Bernoulli **24**(2), 2018, 1233–1265
DOI: 10.3150/16-BEJ898

Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions

KYUNGHEE HAN[1], HANS-GEORG MÜLLER[2] and BYEONG U. PARK[1,*]

[1] *Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea. E-mail:* [*] *bupark@stats.snu.ac.kr*
[2] *Department of Statistics, University of California, 1 Shields Avenue, Davis, CA 95616, USA*

We study smooth backfitting when there are errors-in-variables, which is motivated by functional additive models for a functional regression model with a scalar response and multiple functional predictors that are additive in the functional principal components of the predictor processes. The development of a new smooth backfitting technique for the estimation of the additive component functions in functional additive models with multiple functional predictors requires to address the difficulty that the eigenfunctions and therefore the functional principal components of the predictor processes, which are the arguments of the proposed additive model, are unknown and need to be estimated from the data. The available estimated functional principal components contain an error that is small for large samples but nevertheless affects the estimation of the additive component functions. This error-in-variables situation requires to develop new asymptotic theory for smooth backfitting. Our analysis also pertains to general situations where one encounters errors in the predictors for an additive model, when the errors become smaller asymptotically. We also study the finite sample properties of the proposed method for the application in functional additive regression through a simulation study and a real data example.

Keywords: errors in predictors; functional additive model; functional data analysis; functional principal component; kernel smoothing; smooth backfitting

1. Introduction

There is currently no theory available for smooth backfitting with errors-in-variables for the additive predictors, and we develop such theory in this paper. The need for this is demonstrated in an example from functional regression that is emphasized throughout the paper. Models that pair functional predictors with a scalar response are commonly encountered and constitute one of the central modeling problems in functional data analysis (FDA) (Ramsay and Silverman [20]), motivated by applied problems where one wishes to predict a scalar outcome from observed functional predictors or to model the nature of the relationship. Specifically, we consider the problem where one has d random predictor functions $X_1, \dots, X_d$ defined on intervals $\mathcal{I}_1, \dots, \mathcal{I}_d$, respectively, that are coupled with a continuous scalar response Y. Our goal is to model and implement the regression $E(Y|X_1, \dots, X_d)$.

1350-7265 © 2018 ISI/BS

Many approaches have been developed for the special case $d = 1$. Writing $X_1 = X$ and $E(X) = \mu$, highly structured approaches include the well-established functional linear model (FLM), where one assumes

$$E(Y|X) = \mu_0 + \int \beta(s)\bigl(X(s) - \mu(s)\bigr)\,ds \tag{1}$$

(Cardot *et al.* [2], Bosq [1]) for a smooth fixed parameter function β. This model is a direct extension of the classical linear regression model with multiple predictors. While this model is linear in the predictor process X, extensions to nonlinear cases include the functional quadratic (and polynomial) model (FQM)

$$\begin{aligned} E(Y|X) = \mu_0 &+ \int \beta(s)\bigl(X(s) - \mu(s)\bigr)\,ds \\ &+ \int \gamma(s,t)\bigl(X(s) - \mu(s)\bigr)\bigl(X(t) - \mu(t)\bigr)\,ds\,dt \end{aligned} \tag{2}$$

(Yao and Müller [22]), which contains a second parameter function γ that communicates the quadratic and interaction effects that are part of this model.

These models can be directly represented in terms of functional principal components, as follows. For each predictor function X_j, assumed to be fully observed and to be recorded without noise, let ϕ_{jk}, $1 \le k < \infty$, be the set of the orthonormal eigenfunctions of the integral operator having the auto-covariance surface $C_j(u,v) = E[X_j(u) - \mu_j(u)][X_j(v) - \mu_j(v)]$ as its kernel. We assume that ϕ_{jk} are ordered in terms of the respective eigenvalues $\lambda_{j1} \ge \lambda_{j2} \ge \cdots$. Each set of eigenfunctions $(\phi_{jk} : 1 \le k < \infty)$ forms a basis for $L_2(\mathcal{I}_j)$, $1 \le j \le d$. Let ξ_{jk} denote the functional principal components (FPC) of X_j defined by $\xi_{jk} = \int (X_j(t) - \mu_j(t))\phi_{jk}(t)\,dt$, where $\mu_j(t) = EX_j(t)$. Then it is straightforward to see that the FLM can be written as

$$E(Y|X_1) = \mu_0 + \sum_k \beta_k \xi_{1k} \tag{3}$$

in terms of the FPCs with coefficients β_k, while the FQM can be written as

$$E(Y|X_1) = \mu_0 + \sum_k \beta_k \xi_{1k} + \sum_{k,l} \gamma_{kl} \xi_{1k} \xi_{1l}, \tag{4}$$

with coefficients β_k and γ_{kl}. When implementing these models, only a finite number of predictor scores are used, based on a truncated eigenfunction expansion. Using these representations, these two models can be straightforwardly extended to the case with multiple predictor functions, simply by collecting the FPCs from all predictor functions and then applying a multiple linear or quadratic model to the combined scores.

The perspective of representing the above models in terms of the FPCs suggests a model that is additive in the FPCs,

$$E(Y|X_1) = \mu_0 + \sum_k f_k(\xi_{1k}), \tag{5}$$

where the additive component functions f_k are supposed to be smooth and are required to satisfy $E(f_k(\xi_{1k})) = 0$. This model has been referred to as the Functional Additive Model (FAM) [19]. Our goal is to extend the additive model of Stone [21] to functional regression and in particular, to provide rigorous theoretical justification for this extension. An interesting feature of additive functional regression models is that in case of independent predictor FPCs, as in the case of a Gaussian predictor process, the functions f_k can be consistently estimated by marginal regressions in model (5), that is, consistent estimates of the component functions f_k can be obtained by regressing Y against each of the predictors ξ_{1k} separately, which can be easily done by componentwise nonparametric regression. However, for the case of multivariate predictors that we consider here this simple device is not possible anymore and the marginal regression approach will be biased, due to the dependencies among the predictor components.

Other approaches along these lines include additive regression models for longitudinal data [3] and a "time-additive regression" approach that stands in contrast to the "frequency-additive" regression models described above that are additive in the FPCs, while the time-additive models are additive in time and have been referred to as "continuous additive model" [17,18]. Another related class of models are single index [11,15] and multiple index models that have been considered for single predictor functions [4,8] as well as multiple predictor functions [10]. Additive models with functional predictors and generalized responses have been considered in [7] and with penalized least squares in B spline and reproducing kernel Hilbert space settings in [5,25].

In this paper, we develop the theory for smooth backfitting to fit a functional regression model that is additive in the FPCs of the predictor processes, which are used for dimension reduction of each of the predictor processes X_j. Specifically, we study the *additive functional score model* (AFSM)

$$E(Y|X_1,\dots,X_d) = \sum_{j=1}^{d}\sum_{k} f_{jk}(\xi_{jk}), \tag{6}$$

where f_{jk} are unknown univariate functions. In the case of more than one predictor function, the FPC scores of the different predictors will in general be correlated. This means that the marginal regression approach employed in the Functional Additive Model (FAM) to estimate the various additive component functions is bound to fail, and needs to be replaced with a more complex backfitting method in order to obtain consistent estimates of the component functions in (6). For this, we apply the smooth backfitting technique of [16] for the estimation of the component functions f_{jk} in the AFSM. In the case of real-valued predictors, the smooth backfitting method is known to provide a powerful technique for the estimation of component functions in various structured nonparametric models. Recent work for the non-functional application and implementation of smooth backfitting includes [13,14,23] and [24]. Smooth backfitting avoids the curse of dimensionality and is known to achieve the optimal univariate estimation error rate in multivariate regression.

In the application of the smooth backfitting technique to estimate the additive component functions f_{jk} in the AFSM, we replace the unobserved FPC scores ξ_{jk} by the FPCs that are obtained from the spectral decomposition of the sample covariance surfaces, which is the usual practice in functional data analysis. This is the key feature which motivates the development of a theory of smooth backfitting with errors-in-variables, as existing theory on smooth backfitting is focused on real-valued predictors under the assumption that the predictors are fully observed without error. To address this issue, we develop here an innovative extension of the theory of smooth backfitting methods to cover this case. While we illustrate the extension in the context of functional additive regression, it is also of interest for other applications where one has errors in the predictors that can be assumed to be asymptotically small. To obtain a complete asymptotic theory for this case requires to go deep inside the operation of smooth backfitting.

A related paper is Hildebrandt, Bissantz and Dette [9], where it is assumed that the signal (the regression function) is contaminated (convoluted) by some known function and the goal is to recover the signal. This model differs substantially from the case considered here, where the predictors (FPC scores) are contaminated by estimation errors that are asymptotically small but where the contamination distribution is of unknown nature and therefore the convoluting function is not known. Since the errors vanish asymptotically in our application to functional data, they do not affect the asymptotics of the resulting regression estimators and therefore asymptotically we do not face a deconvolution problem. Another difference lies in the method of estimating additive regression models, where Hildebrandt, Bissantz and Dette [9] adopt the marginal integration technique, while we develop our approach for the smooth backfitting approach. The latter has emerged as a powerful technique that gives reliable estimators for various structural nonparametric regression problems (see, e.g., Yu, Park and Mammen [23], Lee, Mammen and Park [13,14] and Zhang, Park and Wang [24]).

The purpose of this paper is to develop smooth backfitting methodology for the case of asymptotically small errors in the predictors, as motivated by additive functional regression. We present rigorous asymptotic theory and our analysis is complemented by simulations and a data application. The proposed smooth backfitting approach is introduced in the next sections, followed by asymptotic theory in Section 3. Simulation results are reported in Section 4 and an application to bike usage data is in Section 5. This is followed by additional technical details and proofs in Section 6.

2. Methodology

2.1. Range of estimation

We begin by describing briefly the smooth backfitting method in the case where the predictors are real-valued and fully observed. In the classical nonparametric additive regression model, $E(Y|\mathbf{Z}=\mathbf{z}) = f_1(z_1) + \cdots + f_d(z_d)$, where Y is the response variable and $\mathbf{Z} = (Z_1, \ldots, Z_d)$ is the predictor vector having a density that is supported on a bounded set, say $[0, 1]^d$, the component functions f_j are not identifiable, unless one adds a constraint such as $Ef_j(Z_j) = 0$. Then one may rewrite the model as $E(Y|\mathbf{Z}=\mathbf{z}) = f_0 + f_1(z_1) + \cdots + f_d(z_d)$.

Under this model and the constraints, it holds that

$$f_j(z_j) = E(Y|Z_j = z_j) - f_0 - \sum_{k\neq j} \int_0^1 f_k(z_k) \frac{p_{Z_j,Z_k}(z_j, z_k)}{p_{Z_j}(z_j)} \, dz_k, \qquad 1 \le j \le d, \tag{7}$$

and $f_0 = E(Y)$, where p_{Z_j} and p_{Z_j,Z_k} denote the density functions of Z_j and (Z_j, Z_k), respectively. The smooth backfitting estimator $(\hat{f}_j : 0 \le j \le d)$ of the tuple $(f_j : 0 \le j \le d)$ is defined to be the solution of the system of integral equations

$$\hat{f}_j(z_j) = \tilde{f}_j(z_j) - \hat{f}_0 - \sum_{k \neq j} \int_0^1 \hat{f}_k(z_k) \frac{\hat{p}_{Z_j,Z_k}(z_j, z_k)}{\hat{p}_{Z_j}(z_j)} \, dz_k, \qquad 1 \le j \le d$$

subject to $\int_0^1 \hat{f}_j(z_j)\hat{p}_{Z_j}(z_j)\,dx_j = 0$ for $1 \le j \le d$, where $\hat{p}_{Z_j}$ and $\hat{p}_{Z_j,Z_k}$ are estimators of p_{Z_j} and p_{Z_j,Z_k}, respectively, and $\tilde{f}_j$ is an estimator of the marginal regression function $E(Y|Z_j = z_j)$.

For identifiability of the component functions f_{jk} in the AFSM (6) we also need to invoke a constraint for each and may rewrite the model as

$$E(Y|X_1, \ldots, X_d) = f_0 + \sum_{j=1}^{d} \sum_{k=1}^{L_j} f_{jk}(\xi_{jk}) \tag{8}$$

with the constant term f_0 depending the constraints. Here and in the following, we assume that it suffices to approximate the functional predictors X_j by their first L_j functional principal components in the AFSM (6), where the truncation points L_j are tuning parameters. To obtain an analogue of equation (7) for estimating f_{jk} in the AFSM one may consider simply replacing Z_j by the FPC scores ξ_{jk}, f_j by f_{jk}, and the integration over the interval [0, 1] by integrating over the whole real line. One would then need to estimate the conditional means $E(Y|\xi_{jk} = \cdot)$ and the densities of ξ_{jk} and $(\xi_{jk}, \xi_{j'k'})$ on $\mathbb{R}$ or $\mathbb{R}^2$. This is however not feasible since the collection of observed data is bounded. The usual practice in nonparametric regression analysis is to consider a bounded region for the estimation of the regression function, and we adopt this approach by aiming to estimate f_{jk} on bounded intervals.

2.2. Constraints for component functions

We note that we may not use the constraint $Ef_{jk}(\xi_{jk}) = 0$ for the true component function f_{jk} as in smooth backfitting for bounded real-valued predictors, since then the corresponding constraint for the estimator of f_{jk} requires the estimation of f_{jk} on the entire support of the density of ξ_{jk}, which in general has an unbounded support as the ξ_{jk} in general are unbounded random variables. To describe the constraints that we employ instead, let I_{jk} be the bounded intervals on which we estimate f_{jk}, and define $I = I_{11} \times \cdots \times I_{dL_d}$. Let p denote the density function of $\boldsymbol{\xi} = (\xi_{jk} : 1 \le j \le d, 1 \le k \le L_j)$, p_j the density function of $\boldsymbol{\xi}_j = (\xi_{j1}, \ldots, \xi_{jL_j})$, and define

$$p_0^I = \int_I p(\mathbf{u})\,d\mathbf{u}, \qquad p_{jk}^I(u_{jk}) = \int_{I_{-jk}} p(\mathbf{u})\,d\mathbf{u}_{-jk}/p_0^I,$$

where $\mathbf{u}_{-jk}$ for a vector $\mathbf{u} = (u_{jk} : 1 \le j \le d, 1 \le k \le L_j)$ is the resulting vector one obtains from $\mathbf{u}$ after deleting u_{jk}, and $I_{-jk} = \prod_{(j',k') \neq (j,k)} I_{j'k'}$. We then adopt the constraints

$$\int_{I_{jk}} f_{jk}(u) p_{jk}^I(u)\,du = 0, \qquad 1 \le j \le d, 1 \le k \le L_j, \tag{9}$$

which imply $f_0 = \int_I E(Y|\boldsymbol{\xi} = \mathbf{u}) p(\mathbf{u})\, d\mathbf{u}/p_0^I$.

One may prefer to employ other constraints, such as $\int_{I_{jk}} f_{jk}(u) w_{jk}(u)\, du = 0$, instead of (9), for some known weight functions w_{jk}. We choose the constraint (9) since it is natural and gives simpler forms for f_0 and its estimator. The method and theory that we develop here can be easily modified if one uses different constraints.

