Estimation of Sparse Functional Additive Models with Adaptive Group LASSO

Peijun Sang, Liangliang Wang and Jiguo Cao*

Department of Statistics and Actuarial Science

Simon Fraser University

Abstract: We study a flexible model to tackle the issue of lack of fit in the conventional functional linear regression. This model, called the sparse functional additive model, is used to characterize the relationship between a functional predictor and a scalar response of interest. The effect of the functional predictor is represented in a nonparametric additive form, where the arguments are the scaled functional principal component scores. Component selection and smoothing are considered when fitting the model to reduce the variability and enhance the prediction accuracy, while providing an adequate fit. To achieve these goals, we propose using the adaptive group LASSO method to select relevant components and smoothing splines to obtain a smoother estimate of those relevant components. Simulation studies show that the proposed estimation method compares favourably with various conventional methods in terms of prediction accuracy and component selection. The advantage of our estimation method is further demonstrated in two real data examples.

Key words and phrases: Functional Data Analysis; Functional Linear Model; Functional Principal Component Analysis; Group LASSO; Smoothing Spline

Corresponding author. Email: jiguo_cao@sfu.ca

1. Introduction

Functional data analysis has become an important tool for dealing with data collected over multiple time points, spatial locations, or other continua. A fundamental problem in functional data analysis is how to model the relationship between a scalar response of interest and a functional predictor. For instance, the Tecator data (see Section 5.1) consists of 240 meat samples; each of them comprises the spectrum of absorbance, and three contents: water, fat and protein. Researchers have been concerned about how to use the spectrum of absorbance, which can be treated as a functional predictor, to predict one of the three contents. Functional linear regression (FLR) is a conventional and interpretable model for predicting a scalar response from a functional predictor. It has many interesting applications. For instance, Ainsworth *et al.* (2011) applied FLR to explore the effect of river flow on the decline of sockeye salmon. Luo *et al.* (2013) applied FLR to o investigate the effect of the time-varying admission intensity to Emergency Department access block.

In FLR, the relationship between a scalar response and a functional predictor is modelled in a linear form. Hence the key problem in fitting FLR is to estimate the coefficient function of the functional predictor. There has been extensive research to address this problem. For example, Müller and

Stadtmüller (2005) considered representing the coefficient function in terms of Fourier basis functions or the eigenfunctions of the estimated covariance function of the functional predictor. The coefficients of the Fourier basis functions were then obtained from solving a functional estimating equation. Ramsay and Silverman (2005) suggested using spline basis functions to represent the coefficient function and then solving a regularized regression problem, in which the roughness of the spline representation is penalized to obtain a smooth estimate of the coefficient function. Lin *et al.* (2017) proposed a local sparse estimator for the coefficient function to enhance the interpretability of FLR. Liu *et al.* (2017) added a random effect on the coefficient function when repeated measurements are available on multiple subjects. A comprehensive introduction to FLR can be found in Horváth and Kokoszka (2012) and Morris (2015).

Although the studies aforementioned have proposed various estimating methods to fit a FLR model and established some appealing properties of the corresponding estimators, in practice, applications of FLR are sometimes restricted due to its simple linear form. Similar to the multiple linear model, which in some cases may not adequately describe the relationship between a scalar response and scalar covariates, FLR can also suffer from an inadequate flexibility for modelling the relationship between a scalar response and a functional predictor. This phenomenon has been noted by many researchers. For instance, Yao and Müller (2010) extended the FLR model to the case when the scalar response depends on a polynomial of the functional predictor and they mainly focused on the quadratic case. Chen *et al.* (2011) considered using a nonparametric link to connect the scalar response and the functional linear form. A class of flexible functional nonlinear regression models were proposed by Müller *et al.* (2013) by using continuously additive models to characterize the relationship between a functional predictor and a scalar response. Nonlinear and/or nonparametric functional regression models can somewhat address the issue of inadequate fit caused by FLR (see Chen *et al.*, 2011, Müller *et al.*, 2013, Müller and Yao, 2008). However, they can lead to other issues such as over-flexibility and a lack of stability (Zhu *et al.*, 2014). Reiss *et al.* (2017) summarized some of main approaches of regressing a scalar response on a functional predictor. In this paper, we aim to propose a functional regression model which can achieve a satisfactory trade-off between flexibility and simplicity.

