsymbol{\eta}$ since $\tilde{\mathbf{F}}_j(\boldsymbol{\eta})$ is linear in $\boldsymbol{\eta}$. In fact $\tilde{\mathbf{F}}'_j(\boldsymbol{\eta})(\boldsymbol{\delta}) = \hat{\mathbf{F}}'_j(\mathbf{f})(\boldsymbol{\delta})$ for all $\boldsymbol{\eta}$. This means that $\tilde{\mathbf{F}}(\mathbf{f}) = \hat{\mathbf{F}}(\mathbf{f})$ and $\tilde{\mathbf{F}}'(\boldsymbol{\eta}) = \hat{\mathbf{F}}'(\mathbf{f})$ for all $\boldsymbol{\eta}$, so that (A.3) and (A.4) are valid for $\tilde{\mathbf{F}}$. As in the proof of Theorem 2, this implies (A.6).

A.7. Proof of Theorem 4. An application of Newton–Kantorovich theorem gives the first part of the theorem. For the proof of the second part of the theorem, we rewrite a full cycle of the iteration step in (4.5) as $\hat{\boldsymbol{\delta}}^{[s,\ell]} = \hat{\boldsymbol{\delta}}_+^{[s-1]} + \hat{A}^{[s-1]}\hat{\boldsymbol{\delta}}^{[s,\ell-1]}$ with appropriate definitions of $\hat{\boldsymbol{\delta}}_+^{[s-1]}$ and $\hat{A}^{[s-1]}$. Note that $\hat{\boldsymbol{\delta}}_+^{[s-1]}$ differs from the tuple with elements $\tilde{\boldsymbol{\delta}}_j^{[s-1]}$. Also, we can write a full cycle of the iteration step for solving (A.5) as $\check{\boldsymbol{\delta}}^{[\ell]} = \check{\boldsymbol{\delta}}_+ + \check{A}\check{\boldsymbol{\delta}}^{[\ell-1]}$ with appropriate definitions of $\check{\boldsymbol{\delta}}_+$ and $\check{A}$. Finally, we can write $\boldsymbol{\delta}^{[\ell]} = \boldsymbol{\delta}_+ + A\boldsymbol{\delta}^{[\ell-1]}$

with appropriate definitions of $\boldsymbol{\delta}_+$ and A for a full cycle of

$$\boldsymbol{\delta}_j^{[\ell]}(z_j) = \mathbf{W}_{jj}(z_j)^{-1}\mathbf{F}_j(\mathbf{f})(z_j) - \sum_{k\neq j}\int [\mathbf{W}_{jj}(z_j)]^{-1}\mathbf{W}_{jk}(z_j,z_k)\boldsymbol{\delta}_k^{[\ell-1]}(z_k)\,dz_k,$$

$$1 \le j \le p.$$

For the convergence of the last iteration, we note that the projection operators $\pi_{kj}:\mathcal{H}_k(\mathbf{W}) \to \mathcal{H}_j(\mathbf{W})$ for all $1 \le j \neq k \le p$ are Hilbert–Schmidt, where $\mathcal{H}_k(\mathbf{W})$ is a subspace of $\mathcal{H}(\mathbf{W})$ such that $\boldsymbol{\eta} \in \mathcal{H}_k(\mathbf{W})$ if and only if $\boldsymbol{\eta}_l = \mathbf{0}$ for all $l \neq k$, and elements of $\boldsymbol{\eta}_k$ with the configuration at (3.4) satisfy the constraints (3.5). This implies there exist constants C_0 and $0 < \rho_0 < 1$, with

$$\int [\boldsymbol{\delta}^{[\ell]}(\mathbf{z}) - \boldsymbol{\delta}^{[\infty]}(\mathbf{z})]^\top \mathbf{W}(\mathbf{z})[\boldsymbol{\delta}^{[\ell]}(\mathbf{z}) - \boldsymbol{\delta}^{[\infty]}(\mathbf{z})]\,d\mathbf{z} \le C_0\rho_0^{2\ell} \tag{A.7}$$

for some limiting function $\boldsymbol{\delta}^{[\infty]}$.

We now apply $\|\check{A} - A\| = o_p(1)$, $\|\check{\boldsymbol{\delta}}_+ - \boldsymbol{\delta}_+\| = o_p(1)$ and $\sup_s \|\hat{A}^{[s]} - \check{A}\| \le c$, $\sup_s \|\hat{\boldsymbol{\delta}}_+^{[s]} - \check{\boldsymbol{\delta}}_+\| \le c$ with probability tending to one, for some constant $c > 0$ that can be made as small as we like by choosing τ small enough. This and equation (A.7) implies that for some constants C_* and $0 < \rho_* < 1$,

$$\int [\hat{\boldsymbol{\delta}}^{[s,\ell]}(\mathbf{z}) - \hat{\boldsymbol{\delta}}^{[s,\infty]}(\mathbf{z})]^\top \mathbf{W}(\mathbf{z})[\hat{\boldsymbol{\delta}}^{[s,\ell]}(\mathbf{z}) - \hat{\boldsymbol{\delta}}^{[s,\infty]}(\mathbf{z})]\,d\mathbf{z} \le C_*\rho_*^{2\ell},$$

with probability tending to one. This completes the proof of Theorem 4.

REFERENCES

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95** 941–956. MR1804449

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* **66** 289–314. MR1612238

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88** 298–308. MR1212492

Deimling, K. (1985). *Nonlinear Functional Analysis*. Springer, Berlin. MR0787404

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 57–80. MR1959093

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* **27** 715–731. MR1804172

Fengler, M., Härdle, W. and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics* **5** 189–218.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

Heim, S., Fahrmeir, L., Eilers, P. H. C. and Marx, B. D. (2007). 3D space-varying coefficient models with application to diffusion tensor imaging. *Comput. Statist. Data Anal.* **51** 6212–6228. MR2407709

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

Kauermann, G. and Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika* **86** 119–128. MR1688076

Lee, Y. K., Mammen, E. and Park, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458

Lee, Y. K., Mammen, E. and Park, B. U. (2012). Projection-type estimation for varying coefficient regression models. *Bernoulli* **18** 177–205. MR2888703

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

Mammen, E. and Nielsen, J. P. (2003). Generalised structured models. *Biometrika* **90** 551–566. MR2006834

Mammen, E. and Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33** 1260–1294. MR2195635

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035. MR1447739

Noh, H. S. and Park, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statist. Sinica* **20** 1183–1202. MR2730179

Park, B. U., Mammen, E., Härdle, W. and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *J. Amer. Statist. Assoc.* **104** 284–298. MR2504378

Roca-Pardiñas, J. and Sperlich, S. (2010). Feasible estimation in generalized structured models. *Stat. Comput.* **20** 367–379. MR2725394

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

Xue, L. and Yang, L. (2006). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16** 1423–1446. MR2327498

Yang, L., Park, B. U., Xue, L. and Härdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration. *J. Amer. Statist. Assoc.* **101** 1212–1227. MR2328308

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