2.3. Smooth backfitting for the additive functional score model

Suppose that we observe $(X_1^i, \dots, X_d^i, Y^i)$, which are independent across $1 \le i \le n$ and follow model (8). Aiming at estimators for the component functions f_{jk}, define

$$p_{jk,j'k'}^I(u_{jk}, u_{j'k'}) = \int_{I_{-jk,j'k'}} p(\mathbf{u})\, d\mathbf{u}_{-jk,j'k'}/p_0^I,$$

where $\mathbf{u}_{-jk,j'k'}$ is the resulting vector one obtains from $\mathbf{u}$ after deleting u_{jk} and $u_{j'k'}$, and $I_{-jk,j'k'} = \prod_{(j'',k'') \neq (j,k),(j'k')} I_{j''k''}$. Multiplying both sides of (8) by the joint density p and then integrating them over the rectangle I_{-jk} gives the following system of integral equations:

$$\begin{aligned} f_{jk}(u_{jk}) = {} & \frac{1}{p_0^I \cdot p_{jk}^I(u_{jk})} \int_{I_{-jk}} E(Y|\boldsymbol{\xi} = \mathbf{u}) p(\mathbf{u})\, d\mathbf{u}_{-jk} - f_0 \\ & - \sum_{(j',k') \neq (j,k)} \int_{I_{j'k'}} f_{j'k'}(u_{j'k'}) \frac{p_{jk,j'k'}^I(u_{jk}, u_{j'k'})}{p_{jk}^I(u_{jk})}\, du_{j'k'}, \qquad (10) \\ & 1 \le j \le d, 1 \le k \le L_j. \end{aligned}$$

It can be shown that this system of integral equations also follows when minimizing $E[Y - g_0 - g_{11}(\xi_{11}) - \cdots - g_{dL_d}(\xi_{dL_d})]^2 I(\boldsymbol{\xi} \in I)$ over a constant g_0 and univariate functions g_{jk}.

Let ξ_{jk}^i be the FPC score of the jth predictor X_j for the ith subject, that is, $\xi_{jk}^i = \int_{\mathcal{I}_j} (X_j^i(t) - \mu_j(t)) \phi_{jk}(t)\, dt$. We estimate ξ_{jk}^i from the standard eigenanalysis of the estimated auto-covariance surface $\hat{C}_j$, which is defined by

$$\hat{C}_j(s,t) = n^{-1} \sum_{i=1}^n \bigl[X_j^i(s) - \hat{\mu}_j(s)\bigr]\bigl[X_j^i(t) - \hat{\mu}_j(t)\bigr]$$

with $\hat{\mu}_j(t) = n^{-1} \sum_{i=1}^n X_j^i(t)$. Specifically, let $(\hat{\phi}_{jk} : 1 \le k < \infty)$ be the orthonormal eigenfunctions in the spectral decomposition of $\hat{C}_j$, ordered in terms of the respective eigenvalues $\hat{\lambda}_{j1} \ge \hat{\lambda}_{j2} \ge \cdots$. The estimators of ξ_{jk}^i are then obtained by approximating the defining integrals, that is,

$$\hat{\xi}_{jk}^i = \int_{\mathcal{I}_j} \bigl(X_j^i(t) - \hat{\mu}_j(t)\bigr) \hat{\phi}_{jk}(t)\, dt.$$

To consider the estimation of the integral equation (10), we note that p^I_{jk} and $p^I_{jk,j'k'}$ are conditional densities of ξ_{jk} and $(\xi_{jk}, \xi_{j'k'})$, respectively, given that the event $\boldsymbol{\xi} \in I$ occurs. This suggests the following kernel estimators of p^I_{jk} and $p^I_{jk,j'k'}$:

$$\hat{p}^I_{jk}(u) = n^{-1} \sum_{i=1}^{n} K_{h_{jk}}\big(u, \hat{\xi}^i_{jk}\big)\mathbb{I}\big(\hat{\boldsymbol{\xi}}^i \in I\big)/\hat{p}^I_0,$$

$$\hat{p}^I_{jk,j'k'}(u, v) = n^{-1} \sum_{i=1}^{n} K_{h_{jk}}\big(u, \hat{\xi}^i_{jk}\big)K_{h_{j'k'}}\big(v, \hat{\xi}^i_{j'k'}\big)\mathbb{I}\big(\hat{\boldsymbol{\xi}}^i \in I\big)/\hat{p}^I_0,$$

where $\hat{p}^I_0 = n^{-1}\sum_{i=1}^n \mathbb{I}(\hat{\boldsymbol{\xi}}^i \in I)$ and $\mathbb{I}$ is the indicator. The kernel function $K_{h_{jk}}(u, v)$ with a bandwidth h_{jk} and a baseline kernel K is defined by

$$K_{h_{jk}}(u, v) = \mathbb{I}(u \in I_{jk}) \frac{K_{h_{jk}}(u - v)}{\int_{I_{jk}} K_{h_{jk}}(t - v)\,dt} \tag{11}$$

whenever $\int_{I_{jk}} K_{h_{jk}}(t - v)\,dt \neq 0$, and we set $K_{h_{jk}}(u, v) = 0$, otherwise. Here, $K_{h_{jk}}(u - v) = h^{-1}_{jk} K(h^{-1}_{jk}(u - v))$. With these definitions, it follows that

$$\int_{I_{jk}} \hat{p}^I_{jk}(u)\,du = 1, \qquad \int_{I_{j'k'}} \hat{p}^I_{jk,j'k'}(u, v)\,dv = \hat{p}^I_{jk}(u).$$

We also estimate f_0 by $\hat{f}_0 = n^{-1}\sum_{i=1}^n Y^i \mathbb{I}(\hat{\boldsymbol{\xi}}^i \in I)/\hat{p}^I_0$, and the first term on the right-hand side of (10) by

$$\tilde{f}_{jk}(u) = \left[n^{-1} \sum_{i=1}^{n} K_{h_{jk}}\big(u, \hat{\xi}^i_{jk}\big)\mathbb{I}\big(\hat{\boldsymbol{\xi}}^i \in I\big)\right]^{-1} n^{-1} \sum_{i=1}^{n} Y^i K_{h_{jk}}\big(u, \hat{\xi}^i_{jk}\big)\mathbb{I}\big(\hat{\boldsymbol{\xi}}^i \in I\big). \tag{12}$$

The system of smooth backfitting equations for the tuple $(\hat{f}_{jk} : 1 \le j \le d, 1 \le k \le L_j)$ is then given by

$$\hat{f}_{jk}(u) = \tilde{f}_{jk}(u) - \hat{f}_0 - \sum_{(j',k') \neq (j,k)} \int_{I_{j'k'}} \hat{f}_{j'k'}(v) \frac{\hat{p}^I_{jk,j'k'}(u, v)}{\hat{p}^I_{jk}(u)}\,dv, \tag{13}$$
$$1 \le k \le L_j; 1 \le j \le d,$$

with the constraints for $\hat{f}_{jk}$ that

$$\int_{I_{jk}} \hat{f}_{jk}(u)\hat{p}^I_{jk}(u)\,du = 0, \qquad 1 \le k \le L_j, 1 \le j \le d. \tag{14}$$

The solution of equation (13) is obtained by an iteration. With an initial tuple $(\hat{f}_{jk}^{[0]} : 1 \le j \le d, 1 \le k \le L_j)$, the updating formula for $(\hat{f}_{jk}^{[r]} : 1 \le j \le d, 1 \le k \le L_j)$ in the rth cycle of the backfitting iteration is given by

$$\begin{aligned}\hat{f}_{jk}^{[r]}(u) = \tilde{f}_{jk}(u) - \hat{f}_0 &- \sum_{j'=1}^{j-1}\sum_{k'=1}^{L_{j'}}\int_{I_{j'k'}} \hat{f}_{j'k'}^{[r]}(v)\frac{\hat{p}_{jk,j'k'}^{I}(u,v)}{\hat{p}_{jk}^{I}(u)}\,dv \\ &- \sum_{k'=1}^{k-1}\int_{I_{jk'}} \hat{f}_{jk'}^{[r]}(v)\frac{\hat{p}_{jk,jk'}^{I}(u,v)}{\hat{p}_{jk}^{I}(u)}\,dv \\ &- \sum_{k'=k+1}^{L_j}\int_{I_{jk'}} \hat{f}_{jk'}^{[r-1]}(v)\frac{\hat{p}_{jk,jk'}^{I}(u,v)}{\hat{p}_{jk}^{I}(u)}\,dv \\ &- \sum_{j'=j+1}^{d}\sum_{k'=1}^{L_{j'}}\int_{I_{j'k'}} \hat{f}_{j'k'}^{[r-1]}(v)\frac{\hat{p}_{jk,j'k'}^{I}(u,v)}{\hat{p}_{jk}^{I}(u)}\,dv.\end{aligned} \tag{15}$$

We remark that in the estimation of the component functions we use only $(\hat{\boldsymbol{\xi}}^i, Y^i)$ with $\hat{\boldsymbol{\xi}}^i \in I$. One may think of using the full data $(\hat{\boldsymbol{\xi}}^i, Y^i)$, $1 \le i \le n$, to avoid boundary effect near the end points of the set I. This turns out to be not relevant, however. The reason is that the smooth backfitting technique depends on the range of estimation, I, via the integration of the components f_{jk} on I_{jk}. Consequently, the integral equations at (10) involve p_0^I, p_{jk}^I and $p_{jk,j'k'}^I$, instead of the corresponding $p_0 = 1$, p_{jk} and $p_{jk,j'k'}$, the latter two being the densities of ξ_{jk} and $(\xi_{jk}, \xi_{j'k'})$, respectively, and it is not appropriate to estimate these quantities with the full data. If one is interested in estimating the component functions on a compact set I, and also wants to avoid boundary effect when estimating near ∂I, then one may apply the smooth backfitting method that we describe above, to a compact set which is slightly larger than I, and then take the function estimates only on the set I.

3. Theoretical properties

We assume without loss of generality that $I_{jk} = [0, 1]$ for all $1 \le j \le d$, $1 \le k \le L_j$, given the L_j. As already mentioned, we do not assume that ξ_{jk} across k are independent, neither that the predictors X_j are independent.

3.1. Convergence rates of FPC estimators

Here, we derive a uniform (over $1 \le i \le n$) rate of convergence of $\hat{\xi}_{jk}^i$ under a moment condition on the predictor processes X_j. The uniform convergence rate will be used frequently in our theoretical development for the proposed smooth backfitting estimators. Recall that λ_{jk}, $k \ge 1$, denote the ordered eigenvalues in the spectral decomposition of the jth covariance surface

$C_j(s,t) = \mathrm{Cov}(X_j(s), X_j(t))$. In the following, we denote the L_2-norm by $\|\cdot\|$, for spaces of square integrable functions that can have one or two arguments.

We assume:

(A1) For each j, the eigenvalues λ_{jk} for different k are separated and $E\|X_j\|^{2\beta} < \infty$ for some $\beta \geq 2$.

Let $\hat{\mathcal{C}}_j$ denote the covariance operator associated with the covariance surface $\hat{C}_j$, which maps $f \in L_2(\mathcal{I}_j)$ to $g = \hat{\mathcal{C}}_j f \in L_2(\mathcal{I}_j)$ by

$$g(s) = \int_{\mathcal{I}_j} \hat{C}_j(s,t) f(t)\, dt.$$

According to Lemma 4.3 of [1],

$$\|\hat{\phi}_{jk} - \phi_{jk}\| \leq \frac{2\sqrt{2}}{\delta_{jk}} \|\hat{\mathcal{C}}_j - \mathcal{C}_j\|, \qquad 1 \leq k \leq L_j, \tag{16}$$

where $\delta_{jk} = \min_{1\leq \ell \leq k}(\lambda_{j\ell} - \lambda_{j,\ell+1})$, so that

$$E\|\hat{\mathcal{C}}_j - \mathcal{C}_j\|^2 = \int E\big[\hat{C}_j(s,t) - C_j(s,t)\big]^2\, ds\, dt = O\big(n^{-1}\big), \tag{17}$$

whence (16) and (17) imply that $\|\hat{\phi}_{jk} - \phi_{jk}\| = O_p(n^{-1/2})$ for $1 \leq k \leq L_j$. The approximation error of the estimated FPC scores $\hat{\xi}^i_{jk}$ may be then obtained from this. A simple application of Markov inequality gives $\max_{1\leq i\leq n} \|X^i_j - \mu_j\|^2 = O_p(n^{1/\beta})$, where β is the constant in the condition (A1). This with the fact that $\|\hat{\mu}_j - \mu_j\| = O_p(n^{-1/2})$ entails the following proposition, which will be crucial for replacing estimated by true predictor scores when deriving the asymptotic properties of the proposed smooth backfitting method.

Proposition 1. *Under the condition* (A1) *it holds that*

$$\max_{1\leq i\leq n} \big|\hat{\xi}^i_{jk} - \xi^i_{jk}\big| = O_p\big(n^{-(\beta-1)/2\beta}\big), \qquad 1 \leq j \leq d, 1 \leq k \leq L_j. \tag{18}$$

3.2. Theory for smooth backfitting with errors-in-variables

Our first theorem demonstrates that, with probability tending to one, when available predictors $\hat{\xi}^i_{jk}$ satisfy property (18), then the backfitting equation (13) has a unique solution and the iterative algorithm (15) converges to the solution at a geometric rate. We note that these results are independent of our application to functional additive regression and pertain to smooth backfitting in general contexts where one has errors-in-variables. Collecting here the assumptions we use to establish the convergence of the smooth backfitting algorithm, conditions (A2)–(A6) below are typical for nonparametric additive modeling.

(A2) The baseline kernel function K is bounded, has compact support $[-1, 1]$, is symmetric about zero, differentiable and its derivative is Lipschitz continuous.

(A3) The bandwidths h_{jk} satisfy $n^{1/5}h_{jk} \to c_{jk}$ for some positive constants c_{jk}.

(A4) The joint density p of $\boldsymbol{\xi}$ is bounded away from zero and infinity on I.

(A5) $E|Y|^{\alpha} < \infty$ for $\alpha > 5/2$ and $\mathrm{Var}(Y|\xi_{jk} = \cdot, \boldsymbol{\xi} \in I)$ are continuous on $[0, 1]$.

(A6) The component functions f_{jk} are twice continuously differentiable and the densities p_{jk} and $p_{jk,j'k'}$ are (partially) continuously differentiable on [0,1].

Theorem 1. *Assume the conditions* (A1)–(A6), *or alternatively*, (18) *for* $\beta \ge 2$ *and* (A2)–(A6). *Then*, *the following statements hold*: (i) *with probability tending to one*, *there exists a unique solution* $(\hat{f}_{jk} : 1 \le j \le d, 1 \le k \le L_j)$ *of* (13) *subject to the constraints* (14); (ii) *there exists a constant* $0 < \gamma < 1$ *and* $c > 0$ *such that with probability tending to one*

$$\int_0^1 \bigl[\hat{f}_{jk}^{[r]}(u) - \hat{f}_{jk}(u)\bigr]^2 p_{jk}^I(u)\,du \le c \cdot \gamma^{2r}\left(1 + \sum_{j,k}\int_0^1 \hat{f}_{jk}^{[0]}(u)^2 p_{jk}^I(u)\,du\right).$$

The alternative version of this result applies to general errors-in-variables situations where the errors in the predictors decrease with increasing sample size, and therefore is of interest independently of the application to functional regression. Our second theorem gives the rates of convergence and the asymptotic distributions of the estimators of the component functions f_{jk}. Here, we make the moment condition on X_j in (A1) a bit stronger for the effect of estimating ξ_{jk}^i to be negligible in the estimation of the component functions.

(A1$'$) For each j, the eigenvalues λ_{jk} for different k are separated and $E\|X_j\|^{2\beta} < \infty$ for some $\beta > 5$.