Zhu *et al.* (2014) proposed an extended functional additive model, in which the scalar response of interest depends on a transformation of the leading functional principal component (FPC) scores. They assumed that some additive components were vanishing and the nonvanishing components were smooth functions for the sake of simplicity and interpretability while retaining flexibility. To achieve this goal, they adopted the regularization scheme of component selection and smoothing operator (COSSO) proposed by Lin and Zhang (2006), which can select and smooth components simultaneously. While this model turns out to achieve a better trade-off between flexibility and simplicity compared with many other functional regression models, the estimation procedure seems to suffer from several drawbacks. First, only estimation consistency is guaranteed for the proposed estimator. Whether selection consistency holds for this estimator remains an open problem. Another drawback is associated with computational complexity. As noted by Zhang and Lin (2006), when a full basis is employed, the complexity of the algorithm is $O(n^3)$, where n is the sample size. To reduce the computational burden, Zhang and Lin (2006) suggested using a subset basis algorithm instead, which was computationally much more efficient than the full basis algorithm. Zhu *et al.* (2014) seemed to ignore this computational issue when implementing COSSO to fit the proposed model. The computational complexity is demonstrated in simulation studies.

To overcome the drawbacks of the method proposed by Zhu *et al.* (2014), we propose a method to estimate the extended functional additive model. In contrast to representing nonparametric additive components in the framework of RKHS (Zhu *et al.*, 2014), we use B-spline basis functions to represent these components, which are easier to understand and implement. Then selecting nonzero components is equivalent to selecting nonzero coefficients of the B-spline basis functions. The group LASSO method (Yuan and Lin, 2006) has been shown to perform well when selecting grouped variables for accurate prediction in both theory and application. Because an additive component corresponds to a vector of coefficients, which can be treated as a group of variables, we employ the group LASSO method to select nonzero vectors of coefficients. The adaptive group LASSO method is then applied to allow for different shrinkages for different vectors of coefficients. This modification can yield a more accurate estimate of the coefficient vectors, which then leads to a better estimate for the additive components. This method enables us to attain our goal of obtaining a parsimonious model via component selection.

Nevertheless, the estimated nonzero components can be wiggly due to the representation of the additive components using a large number of B-spline basis functions. It may impair predictive performance, which will be demonstrated in simulation studies in Section 4. Thus we suggest refining the selected components via smoothing splines. This extra smoothing step can improve the prediction accuracy of the estimator obtained from the adaptive group LASSO, which will be demonstrated in our simulation studies.

This article makes three main contributions. First, compared with traditional FLR models, our proposed model provides a better trade-off between flexibility and simplicity in modelling the effect of a functional predictor. By selecting and smoothing nonzero components, our proposed method obtains an estimator which has better prediction accuracy. Second, unlike the COSSO regularization scheme adopted in Zhu *et al.* (2014), we employ group LASSO to select components and use the smoothing spline method to smooth nonzero components. As a result, our proposed estimation method is easy to understand and implement. Last but not least, we give both theoretical and empirical demonstrations of the selection consistency and estimation consistency of our proposed estimator, while Zhu *et al.* (2014) only provided a theoretical proof of the estimation consistency of their estimator.

The remainder of this paper is organized as follows. Section 2 introduces a sparse functional additive model and our method to estimate the additive components in the model. Section 3 establishes the selection consistency and the estimation consistency of our proposed estimator. The finite-sample performance of the estimator is investigated empirically in

Section 4, where we conduct simulation studies to compare our proposed estimator and other conventional methods. In Section 5, our method is demonstrated by analyzing two real data examples. In Section 6, we give some conclusions about our method. The procedures to estimate the FPC scores, proofs of the main results in Section 3 and additional empirical studies are provided in the supplementary document.

2. Model and Estimation Method

2.1 Sparse Functional Additive Model

Suppose that $\{X_i(t), y_i\}_{i=1}^n$ are independent and identically distributed (iid) observations from $\{X(t), Y\}$, where $X(t)$ is a random function and Y is a scalar random variable. We assume $X(t)$ to be a square integrable stochastic process over a compact interval $\mathcal{I} = [0, T]$, i.e., $\mathrm{E}\left\{\int_{\mathcal{I}} X^2(t)dt\right\} < \infty$. Let $m(t)$ and $G(s,t)$ denote the mean function and covariance function of $X(t)$, respectively. According to Mercer's theorem, $G(s,t)$ can be represented as $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where λ_k is a nonnegative eigenvalue and $\phi_k(t)$ is the corresponding eigenfunction. For the sake of identifiability, we postulate that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Additionally, $\{\phi_k\}_{k=1}^{\infty}$ is assumed to be a complete orthonormal basis of the space $L^2(\mathcal{I})$, the collection of all square integrable functions on $\mathcal{I}$. Then the stochastic process $X(t)$ admits the Karhunen-Loève expansion:

$$X(t) = m(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \tag{2.1}$$

where $\xi_k = \int_{\mathcal{I}} (X(t) - m(t)) \phi_k(t) dt,\ k = 1, \ldots$, is called the k-th FPC score. The FPC score satisfies $\mathrm{E}\left(\xi_k \xi_{k'}\right) = \lambda_k$ if $k = k'$ and 0 otherwise.

In FLR, Y is treated as the response and $X(t)$ is the functional predictor. Furthermore, the relationship between Y and $X(t)$ is modelled in a linear form:

$$y_i = \int_{\mathcal{I}} X_i(t) b(t) dt + \epsilon_i,$$

where ϵ_i's denote random errors with mean 0 and variance σ_ϵ^2. Given the representation of $X(t)$ in (2.1), we have $y_i = a + \sum_{k=1}^{\infty} b_k \xi_{ik} + \epsilon_i$, where $a = \int_{\mathcal{I}} m(t) b(t) dt$, ξ_{ik} denotes the k-th FPC score of $X_i(t)$, and $b_k = \int_{\mathcal{I}} \phi_k(t) b(t) dt,\ k \geq 1$. To address the curse of dimensionality, a truncated model is usually adopted such that Y only depends on the first d FPC scores. In other words, we get a truncated linear model: $y_i = a + \sum_{j=1}^{d} b_j \xi_{ij} + \epsilon_i$. In practice, d is chosen as the smallest number of FPCs which can explain over 99.9% of the total variability of the functional predictor $X(t)$. As noted by Zhu *et al.* (2014), this choice can, to some extent, circumvent neglecting those FPC scores which play a negligible role in capturing the variability of the functional predictor but are relevant in predicting the response. This truncated model is slightly restrictive since an explicit parametric form is assumed between the response and the leading FPC scores. The linearity assumption is likely to be violated in substantial practical scenarios.

In light of the idea proposed by Hastie and Tibshirani (1986), and the fact that ξ_j's are mutually uncorrelated, a nonparametric functional additive model was proposed by Müller and Yao (2008) to describe the relationship between the response and the first d FPC scores

$$y_i = a + \sum_{j=1}^{d} f_j(\xi_{ij}) + \epsilon_i, \tag{2.2}$$

where we call f_j the j-th component in the nonparametric functional additive model.

FPC scores usually cannot be observed directly. Therefore we need firstly to estimate FPC scores from the observed functional data which may be subject to measurement errors. We assume that $W_{ij} = X_i(t_{ij}) + e_{ij}$, where W_{ij} denotes the observation of the process $X_i(t)$ made at time point $t_{ij},\ j = 1, \ldots, N_i,\ i = 1, \ldots, n$ and e_{ij} denotes the measurement error and is assumed to be independent of $X_i(t)$. Functional principal component analysis (FPCA) is implemented to estimate FPC scores, denoted by $\hat{\xi}_{ij}$'s. The details of this procedure can be found in the supplementary document.

We first scale the FPC scores to [0, 1] via a transformation function F. One possible strategy is to apply the cumulative distribution function (cdf)

$F(z|\lambda_j)$ of the Normal$(0, \lambda_j)$ on ξ_j, where λ_j is the eigenvalue of the covariance function $G(s,t)$, and $\lambda_j = Var(\xi_j)$. We define ζ_j to be the j-th scaled FPC score: $\zeta_j = F(\xi_j|\lambda_j),\ j = 1, \dots, d$. Then the estimated scaled FPC scores are given as $\hat{\zeta}_{ij} = F(\hat{\xi}_{ij}|\hat{\lambda}_j), j = 1, \dots, d, i = 1, \dots, n$. Assumption B in Section 3 gives more general conditions on the transformation function, F. We still use f_j for the j-th component in the nonparametric functional additive model when ξ_j's are replaced by ζ_j's.