For the statement of the theorem, let $p_{jk}^{(1)}(\mathbf{u}) = \partial p(\mathbf{u})/\partial u_{jk}$. Define

$$\tilde{\beta}_{jk}(u) = \int v^2 K(v)\,dv \sum_{j',k'} c_{j'k'}^2 E\left(f'_{j'k'}(\xi_{j'k'})\frac{p_{j'k'}^{(1)}(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\middle|\xi_{jk} = u, \boldsymbol{\xi} \in I\right),$$

$$\sigma_{jk}^2(u) = \frac{1}{p_0^I p_{jk}^I(u)} c_{jk}^{-1}\,\mathrm{Var}(Y|\xi_{jk} = u, \boldsymbol{\xi} \in I)\int K(v)^2\,dv,$$

$$\beta_{jk}(u) = \beta_{jk}^*(u) + \frac{1}{2}c_{jk}^2 f''_{jk}(u)\int u^2 K(u)\,du,$$

where the constants c_{jk} are defined in the condition (A3) above. Let the tuple $(\beta_{jk}^* : 1 \le j \le d, 1 \le k \le L_j)$ be the solution of the system of integral equations

$$\beta_{jk}^*(u) = \tilde{\beta}_{jk}(u) - \sum_{(j',k') \ne (j,k)} \int_{I_{j'k'}} \beta_{j'k'}^*(v)\frac{p_{jk,j'k'}^I(u, v)}{p_{jk}^I(u)}\,dv, \qquad 1 \le j \le d, 1 \le k \le L_j,$$

subject to the constraints

$$\int_0^1 \beta_{jk}^*(u) p_{jk}^I(u)\,du = \mu_2 c_{jk}^2 \int_0^1 f_{jk}'(u) \frac{\partial}{\partial u} p_{jk}^I(u)\,du.$$

Theorem 2. *Assume the conditions* (A1′) *and* (A2)–(A6), *or alternatively*, (18) *for* $\beta > 5$ *and* (A2)–(A6). *Then it follows that for a given vector* $(\mathbf{u} : 0 < u_{jk} < 1, 1 \le j \le d, 1 \le k \le L_j)$, *the estimators* $\hat{f}_{jk}(u_{jk})$ *for different pairs* (j,k) *are asymptotically independent and*

$$n^{2/5}\big(\hat{f}_{jk}(u_{jk}) - f_{jk}(u_{jk})\big) \xrightarrow{d} N\big(\beta_{jk}(u_{jk}), \sigma_{jk}^2(u_{jk})\big).$$

Furthermore, $\|\hat{f}_{jk} - f_{jk}\| = O_p(n^{-2/5})$, $\sup_{u\in[2h_{jk},1-2h_{jk}]} |\hat{f}_{jk}(u) - f_{jk}(u)| = O_p(n^{-2/5}\sqrt{\log n})$ *and* $\sup_{u\in[0,1]} |\hat{f}_{jk}(u) - f_{jk}(u)| = O_p(n^{-1/5})$.

Again, the alternative version includes a result that is of primary interest for additive modeling with smooth backfitting when one has a general error-in-variables situation that is not necessarily related to functional regression. The results of the above theorem also hold for the theoretical estimators of f_{jk}, denoted by $\hat{f}_{jk}^*$, that use the true ξ_{jk}^i instead of the estimated $\hat{\xi}_{jk}^i$, which is seen by a straightforward extension of the standard theory of smooth backfitting.

The proof of the above theorem is based on comparisons of the $\hat{f}_{jk}$ with their theoretical versions $\hat{f}_{jk}^*$. Let $\hat{f}_+ = \sum_{j=1}^d \sum_{k=1}^{L_j} \hat{f}_{jk}$ and $\hat{f}_+^* = \sum_{j=1}^d \sum_{k=1}^{L_j} \hat{f}_{jk}^*$. Then, the additive functions $\hat{f}_+$ and $\hat{f}_+^*$, respectively, are the solutions of the equations

$$\hat{f}_+ = \tilde{f}_\oplus + \hat{T}\hat{f}_+, \qquad \hat{f}_+^* = \tilde{f}_\oplus^* + \hat{T}^*\hat{f}_+^*$$

for appropriately defined additive functions $\tilde{f}_\oplus$ and $\tilde{f}_\oplus^*$, and for appropriately defined linear operators $\hat{T}$ and $\hat{T}^*$, see Section 6.2 for the explicit forms of these functions and operators.

The additive functions $\tilde{f}_\oplus$ and $\tilde{f}_\oplus^*$, as well as $\hat{T}$ and $\hat{T}^*$, differ only in that the former are based on the estimated FPCs, $\hat{\xi}_{jk}^i$, while the latter are based on the true FPCs ξ_{jk}^i. In the proof of Theorem 1 in Section 6, we show that $\|\hat{T} - \hat{T}^*\| = o_p(1)$ and $\|\hat{T}\| \le 1 - \delta$ with probability tending to one for a small constant $\delta > 0$. From this, we can argue that $\hat{f}_+ - \hat{f}_+^*$ is of the same magnitude as $\tilde{f}_\oplus - \tilde{f}_\oplus^*$. To establish that the estimation of the FPC scores has a negligible effect on the first-order properties of $\hat{f}_+$, we need $\hat{f}_+ - \hat{f}_+^*$ to be of an order smaller than $n^{-2/5}$. One might want to prove this by establishing that the $\tilde{f}_{jk}$, of which $\tilde{f}_\oplus$ is composed, differ from the corresponding $\tilde{f}_{jk}^*$ by an order smaller than $n^{-2/5}$. But it turns out that this is not the case. In fact, $\max_{1\le i\le n} |\hat{\xi}_{jk}^i - \xi_{jk}^i| = O_p(n^{-(\beta-1)/2\beta})$ as demonstrated in Proposition 1, which is inflated in $\tilde{f}_{kJ} - \tilde{f}_{jk}^*$ by a factor of the inverse of the bandwidth size, $n^{1/5}$, so that one can only have $\tilde{f}_{jk} - \tilde{f}_{jk}^* = O_p(n^{-(3\beta-5)/10\beta})$. Note that $(3\beta - 5)/10\beta < 3/10 < 2/5$ for all $\beta > 0$. Thus, the proof of Theorem 2 requires a careful asymptotic analysis, deep inside the operation of the smooth backfitting technique. This led us to develop an innovative way of understanding the theory of smooth backfitting methods, which may be also useful for other related problems.

Additional overview on the main steps of the proof which provide further insights how the technical challenges provided by the presence of errors can be overcome and the detailed steps of the proof can be found in Section 6.3.

4. Finite sample performance

In this section, we demonstrate the finite sample performance of the proposed additive functional regression with smooth backfitting. In a simulation setting, we generated a pair of random functions $\mathbf{X} = (X_1, X_2)$ such that

$$X_j(t) = \mu_j(t) + \xi_{j1}\phi_{j1}(t) + \xi_{j2}\phi_{j2}(t), \qquad t \in [0,1],$$

for $j = 1, 2$, where the mean functions μ_j of X_j are given by

$$\mu_1(t) = 2t + 2\cos(3\pi t), \qquad \mu_2(t) = 3t + \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{t-0.6}{\sigma}\right)^2\right\}, \qquad \sigma = 0.1.$$

We chose the normalized Fourier basis $\phi_{11}(t) = \sqrt{2}\sin(2\pi t)$, $\phi_{12}(t) = \sqrt{2}\cos(2\pi t)$, $\phi_{21}(t) = \sqrt{2}\sin(4\pi t)$ and $\phi_{22}(t) = \sqrt{2}\cos(4\pi t)$ on the interval $[0,1]$ and the FPC score vectors $\boldsymbol{\xi}_j = (\xi_{j1}, \xi_{j2})$ to have multivariate normal distributions with mean zero and $\mathrm{var}(\boldsymbol{\xi}_1) = \mathrm{diag}(2.5, 1.5)$, $\mathrm{var}(\boldsymbol{\xi}_2) = \mathrm{diag}(2.2, 1.2)$ and

$$\begin{pmatrix} \mathrm{cov}(\xi_{11}, \xi_{21}) & \mathrm{cov}(\xi_{11}, \xi_{22}) \\ \mathrm{cov}(\xi_{12}, \xi_{21}) & \mathrm{cov}(\xi_{12}, \xi_{22}) \end{pmatrix} = \begin{pmatrix} 0.7 & -0.4 \\ 0.3 & -0.5 \end{pmatrix}.$$

Then the two predictor functions X_1 and X_2 are correlated, as the cross-covariance of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ is nonzero, where the correlation between $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ is given by

$$\begin{pmatrix} \mathrm{corr}(\xi_{11}, \xi_{21}) & \mathrm{corr}(\xi_{11}, \xi_{22}) \\ \mathrm{corr}(\xi_{12}, \xi_{21}) & \mathrm{corr}(\xi_{12}, \xi_{22}) \end{pmatrix} \approx \begin{pmatrix} 0.298 & -0.231 \\ 0.165 & -0.373 \end{pmatrix}.$$

The scalar response Y was generated by $Y = \sum_{j=1}^2 \sum_{k=1}^2 g_{jk}(\xi_{jk}) + \varepsilon$, where $g_{11}(x) = x$, $g_{12}(x) = 2\sin(\pi x)$, $g_{21}(x) = 0.5(x^2 - 4)\cos(\pi x)$, $g_{22}(x) = \frac{1}{3}(x^2 - 4)x^3$ and errors $\varepsilon \sim N(0, 0.1^2)$ were independent of $\boldsymbol{\xi}$. We target the centered component functions $f_{jk} = g_{jk} - \int_{I_{jk}} g_{jk} p_{jk}^I$ satisfying the constraints (9) in the estimation. For the domains I_{jk} in the estimation, we took $I_{jk} = [-2, 2]$. Under this data generating scheme, we obtained $B = 400$ Monte Carlo samples of sizes $n = 100, 200, 400$ and 1000.

Let $\mathcal{X}_n$ denote a generated sample $\{(X_1^i, X_2^i, Y^i) : 1 \le i \le n\}$. For the eigen-analysis of the sample covariance function $\hat{C}_j$, we adopted a standard discretization method [12,20], choosing a dense grid $\{(t_l, t_{l'}) : 0 \le l, l' \le M\}$ on $[0,1]^2$, with equi-spaced $0 = t_0 < t_1 < \cdots < t_M = 1$, calculating for each sample properly normalized eigenvectors $(\hat{\phi}_{jk}(t_l) : 1 \le l \le M)$ of the discretized sample covariance matrices $(\hat{C}_j(t_l, t_{l'}) : 1 \le l, l' \le M)$, then obtaining the estimated FPC scores $\hat{\xi}_{jk}^i = \sum_{l=1}^M (X_j^i(t_l) - \hat{\mu}_j(t_l))\hat{\phi}_{jk}(t_l)(t_l - t_{l-1})$. To determine the number of included components,

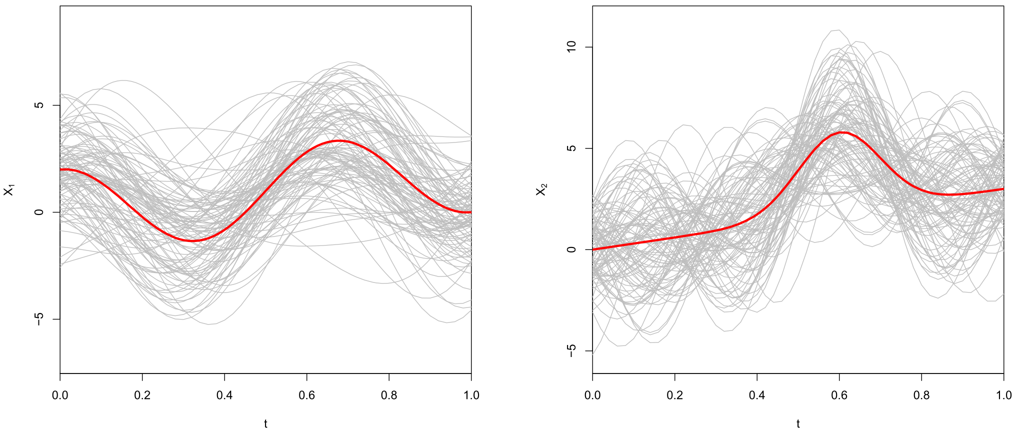

Figure 1. Illustration of a simulation sample. Left panel: Mean function μ_1 (red) and 200 sample trajectories of X_1 (gray) for first predictor functions X_{1i}. Right panel: Same for the second predictor functions X_{2i}.

we employed a "fraction of variance explained" (FVE) criterion, choosing the first L_j eigenfunctions that explained at least 90% of the variation of the sample predictors $\{X_j^i : 1 \le i \le n\}$. In all cases, this criterion led to the choice $L_1 = L_2 = 2$ for all generated Monte Carlo samples.

Figure 1 depicts one randomly chosen sample of the two predictor processes for sample size $n = 200$. We found that covariance estimates, estimated eigenfunctions and estimated functional principal components (FPCs) were close to their targets.

In the practical implementation of our method, one needs to determine $\sum_{j=1}^{d} L_j$ bandwidths for each updating iteration of the smooth backfitting algorithm. Direct application of cross validation (CV) is highly time-consuming and therefore not feasible. We propose an alternative efficient "bandwidth shrinkage scheme" that is based on K-fold CV as follows. Let $(J_\ell : 1 \le \ell \le K)$ be a partition of the index set $\{1, \dots, n\}$ such that $\bigcup_{\ell=1}^{K} J_\ell = \{1, \dots, n\}$ and $J_\ell \cap J_{\ell'} = \varnothing$ for $\ell \ne \ell'$. Let $\mathcal{X}_n^{(\ell)}$ denote the sub-sample corresponding to J_ℓ. Then, (i) compute a baseline bandwidth vector $\mathbf{h}^{(\ell)} = \{h_{jk}^{(\ell)} : 1 \le k \le L_j, 1 \le j \le d\}$ for each ℓ using the sub-sample $\mathcal{X}_n \setminus \mathcal{X}_n^{(\ell)}$; (ii) calculate

$$\mathrm{CV}_\ell(\alpha) = \sum_{i \in J_\ell} \big(Y^i - \hat{f}_{-\ell}(\hat{\boldsymbol{\xi}}^i; \alpha \mathbf{h}^{(\ell)})\big)^2 \mathbb{I}(\hat{\boldsymbol{\xi}}^i \in I)$$

for $\alpha > 0$, where $\hat{f}_{-\ell}(\mathbf{u}; \alpha \mathbf{h}^{(\ell)}) = \hat{f}_{0,-\ell} + \sum_{j=1}^{d} \sum_{k=1}^{L_j} \hat{f}_{jk,-\ell}(u_{jk}; \alpha h_{jk}^{(\ell)})$ is the estimated additive function based on the sub-sample $\mathcal{X}_n \setminus \mathcal{X}_n^{(\ell)}$; (iii) find

$$\hat{\alpha} = \arg\min_{\alpha > 0} \sum_{\ell=1}^{K} \mathrm{CV}_\ell(\alpha)$$

and choose $\hat{\mathbf{h}} = \hat{\alpha}\mathbf{h}^{\text{full}}$, where $\mathbf{h}^{\text{full}}$ is a bandwidth vector obtained by the same method as $\mathbf{h}^{(\ell)}$ but from the entire sample $\mathcal{X}_n$. We call $\hat{\alpha}$ the global bandwidth shrinkage factor. In our simulation, $K = 5$ and the baseline bandwidth vectors $\mathbf{h}^{(\ell)}$ and $\mathbf{h}^{\text{full}}$ were obtained by leave-one-out CV being employed to marginal regression. For example, $h_{jk}^{\text{full}} = \arg\min_{h>0} \sum_{i=1}^{n} (Y^i - \tilde{f}_{jk,-i}(\hat{\xi}_{jk}^{i}; h))^2$, where $\tilde{f}_{jk,-i}$ is the leave-ith-out version of the marginal estimator $\tilde{f}_{jk}$ defined at (12).