The nonparametric functional additive model (2.2) can now be expressed as

$$y_i = a + \sum_{j=1}^{d} f_j(\zeta_{ij}) + \epsilon_i. \tag{2.3}$$

To make the model identifiable, we assume that $\mathrm{E}\{f_j(\zeta_j)\} = 0,\ j = 1, \dots, d$. Models with a parsimonious structure are preferable in practice. Thus we assume that some components, f_j's are vanishing and the rest of the components are nonzero and smooth. Model (2.3) is called a sparse functional additive model in this article.

B-spline functions, due to their nice properties (De Boor, 2001), have been widely used in estimating unknown functions (see Stone, 1985, Stone, 1986, Huang *et al.*, 2010, etc). In this paper we also employ B-spline functions to estimate the additive components in Model (2.3). We present here a brief overview of B-splines. For more information, see De Boor (2001).

Let $0 = \tau_0 < \tau_1 < \cdots < \tau_{L_n} < \tau_{L_n+1} = 1$ be the breakpoints which separate the interval [0, 1] into $L_n + 1$ subintervals. We assume that $L_n = O(n^{\alpha})$, where $0 < \alpha < 0.5$, and define $\delta_n = \max_{0 \le m \le L_n} |\tau_{m+1} - \tau_m| = O(n^{-\alpha})$. Let c_1 be a constant, independent of n, such that $\delta_n < c_1 \min_{0 \le m \le L_n} |\tau_{m+1} - \tau_m|$. Let $\mathscr{S}_n$ be the space of polynomial splines of order l, which is one more than the degree of polynomials, on [0, 1] consisting of functions s satisfying: (i) s is a polynomial of order l at each subinterval $[\tau_m, \tau_{m+1}], m = 0, \ldots, L_n$; (ii) for $0 \le l^{\star} \le l - 2$, the $l^{\star}$-th order derivative of s is continuous on [0, 1]. Then there exist $m_n = L_n + l$ normalized B-spline basis functions $\{B_k, 1 \le k \le m_n\}$ bounded by 0 and 1 on [0, 1], such that any $f \in \mathscr{S}_n$ can be written as

$$f_j(x) = \sum_{k=1}^{m_n} \beta_{jk} B_k(x), \quad j = 1, 2, \ldots, d, \tag{2.4}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})'$ is the spline coefficient vector. Now selecting nonzero components $f_j(\cdot)$ for Model (2.3) amounts to selecting nonzero $\boldsymbol{\beta}_j$.

2.2 Group LASSO

Accounting for the fact that $\mathrm{E}\{f_j(\zeta_j)\} = 0,\ j = 1, \ldots, d$, we define $\psi_{jk}(x) = B_k(x) - \frac{1}{n}\sum_{i=1}^{n} B_k(\hat{\zeta}_{ij}),\ k = 1, \ldots, m_n,\ j = 1, \ldots, d$. For brevity, $\psi_{jk}(x)$ is denoted by $\psi_k(x)$ without causing any confusion. Thus $\sum_{i=1}^{n} \psi_k(\hat{\zeta}_{ij}) = 0,\ j = 1, \ldots, d$. The estimated intercept in Model (2.3) is given as $\bar{y} =$

$\frac{1}{n}\sum_{i=1}^{n} y_i$. Let $\boldsymbol{Z}_{ij} = (\psi_1(\hat{\zeta}_{ij}), \ldots, \psi_{m_n}(\hat{\zeta}_{ij}))^T$, $\boldsymbol{Z}_j = (\boldsymbol{Z}_{1j}, \ldots, \boldsymbol{Z}_{nj})^T$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_d)$. Similarly, define $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T)^T$, where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})^T$, and $\boldsymbol{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$. Nonzero $\boldsymbol{\beta}_j$'s in Model (2.3) can be selected and estimated using the group LASSO (Yuan and Lin, 2006), in which the corresponding estimate $\tilde{\boldsymbol{\beta}}$ minimizes

$$D_1(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{d} ||\boldsymbol{\beta}_j||_2. \tag{2.5}$$

In (2.5), the positive tuning parameter λ_1 determines the magnitude of shrinkage and $||\cdot||_2$ denotes the Euclidean norm of a vector in $\mathbb{R}^{m_n}$. If $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jm_n})^T$, then the corresponding estimate of f_j is denoted by $\tilde{f}_j$, which equals to $\sum_{k=1}^{m_n} \tilde{\beta}_{jk}\psi_k(x)$. Cross validation is employed to choose an "optimal" λ_1, which is chosen to minimize the cross-validation error.

2.3 Adaptive Group LASSO

The group LASSO method penalizes each $\boldsymbol{\beta}_j$ equally in (2.5), which may not be an optimal treatment. To account for different impact on the response of different ζ_j's, we propose an adaptive group LASSO method, which is similar in spirit to the adaptive LASSO method proposed by Zou (2006). More explicitly, a weight vector $(w_1, \ldots w_d)$, which can produce different shrinkages for different $\boldsymbol{\beta}_j$'s, is introduced. Given $\tilde{\boldsymbol{\beta}}$ estimated from group LASSO, for $j = 1, \ldots, d$, w_j is set to be $||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$ if $||\tilde{\boldsymbol{\beta}}_j||_2 > 0$ and ∞

otherwise. Then the adaptive group LASSO estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}$, is obtained by minimizing

$$D_2(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_2 \sum_{j=1}^{d} w_j ||\boldsymbol{\beta}_j||_2, \tag{2.6}$$

where λ_2 denotes a penalty parameter that can be determined by cross validation. Then the corresponding estimate of $f_j(x)$, denoted by $\hat{f}_j(x)$, can be represented in terms of $\boldsymbol{\psi}_j(x) = (\psi_{j1}(x), \ldots, \psi_{jm_n}(x))^T$, i.e., $\hat{f}_j(x) = \widehat{\boldsymbol{\beta}}_j^T \boldsymbol{\psi}_j(x),\ j = 1, \ldots d$. If $\widehat{\boldsymbol{\beta}}_j = \mathbf{0}$ for some j, then the estimate, $\hat{f}_j$, is also zero.

2.4 Smoothing Spline Method

When a large number of B-spline basis functions are employed to estimate f_j, then the adaptive group LASSO estimate may be wiggly. Further smoothing for nonzero estimates obtained from adaptive group LASSO is indispensable if it is the case. This concern was also discussed in Wu *et al.* (2014). To allow for different roughness penalties for different nonzero components, we propose a smoothing spline method. The weight is defined as $w_j = ||\widehat{\boldsymbol{\beta}}_j||_2^{-1}$, where $j \in S$, and $S = \{j : \widehat{\boldsymbol{\beta}}_j \neq \mathbf{0}\}$ is the set of nonzero components. In particular, the updated estimate of $\boldsymbol{\beta}_j$ is obtained from the smoothing spline method by minimizing

$$D_3(\boldsymbol{\beta}) = (\boldsymbol{y} - \sum_{j \in S} \boldsymbol{Z}_j \boldsymbol{\beta}_j)^T(\boldsymbol{y} - \sum_{j \in S} \boldsymbol{Z}_j \boldsymbol{\beta}_j) + \lambda_3 \sum_{j \in S} w_j \int_0^1 \{f_j^{''}(\zeta_j)\}^2 d\zeta_j, \tag{2.7}$$