To assess estimation performance, we computed the Monte Carlo approximation of the mean integrated squared error (MISE):

$$\operatorname{MISE}(f_{jk}) \approx \frac{1}{B} \sum_{b=1}^{B} \frac{1}{4} \int_{I_{jk}} \bigl(\hat{f}_{jk}^{(b)}(u) - f_{jk}(u)\bigr)^2 \, du,$$

where $\hat{f}_{jk}^{(b)}$ is the estimate from the bth Monte Carlo sample and we divide the integrated value by 4 to normalize the integral over the interval $I_{jk} = [-2, 2]$. The results are summarized in Table 1 and Figure 2, the latter exhibiting the bias and variance of the component function estimators $\hat{f}_{jk}$. These results suggest that the bandwidth shrinkage scheme works well, as the global bandwidth shrinkage factor adjusts for the overall smoothness level of the estimated component functions $\hat{f}_{jk}$, while the baseline bandwidths adjust for the smoothness of the individual component functions. This bandwidth shrinkage scheme also gave reliable results for the auxiliary density estimation, demonstrated in Figure 3.

We also compared the prediction performance of the proposed ASFM with that of a naive application of the functional additive model (FAM) based on marginal regression [19], and also with the functional linear model (FLM) and functional quadratic model (FQM). For these latter

Table 1. MISE of $\hat{f}_{jk}$ based on $B = 400$ Monte Carlo samples. The number in parenthesis is the estimated standard error of the Monte Carlo approximation of MISE

	With baseline bandwidths h_{jk}^{full}				With shrinkage bandwidths $\hat{h}_{jk}$			
Sample size (n)	$j = 1$		$j = 2$		$j = 1$		$j = 2$	
	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$
100	0.181 (0.009)	0.890 (0.038)	0.565 (0.019)	0.425 (0.021)	0.172 (0.008)	0.900 (0.038)	0.561 (0.018)	0.413 (0.019)
200	0.082 (0.004)	0.412 (0.020)	0.258 (0.010)	0.236 (0.016)	0.075 (0.004)	0.419 (0.020)	0.254 (0.010)	0.234 (0.015)
400	0.027 (0.002)	0.176 (0.008)	0.121 (0.005)	0.104 (0.004)	0.023 (0.001)	0.174 (0.008)	0.112 (0.004)	0.104 (0.003)
1000	0.014 (0.001)	0.078 (0.004)	0.058 (0.003)	0.049 (0.002)	0.008 (0.000)	0.066 (0.003)	0.044 (0.002)	0.043 (0.001)

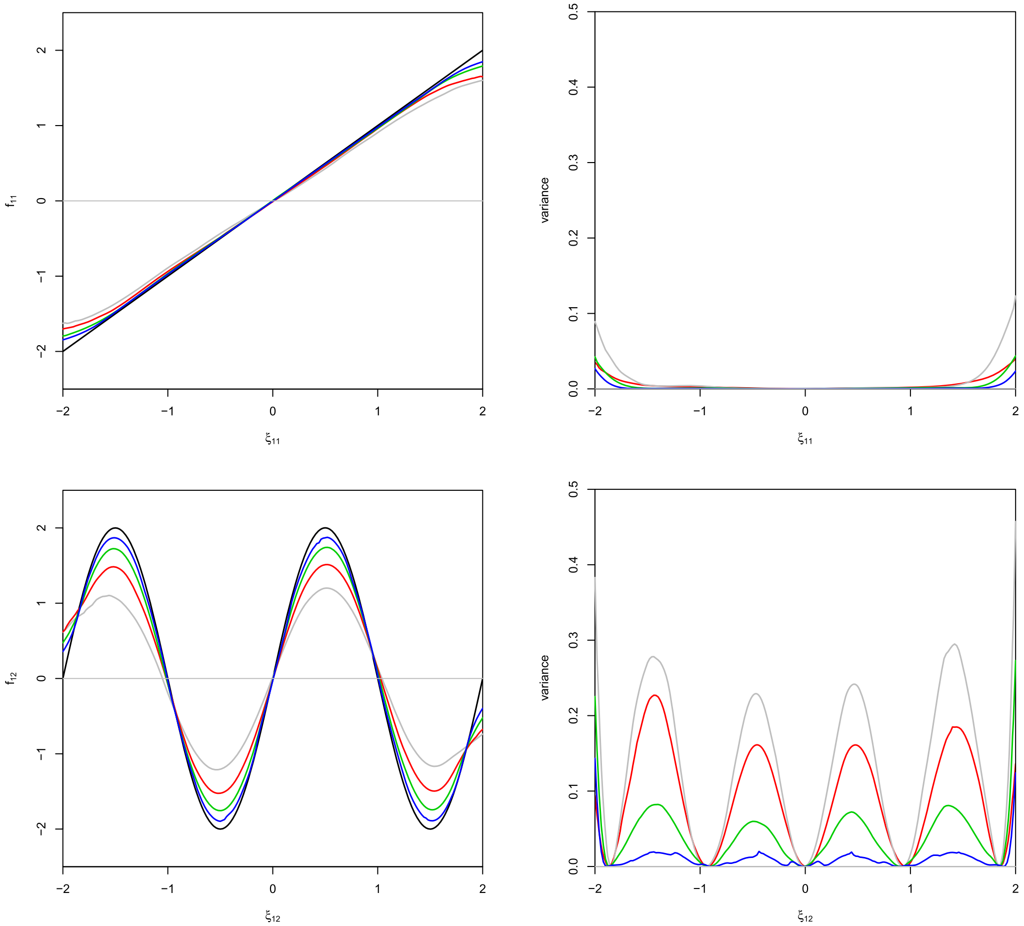

Figure 2. Average functions of $\hat{f}_{jk}$ (left panels) and their variance functions (right panels) for sample sizes $n = 100$ (grey), $n = 200$ (red), $n = 400$ (green) and $n = 1000$ (blue). True targets are black. Based on 400 Monte Carlo replications.

two models, we used the estimated FPCs as predictors in either a linear or a quadratic regular regression model.

We assessed the prediction performance by the mean squared prediction error (MSPE)

$$\mathrm{MSPE} = E\left[\frac{1}{N}\sum_{i=1}^{N}\left(Y^{\mathrm{new},i} - \hat{f}_n\left(\hat{\boldsymbol{\xi}}^{\mathrm{new},i}\right)\right)^2 \mathbb{I}\left(\hat{\boldsymbol{\xi}}^{\mathrm{new},i} \in I\right)\right].$$

Here, $Y^{\mathrm{new},i}$ are the responses in a test sample $\mathcal{X}^{\mathrm{new}}$ of size N that are independent of the training sample $\mathcal{X}_n$ of size n, the FPC score vectors $\hat{\boldsymbol{\xi}}^{\mathrm{new},i}$ are computed from the test sample, and $\hat{f}_n$ is

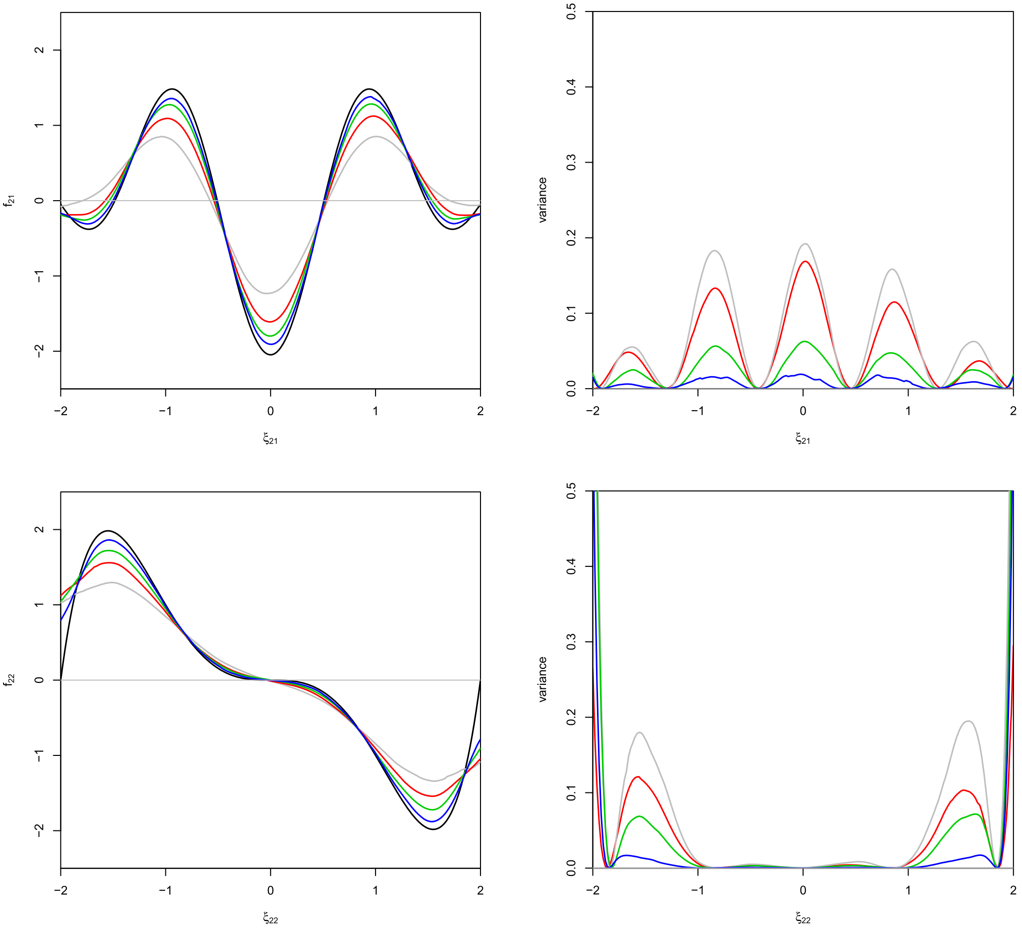

Figure 2. (Continued.)

the estimated model based on the training sample $\mathcal{X}_n$. We report results that are averaged over $B = 400$ Monte Carlo training samples and took $N = 1000$.

The prediction results are provided in Table 2. The unconditional variance of Y was found to be 64.64 in this simulation setting, which is useful to judge the improvement in prediction one obtains from each of the four methods in comparison with the naive prediction provided by the sample mean of the training data responses. The results show that AFSM gave the best performance, followed by FAM, and then the parametric FLM and FQM approaches by a large margin, especially for large sample sizes. The FAM approach performs worse compared to AFSM, because the former neglects the correlation structure between the FPC scores of different predictors, while the latter adjusts for it via the backfitting operation. In implementing the FAM approach, we used the baseline bandwidths h_{jk}^{full}.

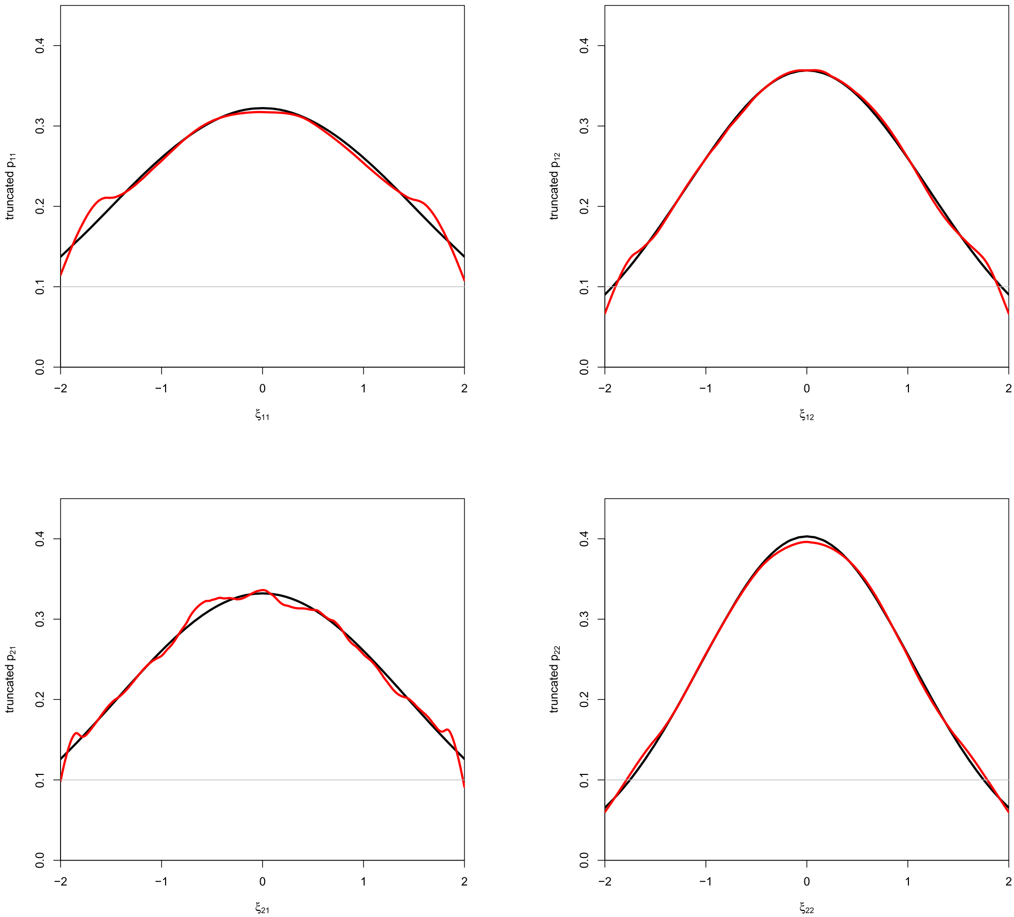

Figure 3. Truncated densities p_{jk}^{I} (black) and the averages of their estimates $\hat{p}_{jk}^{I}$ for $B = 400$ Monte Carlo samples of size $n = 200$.

5. Illustration with bike sharing data

Bike sharing systems increasingly replace traditional bike rentals, where the whole process of membership, rental and return is automatic [6]. More than 500 bike sharing systems are operating around the world (http://www.earth-policy.org/plan_b_updates/2013/update112). We analyze data from the Capital Bike Share System (http://capitalbikeshare.com/) that is operating in Washington, DC These data include the hourly counts of bike pick-ups between 2011 and 2012 and are available at https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. Initial exploration shows that hourly bike pick-up counts on weekdays and weekends tend to have quite distinct patterns. We focus on a prediction model for the total number of bike pick-ups on a Sunday as scalar response, based on the observed hourly bike pick-up counts on the preceding Friday

Table 2. Comparison of MSPE for the AFSM, FAM, FLM and FQM approaches based on $B = 400$ Monte Carlo samples. The number in parenthesis is the estimated standard error of the Monte Carlo approximation of MSPE

Sample size (n)	Mean squared prediction error			
	AFSM	FAM	FLM	FQM
100	1.345 (0.036)	7.441 (0.628)	9.337 (0.578)	10.871 (0.806)
200	0.877 (0.016)	2.447 (0.143)	8.889 (0.567)	9.224 (0.765)
400	0.589 (0.010)	1.379 (0.057)	7.988 (0.300)	8.856 (0.346)
1000	0.343 (0.008)	0.770 (0.037)	6.124 (0.098)	6.111 (0.092)

and Saturday, which constitute the two distinct but clearly correlated functional predictors. The total bike pick-up counts on Sunday and the hourly count curves on the preceding Friday and Saturday are in Figure 4. On Fridays the bike pick-ups have two well-defined peaks around 8am and 6pm that reflect bike usage for commuting to and from work. Saturday bike usage indicates a leisure use pattern, where most of the usage occurs during the daytime between 10am and 6pm, with a flat peak in usage during the afternoon.

Applying the FVE method to determine the number of FPC scores to enter into the model led to the choice of three eigencomponents for Friday and one for Saturday. Each set of eigencomponents explains more than 90% of the variation of the corresponding observed trajectories. The estimated covariance functions and the estimated eigenfunctions are displayed in Figures 5 and 6. The first eigenfunction $\hat{\phi}_{11}$ of the Friday data has a similar pattern as the average of the hourly Friday bike usage, depicted by the thick red curve in Figure 4, so this corresponds to a baseline factor that is likely due to seasonal effects. The second eigenfunction $\hat{\phi}_{12}$ presents a contrast between morning and afternoon bike usage, while the third eigenfunction $\hat{\phi}_{13}$ further differentiates usage around midday and evening. Analogously to the situation for the Friday data, the first eigenfunction $\hat{\phi}_{21}$ for the Saturday data, depicted in Figure 6, also is similar to the mean bike usage.

We applied the proposed additive functional regression with smooth backfitting method as well as the functional linear model (FLM) and functional additive model (FAM) approaches to the Capital Bike Share dataset, aiming to predict the Sunday total bike usage from the functional profiles generated by the hourly count data observed for the preceding Friday and Saturday. Figure 7 depicts the estimated component functions by the proposed method. In view of our interpretation for the estimated eigenfunctions, the patterns of $\hat{f}_{11}$ and $\hat{f}_{21}$ in the top left and bottom right panels indicate that the overall hourly profiles on Friday and Saturday are positively and monotonously associated with daily bike usage on the following Sunday, noting that the corresponding eigenfunctions $\hat{\phi}_{11}$ and $\hat{\phi}_{21}$ fall in roughly the same direction as the mean profile.

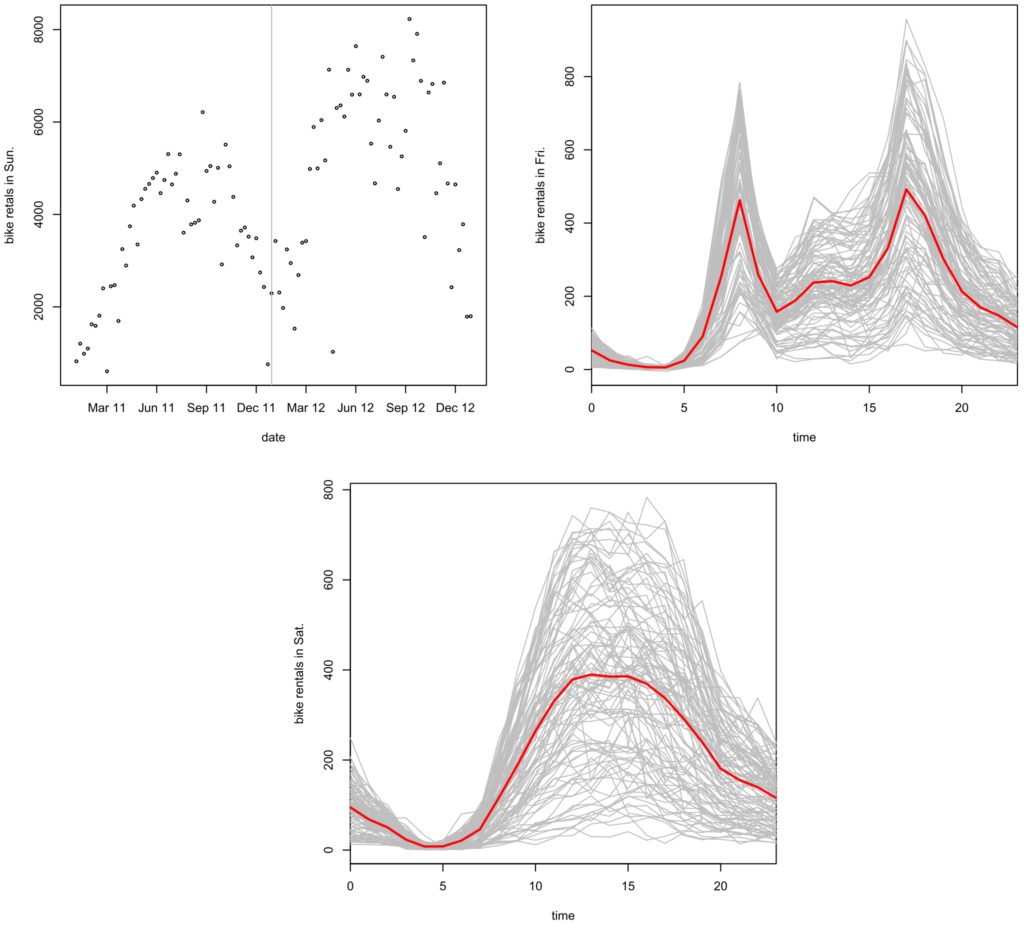

Figure 4. Total daily bike pick-up counts for 104 Sundays (top left) and pick-up curves for the preceding Fridays (top right) and Saturdays (bottom). The vertical grey line in the Sunday data (top left) indicates January 1st, 2012.

The second component function for the Friday usage profile $\hat{f}_{12}$ in the top right panel is nonlinear and shows that only negative functional principal components as predictor scores have a positive effect on Sunday bike usage, but that this effect tapers off as the predictor score increases beyond 0. In view of the shape of $\hat{\phi}_{12}$ in Figure 5, this indicates that substantially larger bike usage in the morning and reduced bike usage over noon on the preceding Friday is associated with increased bike usage on Sunday, as these features are associated with a negative value of $\hat{\xi}^i_{12}$. The third additive function $\hat{f}_{13}$ for the Friday predictor profile in the bottom left panel is monotonously declining, and viewing it in conjunction with the shape of $\hat{\phi}_{13}$ suggests that relatively more evening bike usage on Fridays, which probably is related to leisure use, relatively to

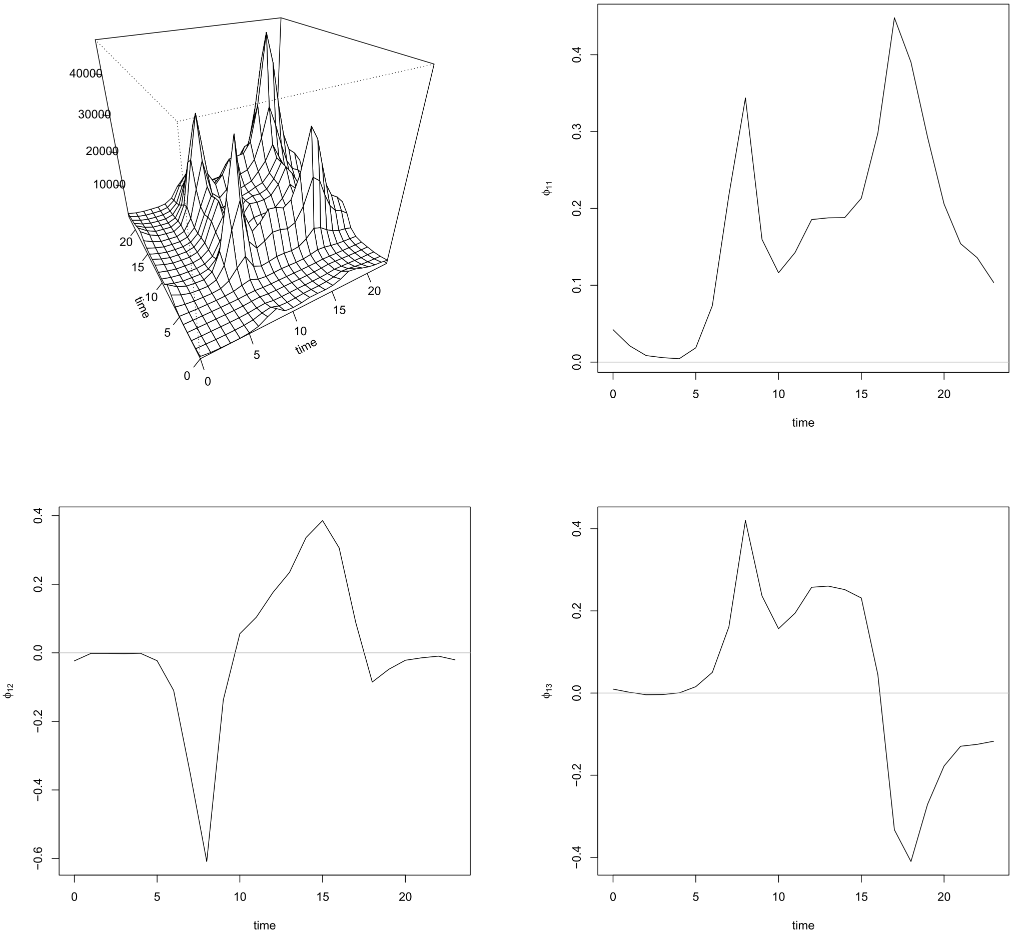

Figure 5. Estimated covariance function of bike usage for Fridays (top left) and its first (top right), second (bottom left) and third (bottom right) eigenfunction. The fraction of variance explained (FVE) by each of these three eigencomponents is 83.9%, 5.9% and 5.0%, respectively.

morning and mid day use, which probably is related to work usage, is associated with increased subsequent Sunday bike usage.

We also observe that the additive functions except for the second component of Friday have an overall linear trend, which suggest that the FLM might also work well for this functional regression problem. Using all the data, we found that the prediction performance of the FLM approach was similar to that of the proposed method, as $\sqrt{\mathrm{MSPE}} = 1143.3$ for the proposed method and $\sqrt{\mathrm{MSPE}} = 1145.2$ for the FLM, based on 10-fold CV. In contrast, the prediction error of the FAM was $\sqrt{\mathrm{MSPE}} = 1815.6$, which suggests that FAM is strongly biased for this prediction task, as it assumes independent predictor scores, an assumption that is violated for

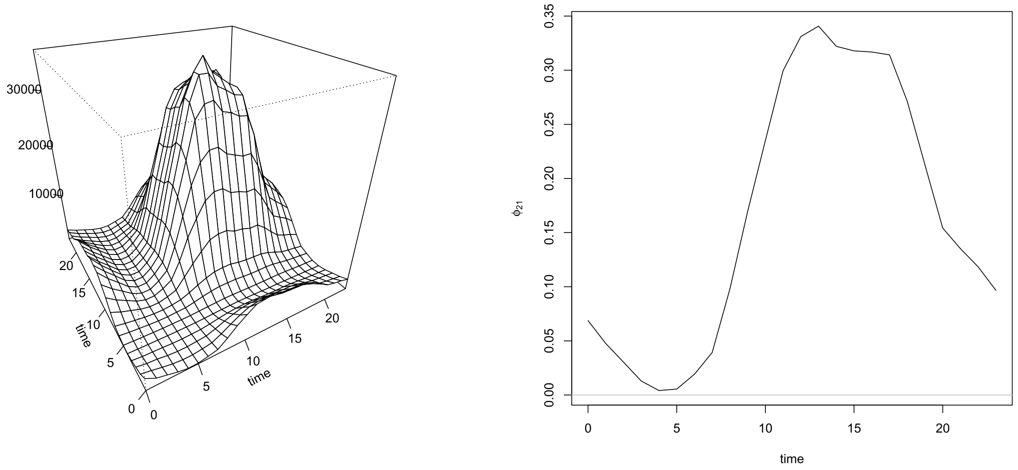

Figure 6. Estimated covariance function of bike usage for Saturdays (left) and its first eigenfunction (right). The fraction of variance explained (FVE) by the first eigencomponent is 90.2%.

these data. For the naive prediction made by the mean of the Sunday counts in the training sample, $\sqrt{\text{MSPE}} = 1901.7$.

The left panel of Figure 8 plots the predicted counts versus the observed ones. The three circled cases are outlying in the scatter plot. The circled case in the upper left area corresponds to the Sunday count on August 28, 2011. It was right after the Saturday when the hurricane "Irene" hit Washington, DC, so that the prediction is likely off for this reason. The circled case on the bottom right represents the count on October 7, 2012, right before Columbus Day, which probably also upended the prediction. The circled data point in the middle right area is for the 2012 George Washington Running Classic in Arlington and Alexandria, Virginia, which also likely affected bike usage on that Saturday. If we remove these three outliers from the data, we obtain the prediction results as depicted in the right panel of Figure 8, and the prediction accuracy was much improved, with $\sqrt{\text{MSPE}} = 819.1$ for our method, $\sqrt{\text{MSPE}} = 861.4$ for FLM and $\sqrt{\text{MSPE}} = 1348.2$ for FAM.

6. Technical details

We may estimate the mean functions μ_j by $\bar{X}_j = n^{-1}\sum_{i=1}^{n} X_j$ and these estimates will have the parametric $\sqrt{n}$ rate of convergence to μ_j. Thus, we assume without loss of generality that $\mu_j \equiv 0$ and neglect their estimation. As mentioned earlier, we consider in our asymptotic analysis the theoretical estimators $\hat{f}_{jk}^{*}$ that use the true unknown FPC scores ξ_{jk}^{i}. Below, we introduce some terminology for related terms. We let $\hat{f}_0^{*}$, $\hat{p}_0^{*I}$, $\hat{p}_{jk}^{*I}$, $\hat{p}_{jk,j'k'}^{*I}$ and $\tilde{f}_{jk}^{*}$, respectively, denote versions of $\hat{f}_0$, $\hat{p}_0^{I}$, $\hat{p}_{jk}^{I}$, $\hat{p}_{jk,j'k'}^{I}$ and $\tilde{f}_{jk}$ defined in Section 2.3 with $\hat{\xi}_{jk}^{i}$ being replaced by ξ_{jk}^{i}. The theoretical estimators $\hat{f}_{jk}^{*}$ are then defined to be the solution of a version of (13) subject to

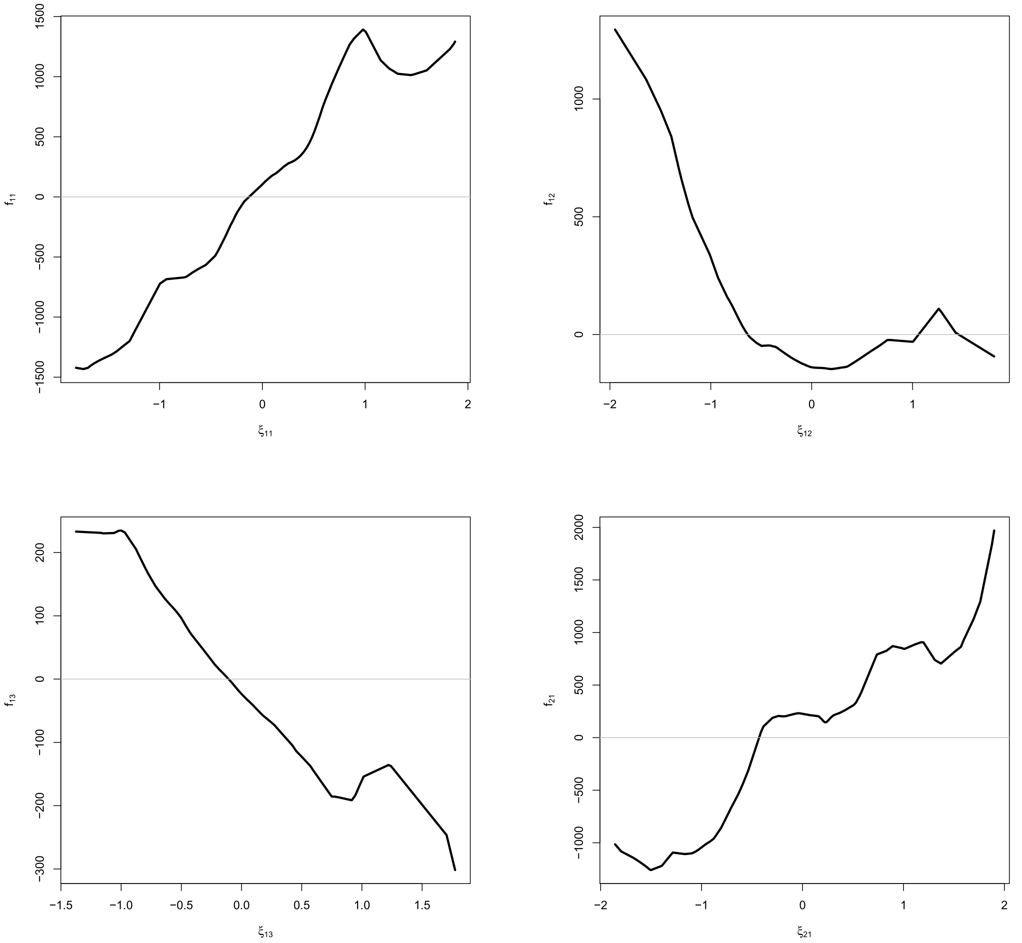

Figure 7. Estimated component functions $\hat{f}_{11}$ (top left), $\hat{f}_{12}$ (top right), $\hat{f}_{13}$ (bottom left) and $\hat{f}_{21}$ (bottom right) by our method applied to the bike rental data.

a version of (14), with $\hat{f}_0$, $\hat{p}^I_{jk}$, $\hat{p}^I_{jk,j'k'}$ and $\tilde{f}_{jk}$ being replaced by $\hat{f}^*_0$, $\hat{p}^{*I}_{jk}$, $\hat{p}^{*I}_{jk,j'k'}$ and $\tilde{f}^*_{jk}$, respectively. The corresponding theoretical versions of the rth update $\hat{f}^{[r]}_{jk}$ are defined analogously for the backfitting iteration (15).