where λ_3 denotes the smoothing parameter. The roughness penalty term $\int_0^1 \{f_j^{''}(\zeta_j)\}^2 d\zeta_j = \boldsymbol{\beta}_j^T \boldsymbol{Q}_j \boldsymbol{\beta}_j$, where $\boldsymbol{Q}_j$ is an $m_n \times m_n$ penalty matrix with the pq-th element $\boldsymbol{Q}_j^{pq} = \int_0^1 B_p^{''}(\zeta_j) B_q^{''}(\zeta_j) d\zeta_j$. When the second derivative of $f_j(\zeta_j)$ does not exist, the penalty matrix $\boldsymbol{Q}_j$ can be replaced by the difference matrix introduced by Eilers and Marx (1996). Minimization of (2.7) is equivalent to a classical smoothing spline problem except that there is a weight vector in this problem. Let $\sum_{j \in S} \boldsymbol{Z}_j \boldsymbol{\beta}_j = \boldsymbol{Z}_S \boldsymbol{\beta}$, where $\boldsymbol{Z}_S = (\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_{|S|}}) \in \mathbb{R}^{n \times m_n |S|}$, $i_1, \ldots, i_{|S|}$ are all elements of S and $|S|$ denotes the cardinality of the set S. Let $\boldsymbol{Q} = \text{diag}(w_{i_1} \boldsymbol{Q}_{i_1}, \ldots, w_{i_{|S|}} \boldsymbol{Q}_{i_{|S|}})$. Then the estimate of $\boldsymbol{\beta}$, still denoted by $\widehat{\boldsymbol{\beta}}$, is given as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}_S^T \boldsymbol{Z}_S + \lambda_3 \boldsymbol{Q})^{-1} \boldsymbol{Z}_S^T \boldsymbol{y}$. The corresponding estimate of f_j is $\hat{f}_j = \widehat{\boldsymbol{\beta}}_j^T \boldsymbol{\psi}_j(x), \ j \in S$.

The smoothing parameter λ_3 can be determined by the generalized cross-validation (GCV) measure. For a given λ_3, the corresponding measure can be expressed as

$$\text{GCV}(\lambda_3) = \frac{n \cdot \text{SSE}}{(n - df(\lambda_3))^2},$$

where $\text{SSE} = (\boldsymbol{y} - \boldsymbol{Z}_S \widehat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{Z}_S \widehat{\boldsymbol{\beta}})$ and $df(\lambda_3) = \text{trace}(\boldsymbol{Z}_S (\boldsymbol{Z}_S^T \boldsymbol{Z}_S + \lambda_3 \boldsymbol{Q})^{-1} \boldsymbol{Z}_S^T)$. The optimal smoothing parameter is chosen to minimize the GCV measure. Our whole estimating procedure is called Components Selection and Smoothing in a sparse Functional Additive Model (CSS-FAM in short) in this article.

Remark: Note that Model 2.3 and the corresponding estimation scheme presented above only account for the effect of a single functional predictor on a scalar response. Examples show that incorporating scalar predictors is likely to improve prediction accuracy in practice (Sang *et al.*, 2018; Wong *et al.*, 2018). Therefore, when prediction of a scalar response is the main goal, it would be desirable to incorporate both scalar predictors and multiple functional predictors into the current model structure using adaptive group LASSO and smoothing spline for estimation. Actually we can follow the idea in Sang *et al.* (2018) to extend the proposed framework to allow for scalar covariates and multiple functional predictors.

3. Theoretical Properties

To ensure that the estimated scaled FPC scores, $\hat{\boldsymbol{\zeta}}$, are consistent estimators of the true scaled FPC scores, we need to impose some regularity conditions on the design of the functional predictor $X(t)$. The following conditions follow Zhu *et al.* (2014). As stated in Section 2.1, $\{t_{ij}, j = 1, \ldots, N_i; i = 1, \ldots, n\} \subset \mathcal{I}$ denote the time points when the functional predictor $X_i(t)$ is observed. We assume that $t_{i0} = 0$ and $t_{iN_i} = T$ for each $X_i(t)$. Let $\mathcal{I}_\tau = [-\tau, T + \tau]$ for some $\tau > 0$, and let h_i and $K(\cdot)$ denote the bandwidth and the kernel function used in smoothing the i-th trajectory, respectively. Note that the same kernel function is employed in the local linear smoother for each trajectory, when estimating FPC scores. Below is a list of regularity conditions which can guarantee that the estimated FPC scores and eigenvalues of the covariance function of $X(t)$ converge in probability to the corresponding population values.

Assumption A

(A1) $X(t)$ has a continuous second derivative on $\mathcal{I}_d$ with probability 1 and for $k = 0, 2$, $\int \mathrm{E}\{X^{(k)}(t)\}^4 dt < \infty$. The measurement errors e_{ij}'s of $X_i(t)$ satisfy $\mathrm{E}\,(e_{ij}^4) < \infty$ and they are identically and independently distributed.