We write $K^i_{jk}(u) = K_{h_{jk}}(u, \hat{\xi}^i_{jk})$, $K^{*i}_{jk}(u) = K_{h_{jk}}(u, \xi^i_{jk})$, $\mathbb{I}^i = \mathbb{I}(\hat{\boldsymbol{\xi}}^i \in I)$ and $\mathbb{I}^{*i} = \mathbb{I}(\boldsymbol{\xi}^i \in I)$. Let $\varepsilon^i = Y^i - \sum_{j=1}^d \sum_{k=1}^{L_j} f_{jk}(\xi^i_{jk})$ and define

$$\tilde{f}^A_{jk}(u) = \frac{1}{\hat{p}^I_0 \hat{p}^I_{jk}(u)} n^{-1} \sum_{i=1}^n K^i_{jk}(u) \mathbb{I}^i \varepsilon^i,$$

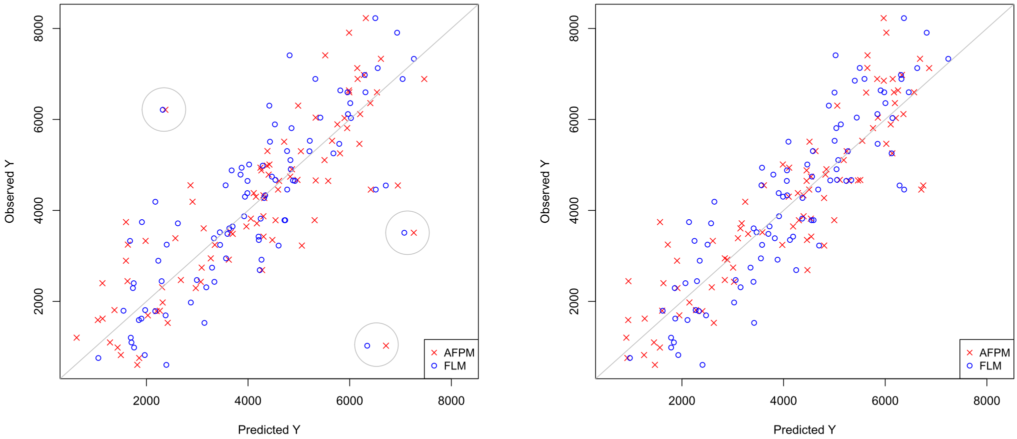

Figure 8. Prediction results for the proposed additive method and the FLM approach, where circles mark outliers, as described in the text.

$$\tilde{f}^B_{jk}(u) = \frac{1}{\hat{p}^I_0 \hat{p}^I_{jk}(u)} n^{-1} \sum_{i=1}^{n} K^i_{jk}(u)\mathbb{I}^i \big[f_{jk}\big(\xi^i_{jk}\big) - f_{jk}(u)\big], \tag{19}$$

$$\tilde{f}^C_{jk,j'k'}(u) = n^{-1} \sum_{i=1}^{n} K^i_{jk}(u)\mathbb{I}^i \int_0^1 K^i_{j'k'}(v)\big[f_{j'k'}\big(\xi^i_{j'k'}\big) - f_{j'k'}(v)\big]\,dv.$$

Likewise, define $\tilde{f}^{*A}_{jk}$, $\tilde{f}^{*B}_{jk}$ and $\tilde{f}^{*C}_{jk,j'k'}$, replacing $\hat{p}^I_0$, $\hat{p}^I_{jk}$, $\mathbb{I}^i$, K^i_{jk}, $K^i_{j'k'}$ in the definitions of $\tilde{f}^A_{jk}$, $\tilde{f}^B_{jk}$ and $\tilde{f}^C_{jk,j'k'}$ by $\hat{p}^{*I}_0$, $\hat{p}^{*I}_{jk}$, $\mathbb{I}^{*i}$, K^{*i}_{jk}, $K^{*i}_{j'k'}$, respectively. Note that we have $f_{jk}(\xi^i_{jk})$ and $f_{j'k'}(\xi^i_{j'k'})$, instead of $f_{jk}(\hat{\xi}^i_{jk})$ and $f_{j'k'}(\hat{\xi}^i_{j'k'})$, in the definitions of $\tilde{f}^B_{jk}$ and $\tilde{f}^C_{jk,j'k'}$.

6.1. Preliminary results

Here, we present two lemmas for the approximation of some relevant terms in the analysis of the backfitting equations. The lemmas are based on Proposition 1 in Section 3.1.

Lemma 1. *Under the conditions* (A1)–(A6), *we have*

$$\hat{p}^I_0 - \hat{p}^{*I}_0 = O_p\big(n^{-(\beta-1)/(2\beta)}\big),$$

$$\sup_{u\in[0,1]} \big|\hat{p}^I_{jk}(u) - \hat{p}^{*I}_{jk}(u)\big| = O_p\big(n^{-(3\beta-5)/(10\beta)}\big) \qquad \textit{for all } j,k,$$

$$\sup_{u,v\in[0,1]} \big|\hat{p}^I_{jk,j'k'}(u,v) - \hat{p}^{*I}_{jk,j'k'}(u,v)\big| = O_p\big(n^{-(3\beta-5)/(10\beta)}\big) \qquad \textit{for all } \big(j',k'\big) \neq (j,k).$$

Proof. For the proof of the first claim, we may assume that $\max_{i,j,k} |\hat{\xi}^i_{jk} - \xi^i_{jk}| \le C_0 n^{-(\beta-1)/(2\beta)}$ for some positive constant C_0, due to Proposition 1. Define $I_n = I^L_n / I^S_n$, where

$$I^L_n = \{\mathbf{u} : -C_0 n^{-(\beta-1)/(2\beta)} \le u_{jk} \le 1 + C_0 n^{-(\beta-1)/(2\beta)}, 1 \le j \le d, 1 \le k \le L_j\},$$
$$I^S_n = \{\mathbf{u} : C_0 n^{-(\beta-1)/(2\beta)} \le u_{jk} \le 1 - C_0 n^{-(\beta-1)/(2\beta)}, 1 \le j \le d, 1 \le k \le L_j\}.$$

The volume of I_n in $\mathbb{R}^{L_1+\cdots+L_d}$ is of order $n^{-(\beta-1)/(2\beta)}$. Thus, we have

$$|\hat{p}^I_0 - \hat{p}^{*I}_0| \le n^{-1} \sum_{i=1}^n \mathbb{I}(\boldsymbol{\xi}^i \in I_n) = O_p(n^{-(\beta-1)/(2\beta)}).$$

Among the last two claims of the lemma, we only prove the third one. The second one follows by similar arguments. From the condition (A2) and Proposition 1, we get

$$\sup_{u \in [0,1]} |K^i_{jk}(u) - K^{*i}_{jk}(u) - (\hat{\xi}^i_{jk} - \xi^i_{jk}) K'_{h_{jk}}(u, \xi^i_{jk})| \le L(\hat{\xi}^i_{jk} - \xi^i_{jk})^2 h^{-3}_{jk} \tag{20}$$

for all $1 \le i \le n$, where $L > 0$ is an absolute constant and $K'_g(u, v) = \partial K_g(u, v)/\partial v$. We also obtain

$$\begin{aligned}
\sup_{u \in [0,1]} n^{-1} \sum_{i=1}^n |K^{*i}_{jk}(u)| &= O_p(1), \\
\sup_{u \in [0,1]} n^{-1} \sum_{i=1}^n |K'_{h_{jk}}(u, \xi^i_{jk})| &= O_p(h^{-1}_{jk}), \\
\sup_{u,v \in [0,1]} n^{-1} \sum_{i=1}^n |K'_{h_{jk}}(u, \xi^i_{jk}) K^{*i}_{j'k'}(v)| &= O_p(h^{-1}_{jk}), \\
\sup_{u,v \in [0,1]} n^{-1} \sum_{i=1}^n |K'_{h_{jk}}(u, \xi^i_{jk}) K'_{h_{j'k'}}(v, \xi^i_{j'k'})| &= O_p(h^{-1}_{jk} h^{-1}_{j'k'}).
\end{aligned} \tag{21}$$

From (20) and (21), we can deduce

$$n^{-1} \sum_{i=1}^n K^i_{jk}(u) K^i_{j'k'}(v) \mathbb{I}^i = n^{-1} \sum_{i=1}^n K^{*i}_{jk}(u) K^{*i}_{j'k'}(v) \mathbb{I}^i + O_p(n^{-(3\beta-5)/(10\beta)})$$

uniformly for $u, v \in [0, 1]$. Now, assuming $\max_{i,j,k} |\hat{\xi}^i_{jk} - \xi^i_{jk}| \le C_0 n^{-(\beta-1)/(2\beta)}$ as in the proof of the first claim and making the following approximation completes the proof of the third part of the lemma:

$$\begin{aligned}
n^{-1} \sum_{i=1}^n K^{*i}_{jk}(u) K^{*i}_{j'k'}(v) |\mathbb{I}^i - \mathbb{I}^{*i}| &\le n^{-1} \sum_{i=1}^n K^{*i}_{jk}(u) K^{*i}_{j'k'}(v) \mathbb{I}(\boldsymbol{\xi}^i \in I_n) \\
&= O_p(n^{-(\beta-1)/(2\beta)+(1/5)}) = O_p(n^{-(3\beta-5)/(10\beta)})
\end{aligned} \tag{22}$$

uniformly for $u, v \in [0,1]$. □

Lemma 2. *Under the conditions* (A1)–(A6), *we have*

$$\sup_{u\in[0,1]} \bigl|\tilde{f}^A_{jk}(u) - \tilde{f}^{*A}_{jk}(u)\bigr| = O_p\bigl(n^{-(11\beta-5)/(20\beta)}(\log n)^{1/2}\bigr) \qquad \textit{for all } j,k,$$

$$\sup_{u\in[0,1]} \bigl|\tilde{f}^B_{jk}(u) - \tilde{f}^{*B}_{jk}(u)\bigr| = O_p\bigl(n^{-(\beta-1)/(2\beta)}\bigr) \qquad \textit{for all } j,k,$$

$$\sup_{u\in[0,1]} \bigl|\tilde{f}^C_{jk,j'k'}(u) - \tilde{f}^{*C}_{jk,j'k'}(u)\bigr| = O_p\bigl(n^{-(\beta-1)/(2\beta)}\bigr) \qquad \textit{for all } \bigl(j',k'\bigr) \neq (j,k).$$

Proof. We prove the first and third parts only. The second part follows by the arguments used in the proof of the third part. For the first part, we note that from (20)

$$\begin{aligned} n^{-1}\sum_{i=1}^n K^i_{jk}(u)\mathbb{I}^i\varepsilon^i &= n^{-1}\sum_{i=1}^n K^{*i}_{jk}(u)\mathbb{I}^i\varepsilon^i + n^{-1}\sum_{i=1}^n \bigl(\hat{\xi}^i_{jk} - \xi^i_{jk}\bigr)K'_{h_{jk}}\bigl(u,\xi^i_{jk}\bigr)\mathbb{I}^i\varepsilon^i \\ &\quad + O_p\bigl(n^{-(9\beta-10)/(10\beta)}\bigr) \end{aligned} \tag{23}$$

uniformly for $u \in [0,1]$. By an application of an exponential inequality conditioning on $(\mathbf{X}^i : 1 \le i \le n)$ and the use of Proposition 1, we may show that the second term on the right-hand side of (23) is of order $O_p(n^{-1/2}h^{-3/2}_{jk}(\log n)^{1/2}n^{-(\beta-1)/(2\beta)}) = O_p(n^{-(7\beta-5)/(10\beta)}(\log n)^{1/2})$ uniformly for $u \in [0,1]$. Similarly,

$$n^{-1}\sum_{i=1}^n K^{*i}_{jk}(u)\bigl(\mathbb{I}^i - \mathbb{I}^{*i}\bigr)\varepsilon^i = O_p\bigl(n^{-1/2}h^{-1}_{jk}n^{-(\beta-1)/(4\beta)}(\log n)^{1/2}\bigr)$$

uniformly for $u \in [0,1]$, where we have used

$$n^{-1}\sum_{i=1}^n K^{*i}_{jk}(u)^2\bigl(\mathbb{I}^i - \mathbb{I}^{*i}\bigr)^2 = O_p\bigl(h^{-2}_{jk}n^{-(\beta-1)/(2\beta)}\bigr)$$

uniformly for $u \in [0,1]$. This completes the proof of the first part.

To prove the third part, we replace $K^i_{jk}(u)$ in $\tilde{f}^C_{jk,j'k'}(u)$ at (19) by its approximation $K^{*i}_{jk}(u) + (\hat{\xi}^i_{jk} - \xi^i_{jk})K'_{h_{jk}}(u,\xi^i_{jk}) + (\text{remainder})$, with the remainder being of order $n^{-(2\beta-5)/(5\beta)}$ uniformly for $u \in [0,1]$. Likewise, we replace $K^i_{j'k'}(v)$ in $\tilde{f}^C_{jk,j'k'}(u)$ by similar terms. This gives a decomposition of $\tilde{f}^C_{jk,j'k'}(u) - \tilde{f}^{*C}_{jk,j'k'}(u)$ into several terms. The three leading terms are

$$\begin{aligned} \mathbf{I}+\mathbf{II}+\mathbf{III} \equiv{}& n^{-1}\sum_{i=1}^n K^{*i}_{jk}(u)\mathbb{I}^i\bigl(\hat{\xi}^i_{j'k'} - \xi^i_{j'k'}\bigr)\int_0^1 K'_{h_{j'k'}}\bigl(v,\xi^i_{j'k'}\bigr)\bigl[f_{j'k'}\bigl(\xi^i_{j'k'}\bigr) - f_{j'k'}(v)\bigr]\,dv \\ &+ n^{-1}\sum_{i=1}^n K'_{h_{jk}}\bigl(u,\xi^i_{jk}\bigr)\mathbb{I}^i\bigl(\hat{\xi}^i_{jk} - \xi^i_{jk}\bigr)\int_0^1 K^{*i}_{j'k'}(v)\bigl[f_{j'k'}\bigl(\xi^i_{j'k'}\bigr) - f_{j'k'}(v)\bigr]\,dv \end{aligned}$$

$$+ n^{-1} \sum_{i=1}^{n} K_{jk}^{*i}(u)\big(\mathbb{I}^i - \mathbb{I}^{*i}\big) \int_0^1 K_{j'k'}^{*i}(v)\big[f_{j'k'}\big(\xi_{j'k'}^i\big) - f_{j'k'}(v)\big]\, dv.$$

All others are of smaller order. Using the third property of (21) and the fact that

$$\big|K'_{h_{j'k'}}\big(v, \xi_{j'k'}^i\big)\big(f_{j'k'}\big(\xi_{j'k'}^i\big) - f_{j'k'}(v)\big)\big| \le C h_{j'k'} K'_{h_{j'k'}}\big(v, \xi_{j'k'}^i\big) \tag{24}$$

for some constant $C > 0$, we get that both **I** and **II** are of order $O_p(n^{-(\beta-1)/(2\beta)})$ uniformly for $u \in [0, 1]$. Note that (24) also holds with $K_{j'k'}^{*i}(v)$ replacing $K'_{h_{j'k'}}(v, \xi_{j'k'}^i)$. This together with (22) gives $\mathbf{I}II = O_p(n^{-(\beta-1)/(2\beta)})$ uniformly for $u \in [0, 1]$. This completes the proof of the lemma. □

6.2. Proof of Theorem 1

Define linear operators

$$\pi_{jk}(g) = \int_{I_{-jk}} g(\mathbf{u}) \frac{p^I(\mathbf{u})}{p_{jk}^I(u_{jk})}\, d\mathbf{u}_{-jk},$$

and likewise $\hat{\pi}_{jk}$ and $\hat{\pi}_{jk}^*$, respectively, replacing (p^I, p_{jk}^I) by $(\hat{p}^I, \hat{p}_{jk}^I)$ and $(\hat{p}^{*I}, \hat{p}_{jk}^{*I})$, where $p^I(\mathbf{u}) = p(\mathbf{u})/p_0^I$, $\hat{p}^I(\mathbf{u}) = \hat{p}(\mathbf{u})/\hat{p}_0^I$ and $\hat{p}^{*I}(\mathbf{u}) = \hat{p}^*(\mathbf{u})/\hat{p}_0^{*I}$. Define a linear operator

$$T = (I - \pi_{d,L_d})(I - \pi_{d,L_d-1}) \cdots (I - \pi_{1,2})(I - \pi_{1,1}),$$

and likewise $\hat{T}$ and $\hat{T}^*$ with π_{jk} being replaced by $\hat{\pi}_{jk}$ and $\hat{\pi}_{jk}^*$, respectively. For a linear operator F that maps the space of additive functions to itself, we define its norm $\|F\|$ by

$$\|F\|^2 = \sup\left\{\int F(g)(\mathbf{u})^2 p^I(\mathbf{u})\, d\mathbf{u} : g \text{ is additive and } \int g(\mathbf{u})^2 p^I(\mathbf{u})\, d\mathbf{u} = 1\right\}.$$

Then, along the lines of the proof of Theorem 1 in Mammen, Linton and Nielsen [16] it can be proved that

$$\big\|\hat{T}^* - T\big\| = o_p(1), \qquad \|T\| < \gamma \tag{25}$$

for some constant $0 < \gamma < 1$.

Next, we define

$$\begin{aligned} \tilde{f}_\oplus = {} & \tilde{f}_{d,L_d} + (I - \hat{\pi}_{d,L_d})\tilde{f}_{d,L_d-1} + (I - \hat{\pi}_{d,L_d})(I - \hat{\pi}_{d,L_d-1})\tilde{f}_{d,L_d-2} \\ & + \cdots + (I - \hat{\pi}_{d,L_d}) \cdots (I - \hat{\pi}_{1,2})\tilde{f}_{1,1} \end{aligned} \tag{26}$$

and likewise $\tilde{f}_\oplus^*$ with $\hat{\pi}_{jk}$ being replaced by $\hat{\pi}_{jk}^*$. With the additive functions

$$\hat{f}_+(\mathbf{u}) \equiv \sum_{j=1}^{d} \sum_{k=1}^{L_j} \hat{f}_{jk}(u_{jk}), \qquad \hat{f}_+^*(\mathbf{u}) \equiv \sum_{j=1}^{d} \sum_{k=1}^{L_j} \hat{f}_{jk}^*(u_{jk}),$$

$$\hat{f}_+^{[r]}(\mathbf{u}) \equiv \sum_{j=1}^{d}\sum_{k=1}^{L_j} \hat{f}_{jk}^{[r]}(u_{jk}), \qquad \hat{f}_+^{*[r]}(\mathbf{u}) \equiv \sum_{j=1}^{d}\sum_{k=1}^{L_j} \hat{f}_{jk}^{*[r]}(u_{jk}),$$

the whole system of the backfitting equations at (13) and its theoretical version can be written as $\hat{f}_+ = \tilde{f}_\oplus + \hat{T}\hat{f}_+$ and $\hat{f}_+^* = \tilde{f}_\oplus^* + \hat{T}^*\hat{f}_+^*$, respectively. It also follows that $\hat{f}_+^{[r]} = \tilde{f}_\oplus + \hat{T}\hat{f}_+^{[r-1]}$ and $\hat{f}_+^{*[r]} = \tilde{f}_\oplus^* + \hat{T}^*\hat{f}_+^{*[r-1]}$. Because of (25) and from the standard theory of smooth backfitting, the theorem follows if we prove $\|\hat{T} - \hat{T}^*\| = o_p(1)$. The latter is a direct consequence of Lemma 1 since the second and third parts of the lemma imply $\|\hat{\pi}_{jk} - \hat{\pi}_{jk}^*\| = O_p(n^{-(3\beta-5)/(10\beta)}) = o_p(1)$.

6.3. Proof of Theorem 2

We assume $f_0 = 0$ and ignore $\hat{f}_0$ and $\hat{f}_0^*$ in the backfitting equation (13) and its theoretical version, respectively. This is justified since $\hat{f}_0^* - f_0$ is of order $n^{-1/2}$ and $\hat{f}_0 - \hat{f}_0^*$ is of order $n^{-(\beta-1)/(2\beta)} = o(n^{-2/5})$ if $\beta > 5$.

The main idea of the proof is to extract the key stochastic terms from $\tilde{f}_{jk}$, put them into the smooth backfitting operation together with $\hat{f}_{jk}$ and then prove that the estimation of the FPC scores ξ_{jk}^i by $\hat{\xi}_{jk}^i$ has a negligible effect on the resulting smooth backfitting equation. Indeed, with those terms defined in (19) we may express the backfitting equation (13) as follows.

$$\begin{aligned}\hat{f}_{jk}(u) = f_{jk}(u) + \tilde{f}_{jk}^A(u) + \tilde{f}_{jk}^B(u) + \frac{1}{\hat{p}_0^I\hat{p}_{jk}^I(u)} \sum_{(j',k')\neq(j,k)} \tilde{f}_{jk,j'k'}^C(u) \\ - \sum_{(j',k')\neq(j,k)} \int_0^1 \big[\hat{f}_{j'k'}(v) - f_{j'k'}(v)\big]\frac{\hat{p}_{jk,j'k'}^I(u,v)}{\hat{p}_{jk}^I(u)}\,dv.\end{aligned} \tag{27}$$

For the equation (27), we have used $\int K_{jk}^i(u)\,du = 1$. We take some parts of $\tilde{f}_{jk,j'k'}^C(u)$ and put them into the integral term in (27) and then approximate the solution of the resulting backfitting equation to get

$$\hat{f}_{jk}(u) = f_{jk}(u) + \tilde{f}_{jk}^{*A}(u) + h_{jk}\frac{a_{jk}(u)}{\mu_{0,jk}(u)} + \frac{1}{2}h_{jk}^2\mu_2 f_{jk}''(u) + \Delta_{jk}(u) + r_{jk}(u). \tag{28}$$

Here and below, $\mu_{l,jk}(z) = h_{jk}^{-l}\int(w-z)^l K_{h_{jk}}(z,w)\,dw$ and $a_{jk}(z) = \mu_{1,jk}(z)f_{jk}'(z)$. The remainder r_{jk} denotes a generic stochastic term such that

$$\sup_{u\in[2h_{jk},1-2h_{jk}]}\big|r_{jk}(u)\big| = o_p\big(n^{-2/5}\big), \qquad \sup_{u\in[0,1]}\big|r_{jk}(u)\big| = O_p\big(n^{-2/5}\big).$$

Also, the tuple $(\Delta_{jk} : 1 \le j \le d, 1 \le k \le L_j)$ is defined to be the solution of the system of equations

$$\Delta_{jk}(u) = \tilde{\Delta}_{jk}(u) - \sum_{(j',k')\neq(j,k)} \int_0^1 \Delta_{j'k'}(v)\frac{p_{jk,j'k'}^I(u,v)}{p_{jk}^I(u)}\,dv \tag{29}$$

subject to

$$\int_0^1 \Delta_{jk}(u)p^I_{jk}(u)\,du = \mu_2 h^2_{jk}\int_0^1 f'_{jk}(u)\frac{\partial}{\partial u}p^I_{jk}(u)\,du, \tag{30}$$

where $\mu_l = \int u^l K(u)\,du$ and

$$\tilde{\Delta}_{jk}(u) = \mu_2 \sum_{j'=1}^{d}\sum_{k'=1}^{L_{j'}} h^2_{j'k'} E\biggl[f'_{j'k'}(\xi_{j'k'})\frac{p^{(1)}_{j'k'}(\boldsymbol{\xi})}{p(\boldsymbol{\xi})}\bigg|\xi_{jk}=u, \boldsymbol{\xi}\in I\biggr].$$

The tuple that satisfies the system of equations (29) is unique up to an additive constant vector. This can be seen from the fact that replacing $\Delta_{jk}(u)$ by $\Delta_{jk}(u)+c$ on the left-hand side and $\Delta_{j'k'}(v)$ by $\Delta_{j'k'}(v)-c$ for a particular (j',k') on the right-hand side gives another solution. With the constraints at (30), however, the tuple $(\Delta_{jk} : 1 \le j \le d, 1 \le k \le L_j)$ is uniquely determined.

The first part of the theorem follows immediately from (28) since $\tilde{f}^{*A}_{jk}$ for different pairs (j,k) are asymptotically independent and $n^{2/5}(\tilde{f}^{*A}_{jk}(u) - f_{jk}(u))$ converges in distribution to $N(0,\sigma^2_{jk}(u))$. For the second part of the theorem, we note

$$\|\tilde{f}^{*A}_{jk}\| = O_p\bigl(n^{-2/5}\bigr), \qquad \sup_{u\in[0,1]}\bigl|\tilde{f}^{*A}_{jk}(u)\bigr| = O_p\bigl(n^{-2/5}\sqrt{\log n}\bigr)$$

from standard results of kernel smoothing. Since we also have $a_{jk}(u) = 0$ for $u \in [2h_{jk}, 1-2h_{jk}]$ and $\sup_{u\in[0,1]}|a_{jk}(u)| = O(1)$, the second part of the theorem follows from (28). We now prove (28). The proof is decomposed into several steps.

Approximation of $\tilde{f}^C_{jk,j'k'}(u)$: First of all, we note from Lemma 2 that $\tilde{f}^C_{jk,j'k'}(u) = \tilde{f}^{*C}_{jk,j'k'}(u) + o_p(n^{-2/5})$ uniformly for $u \in [0,1]$. To approximate $\tilde{f}^{*C}_{jk,j'k'}(u)$ further, define

$$\delta^i_{jk} = \int \bigl[f_{jk}\bigl(\xi^i_{jk}\bigr) - f_{jk}(z)\bigr]K^i_{jk}(z)\,dz.$$

Then, $\tilde{f}^{*C}_{jk,j'k'}(u) = n^{-1}\sum_{i=1}^n \delta^i_{j'k'}\mathbb{I}^{*i}K^{*i}_{jk}(u)$. Using standard results for kernel smoothing, it can be shown that

$$\sup_{u\in[0,1]}\left|n^{-1}\sum_{i=1}^n\bigl[\delta^i_{j'k'} - E\bigl(\delta^i_{j'k'}|\xi^i_{jk},\mathbb{I}^{*i}\bigr)\bigr]K^{*i}_{jk}(u)\mathbb{I}^{*i}\right| = O_p\bigl(n^{-3/5}\sqrt{\log n}\bigr). \tag{31}$$

We compute $E(\delta^i_{j'k'}|\xi^i_{jk} = v, \boldsymbol{\xi}^i \in I)$. We note that $\mu_{l,jk}(z) = 0$ for $z \in [2h_{jk}, 1-2h_{jk}]$ if l is an odd positive integer and the baseline kernel K is symmetric. Let $b_{jk}(z) = \mu_2 f''_{jk}(z)/2$ and

$$c_{jk,j'k'}(v,z) = \mu_2 f'_{j'k'}(z)p^I_{jk,j'k'}(v,z)^{-1}\partial p^I_{jk,j'k'}(v,z)/\partial z.$$

By expanding $f_{jk}(w) - f_{jk}(z)$ and the conditional density $p^I_{jk,j'k'}(v,w)/p^I_{jk}(v)$ for w near z, we get

$$E\big(\delta^i_{j'k'}|\xi^i_{jk} = v, \boldsymbol{\xi}^i \in I\big) = \int_0^1 \frac{p^I_{jk,j'k'}(v,z)}{p^I_{jk}(v)} \big[h_{j'k'}a_{j'k'}(z) + h^2_{j'k'}b_{j'k'}(z) \\ + h^2_{j'k'}c_{jk,j'k'}(v,z)\big]\,dz + o_p\big(n^{-2/5}\big)$$

uniformly for $u \in [0,1]$. For this, we have used the formula

$$E\big[g(\boldsymbol{\xi})|\xi_{jk} = u_{jk}, \boldsymbol{\xi} \in I\big] = \left(\int_{I_{-jk}} p(\mathbf{u})\,d\mathbf{u}_{-jk}\right)^{-1} \int_{I_{-jk}} g(\mathbf{u})p(\mathbf{u})\,d\mathbf{u}_{-jk} \tag{32}$$

for $u_{jk} \in [0,1]$. This together with (31) gives that, uniformly for $u \in [0,1]$,

$$\begin{aligned}\tilde{f}^{*C}_{jk,j'k'}(u) &= n^{-1}\sum_{i=1}^n E\big(\delta^i_{j'k'}|\xi^i_{jk}, \mathbb{I}^{*i}\big)\mathbb{I}^{*i}K^{*i}_{jk}(u) + o_p\big(n^{-2/5}\big) \\ &= n^{-1}\sum_{i=1}^n \mathbb{I}^{*i}\int_0^1 \frac{p^I_{jk,j'k'}(\xi^i_{jk},z)}{p^I_{jk}(\xi^i_{jk})}\big[h_{j'k'}a_{j'k'}(z) + h^2_{j'k'}b_{j'k'}(z) \\ &\quad + h^2_{j'k'}c_{jk,j'k'}(u,z)\big]K^{*i}_{jk}(u)\,dz + o_p\big(n^{-2/5}\big).\end{aligned} \tag{33}$$

We further approximate the main term on the right-hand side of (33). From (32), we get

$$E\big(K^{*i}_{j'k'}(z)|\xi^i_{jk}, \mathbb{I}^{*i}\big) = \int_0^1 K_{h_{j'k'}}(z,w)\frac{p^I_{jk,j'k'}(\xi^i_{jk},w)}{p^I_{jk}(\xi^i_{jk})}\,dw. \tag{34}$$

Note that $a_{jk}(z) = O(1), b_{jk}(z) = O(1)$ and $c_{jk,j'k'}(u,z) = O(1)$ uniformly for $u, z \in [0,1]$. Also, $a_{jk}(z) = 0$ for $z \in [2h_{jk}, 1-2h_{jk}]$. Thus, from (34) the main term on the right-hand side of (33) equals $\mathbf{I}V + \mathbf{V} + o_p(n^{-2/5})$ uniformly for $u \in [0,1]$, where

$$\begin{aligned}\mathbf{IV} &= \hat{p}^{*I}_0 \int_0^1 \left[h_{j'k'}\frac{a_{j'k'}(z)}{\mu_{0,j'k'}(z)} + h^2_{j'k'}b_{j'k'}(z) + h^2_{j'k'}c_{jk,j'k'}(u,z)\right]\hat{p}^{*I}_{jk,j'k'}(u,z)\,dz, \\ \mathbf{V} &= -n^{-1}\sum_{i=1}^n \mathbb{I}^{*i}\int_0^1 \left[h_{j'k'}\frac{a_{j'k'}(z)}{\mu_{0,j'k'}(z)} + h^2_{j'k'}b_{j'k'}(z) + h^2_{j'k'}c_{jk,j'k'}(u,z)\right] \\ &\quad \times \big[K^{*i}_{j'k'}(z) - E\big(K^{*i}_{j'k'}(z)|\xi^i_{jk}, \mathbb{I}^{*i}\big)\big]K^{*i}_{jk}(u)\,dz.\end{aligned} \tag{35}$$

By Lemma 1, we may replace $\hat{p}^{*I}_0$ and $\hat{p}^{*I}_{jk,j'k'}$ by $\hat{p}^I_0$ and $\hat{p}^I_{jk,j'k'}$, respectively, in **IV** with an approximation error $o_p(n^{-2/5})$ uniformly for $u \in [0,1]$. Also, we get $\mathbf{V} = o_p(n^{-2/5})$ uniformly for $u \in [0,1]$. This gives

$$\begin{aligned}\tilde{f}^C_{jk,j'k'}(u) = \hat{p}^I_0 \int_0^1 \biggl[h_{j'k'} \frac{a_{j'k'}(z)}{\mu_{0,j'k'}(z)} + h^2_{j'k'} b_{j'k'}(z) + h^2_{j'k'} c_{jk,j'k'}(u,z) \biggr] \\ \times \hat{p}^I_{jk,j'k'}(u,z)\,dz + o_p\bigl(n^{-2/5}\bigr)\end{aligned} \tag{36}$$

uniformly for $u \in [0.1]$.

Derivation of a new backfitting equation: From Lemma 2, we have $\tilde{f}^B_{jk}(u) = \tilde{f}^{*B}_{jk}(u) + o_p(n^{-2/5})$ uniformly for $u \in [0,1]$. Furthermore,

$$\tilde{f}^{*B}_{jk}(u) = h_{jk} \frac{a_{jk}(u)}{\mu_{0,jk}(u)} + h^2_{jk} b_{jk}(u) + h^2_{jk} c_{jk}(u) + r_{jk}(u), \tag{37}$$

where $c_{jk}(u) = \mu_2 f'_{jk}(u) p^I_{jk}(u)^{-1} \partial p^I_{jk}(u)/\partial u$ and r_{jk} denotes a generic stochastic term defined above. Then, (27), (36), (37) and Lemma 2 give

$$\begin{aligned}\hat{f}_{jk}(u) = f_{jk}(u) + \tilde{f}^{*A}_{jk}(u) + h_{jk} \frac{a_{jk}(u)}{\mu_{0,jk}(u)} + h^2_{jk} b_{jk}(u) + \tilde{\Delta}_{jk}(u) \\ - \sum_{(j',k') \neq (j,k)} \int_0^1 \biggl[\hat{f}_{j'k'}(v) - f_{j'k'}(v) - \tilde{f}^{*A}_{j'k'}(v) - h_{j'k'} \frac{a_{j'k'}(v)}{\mu_{0,j'k'}(v)} \\ - h^2_{j'k'} b_{j'k'}(v) \biggr] \frac{\hat{p}^I_{jk,j'k'}(u,v)}{\hat{p}^I_{jk}(u)}\,dv + r_{jk}(u).\end{aligned} \tag{38}$$

In the above equation, we have used

$$\sup_{u \in [0,1]} \biggl| \int_0^1 \tilde{f}^{*A}_{j'k'}(v) \frac{\hat{p}^I_{jk,j'k'}(u,v)}{\hat{p}^I_{jk}(u)}\,dv \biggr| = o_p\bigl(n^{-2/5}\bigr),$$

$$\sup_{u \in [0,1]} \biggl| \int_0^1 c_{jk,j'k'}(u,v) \biggl(\frac{\hat{p}^I_{jk,j'k'}(u,v)}{\hat{p}^I_{jk}(u)} - \frac{p^I_{jk,j'k'}(u,v)}{p^I_{jk}(u)} \biggr)\,dv \biggr| = o_p(1),$$

which are due to Lemma 1 and standard results for kernel smoothing. With

$$\hat{\Delta}_{jk}(u) \equiv \hat{f}_{jk}(u) - f_{jk}(u) - \tilde{f}^{*A}_{jk}(u) - h_{jk} \frac{a_{jk}(u)}{\mu_{0,jk}(u)} - h^2_{jk} b_{jk}(u) - r_{jk}(u),$$

(38) implies that, up to a remainder that is uniformly of order $o_p(n^{-2/5})$, the tuple $(\hat{\Delta}_{jk} : 1 \leq j \leq d, 1 \leq k \leq L_j)$ satisfies the system of integral equations

$$\hat{\Delta}_{jk}(u) = \tilde{\Delta}_{jk}(u) - \sum_{(j',k') \neq (j,k)} \int_0^1 \hat{\Delta}_{j'k'}(v) \frac{\hat{p}^I_{jk,j'k'}(u,v)}{\hat{p}^I_{jk}(u)}\,dv. \tag{39}$$

Approximation of $\hat{\Delta}_{jk}(u)$: We prove

$$\hat{\Delta}_{jk}(u) = \Delta_{jk}(u) + r_{jk}(u). \tag{40}$$

Define $\Delta_\oplus$ as $\tilde{f}_\oplus$ at (26) with $\tilde{f}_{jk}$ and $\hat{\pi}_{jk}$, respectively, being replaced by $\tilde{\Delta}_{jk}$ and π_{jk}. Also, define $\hat{\Delta}_\oplus$ with only $\tilde{f}_{jk}$ being replaced by $\tilde{\Delta}_{jk}$. For

$$\hat{\Delta}_+(\mathbf{u}) \equiv \sum_{j=1}^{d}\sum_{k=1}^{L_j} \hat{\Delta}_{jk}(u_{jk}), \qquad \Delta_+(\mathbf{u}) \equiv \sum_{j=1}^{d}\sum_{k=1}^{L_j} \Delta_{jk}(u_{jk}),$$

the backfitting equations (39) and (29) can be written as $\hat{\Delta}_+ = \hat{\Delta}_\oplus + \hat{T}\hat{\Delta}_+$ and $\Delta_+ = \Delta_\oplus + T\Delta_+$, respectively. Since $\sup_{u\in[0,1]}|\tilde{\Delta}_{jk}(u)| = O(n^{-2/5})$, and $\hat{\Delta}_\oplus$ differs from $\Delta_\oplus$ only in that it uses $\hat{\pi}_{jk}$ instead of π_{jk}, it follows from Lemma 1 that

$$\sup_{\mathbf{u}\in I_0}\bigl|\hat{\Delta}_\oplus(\mathbf{u}) - \Delta_\oplus(\mathbf{u})\bigr| = o_p\bigl(n^{-2/5}\bigr), \tag{41}$$

where $I_0 = \{\mathbf{u} : 2h_{jk} < u_{jk} < 1 - 2h_{jk}, 1 \le j \le d, 1 \le k \le L_j\}$. From (25) and the fact $\|\hat{T} - \hat{T}^*\| = o_p(1)$, we also have $\|\hat{T} - T\| = o_p(1)$ and $\|T\| < 1$. These with (41) entail

$$\sup_{\mathbf{u}\in I_0}\bigl|\hat{\Delta}_+(\mathbf{u}) - \Delta_+(\mathbf{u})\bigr| = o_p\bigl(n^{-2/5}\bigr), \qquad \sup_{\mathbf{u}\in I}\bigl|\hat{\Delta}_+(\mathbf{u}) - \Delta_+(\mathbf{u})\bigr| = O_p\bigl(n^{-2/5}\bigr),$$

so that

$$\hat{\Delta}_{jk}(u) = \Delta_{jk}(u) + n^{-2/5}C_{jk} + r_{jk}(u) \tag{42}$$

for some random variables C_{jk} such that $\sum_{j=1}^{d}\sum_{k=1}^{L_j} C_{jk} = o_p(1)$. Below we prove $C_{jk} = o_p(1)$ for all j and k. This establishes (40).

From the definition of $\hat{\Delta}_{jk}$, its expansion at (42) and the constraints for $\hat{f}_{jk}$ at (14), we have

$$\begin{aligned} 0 = {} & \int_0^1 f_{jk}(u)\hat{p}^I_{jk}(u)\,du + h_{jk}\int_0^1 \frac{a_{jk}(u)}{\mu_{0,jk}(u)}\hat{p}^I_{jk}(u)\,du \\ & + \frac{1}{2}h^2_{jk}\mu_2\int_0^1 f''_{jk}(u)p^I_{jk}(u)\,du + \int_0^1 \Delta_{jk}(u)p^I_{jk}(u)\,du + n^{-2/5}C_{jk} + o_p\bigl(n^{-2/5}\bigr). \end{aligned} \tag{43}$$

Here, we have used Lemma 1 and the fact that $\sup_{u\in[0,1]}|\Delta_{jk}(u)| = O(n^{-2/5})$. Using $\int K_{h_{jk}}(u, v)\,du = 1$ for all $v \in [0, 1]$, we get

$$\begin{aligned} \int_0^1 f_{jk}(u)\hat{p}^I_{jk}(u)\,du = {} & n^{-1}\sum_{i=1}^{n}\mathbb{I}^i\int_0^1 \bigl[f_{jk}(u) - f_{jk}\bigl(\xi^i_{jk}\bigr)\bigr]K_{h_{jk}}\bigl(u, \hat{\xi}^i_{jk}\bigr)\,du/\hat{p}^I_0 \\ & + n^{-1}\sum_{i=1}^{n}\mathbb{I}^i f_{jk}\bigl(\xi^i_{jk}\bigr)/\hat{p}^I_0. \end{aligned}$$

The second term on the right-hand side of the above equation is of order $n^{-(\beta-1)/(2\beta)}$. This is due to the constraint (9), $n^{-1}\sum_{i=1}^{n}|\mathbb{I}^i - \mathbb{I}^{*i}| = O_p(n^{-(\beta-1)/(2\beta)})$ and

$$n^{-1}\sum_{i=1}^{n}\mathbb{I}^{*i}\big[f_{jk}\big(\xi_{jk}^i\big) - E\big(f_{jk}\big(\xi_{jk}^i\big)|\boldsymbol{\xi}^i \in I\big)\big] = O_p\big(n^{-1/2}\big).$$

For the first term, denoted by **VI**, we get

$$\begin{aligned}\mathbf{VI} &= n^{-1}\sum_{i=1}^{n}\mathbb{I}^{*i}\int_0^1\big[f_{jk}(u) - f_{jk}\big(\xi_{jk}^i\big)\big]K_{h_{jk}}\big(u,\xi_{jk}^i\big)\,du/\hat{p}_0^{*I} + o_p\big(n^{-2/5}\big)\\ &= \int_0^1\big[f_{jk}(u) - f_{jk}(v)\big]K_{h_{jk}}(u,v)p_{jk}^I(v)\,dv\,du + o_p\big(n^{-2/5}\big)\\ &= -h_{jk}\int_0^1 a_{jk}(u)p_{jk}^I(u)\,du - \frac{1}{2}h_{jk}^2\mu_2\int_0^1 f_{jk}''(u)p_{jk}(u)\,du\\ &\quad - h_{jk}^2\mu_2\int_0^1 f_{jk}'(u)\frac{\partial}{\partial u}p_{jk}^I(u)\,du + o_p\big(n^{-2/5}\big).\end{aligned}$$

In the first approximation of $\mathbf{V}I$, we have used Proposition 1 and Lemma 1. We also have

$$h_{jk}\int_0^1\frac{a_{jk}(u)}{\mu_{0,jk}(u)}\hat{p}_{jk}^I(u)\,du = h_{jk}\int_0^1 a_{jk}(u)p_{jk}^I(u)\,du + o_p\big(n^{-2/5}\big).$$

These approximations of the terms in (43) and the constraint of Δ_{jk} at (30) give $C_{jk} = o_p(1)$.

Acknowledgements

Kyunghee Han and Byeong U. Park research supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF- 2015R1A2A1A05001753). Hans-Georg Müller research supported by NSF Grants DMS-11-04426, DMS-12-28369 and DMS-14-07852.

References

[1] Bosq, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. New York: Springer. MR1783138

[2] Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

[3] Carroll, R.J., Maity, A., Mammen, E. and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika* **96** 383–398. MR2507150

[4] Chen, D., Hall, P. and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.* **39** 1720–1747. MR2850218

[5] Fan, Y., James, G.M. and Radchenko, P. (2013). Functional additive regression. *Ann. Statist.* **40** 2296–2325. MR3396986

[6] Fanaee-T, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *J. Prog. Artif. Intell.* **2** 113–127.

[7] Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *TEST* **22** 278–292. MR3062258

[8] Ferraty, F., Goia, A., Salinelli, E. and Vieu, P. (2013). Functional projection pursuit regression. *TEST* **22** 293–320. MR3062259

[9] Hildebrandt, T., Bissantz, N. and Dette, H. (2014). Additive inverse regression models with convolution-type operators. *Electron. J. Stat.* **8** 1–40. MR3161732

[10] James, G.M. and Silverman, B.W. (2005). Functional adaptive model estimation. *J. Amer. Statist. Assoc.* **100** 565–576. MR2160560

[11] Jiang, C.-R. and Wang, J.-L. (2011). Functional single index model. *Ann. Statist.* **39** 362–388. MR2797850

[12] Kneip, A. and Utikal, K.J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.* **96** 519–542. MR1946423

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

[14] Lee, Y.K., Mammen, E. and Park, B.U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048

[15] Ma, S. (2014). Estimation and inference in functional single-index models. *Ann. Inst. Statist. Math.* **68** 181–208 1–28. MR3440219

[16] 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

[17] McLean, M.W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014). Functional generalized additive models. *J. Comput. Graph. Statist.* **23** 249–269. MR3173770

[18] Müller, H.-G., Wu, Y. and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika* **100** 607–622. MR3094440

[19] Müller, H.-G. and Yao, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202

[20] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. New York: Springer. MR2168993

[21] Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

[22] Yao, F. and Müller, H.-G. (2010). Functional quadratic regression. *Biometrika* **97** 49–64. MR2594416

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

[24] Zhang, X., Park, B.U. and Wang, J.-L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108** 983–998. MR3174678

[25] Zhu, H., Yao, F. and Zhang, H.H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 581–603. MR3210729

Received January 2016 and revised June 2016