(A2) We define T_n to be the lower bound of the number of observations for each trajectory $X_i(t)$. As $n \to \infty$, $T_n \to \infty$. Let $\triangle_i$ denote the largest time difference of two consecutive observations for each trajectory $X_i(t)$, i.e., $\triangle_i = \max\{t_{ij} - t_{i,j-1} : j = 1, \ldots, N_i\}$. The maximal value of these satisfies $\max_i \triangle_i = O(T_n^{-1})$.

(A3) There is a sequence $b_n \to 0$, such that $c_1 b_n \leq \min_i h_i \leq \max_i h_i \leq c_2 b_n$ for some constants $c_2 \geq c_1 > 0$ as $n \to \infty$. In addition, b_n and T_n satisfy that $(T_n b_n)^{-1} + b_n^4 + T_n^{-2} = O(n^{-1})$.

(A4) The kernel function $K(\cdot)$ has a compact support and satisfies $|K(s) - K(t)| \leq C|s - t|$ for $s,\ t$ in its domain and some positive constant

C.

For Model (2.3), let A_1 and A_0 denote the set of non-vanishing and vanishing components, respectively; i.e., $A_1 = \{j : f_j \neq 0, j = 1, \ldots, d\}$ and $A_0 = \{j : f_j \equiv 0, 1 \leq j \leq d\}$. Regarding the transformation function $F(x|\lambda)$, a cdf with variance λ, we make the following assumptions.

Assumption B

(B1) The transformation function $F(x|\lambda)$ is differentiable at x and λ. Furthermore, there exist a positive constant C and a negative constant γ, such that $\frac{\partial F(x|\lambda)}{\partial x} \leq C\lambda^{\gamma}$ and $\frac{\partial F(x|\lambda)}{\partial \lambda} \leq C\lambda^{\gamma}|x|$.

(B2) The cdf of each scaled score ζ_j is absolutely continuous and there exist positive constants C_1 and C_2 such that the probability density function of ζ_j, g_j, satisfies $C_1 \leq g_j(x) \leq C_2$ for $x \in [0, 1]$ and $j \in A_1$.

Assumption (B1) is from Zhu *et al.* (2014) as well. Together with Assumptions (A1)-(A4), it can guarantee that the $\hat{\zeta}_j$ is a consistent estimator of ζ_j, $1 \leq j \leq d$. Assumption (B2) is a standard assumption in nonparametric additive models according to Stone (1985).

Define $||f||_2 = \{\int_0^1 f^2(x)dx\}^{1/2}$ whenever the integral is finite. Let $L > 0$, r be a nonnegative integer, and $\nu \in (0, 1]$ such that $\rho = r + \nu > 0.5$. Let $\mathscr{F}$ be the class of functions h on $[0, 1]$ whose r-th derivative exists and satisfies the Hölder condition with exponent ν: $|h^{(r)}(s) - h^{(r)}(t)| \leq L|s - t|^{\nu}$ for any $0 \leq s, t \leq 1$. Other standard assumptions for additive nonparametric models (see Huang *et al.*, 2010) include:

Assumption C

(C1) $\min_{j \in A_1} ||f_j|| \geq c_f$ for some $c_f > 0$.

(C2) The random variables $\epsilon_1, \ldots \epsilon_n$ are iid with mean 0 and variance σ_ϵ^2. Furthermore, the tail probability satisfies $P(|\epsilon_1| > x) \leq K \exp(-Cx^2)$, for $\forall x \geq 0$ and some constants C and K.

(C3) $\mathrm{E}\{f_j(\zeta_j)\} = 0$ and $f_j \in \mathscr{F}$, $j \in A_1$.

The following proposition explains why it is reasonable to employ B-spline functions to approximate each nonparametric component f_j in Model (2.3). To guarantee that B-spline functions in $\mathscr{S}_n$ can provide a satisfactory approximation of functions in $\mathscr{F}$, throughout the article we assume that l, the order of polynomial functions in $\mathscr{S}_n$, satisfies $l > \max\{r, 1\}$. Write the centered version of $\mathscr{S}_n$ as

$$\mathscr{S}_{nj}^0 = \left\{ f_{nj} : f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} \psi_k(x), (\beta_{j1}, \ldots, \beta_{jm_n}) \in \mathbb{R}^{m_n} \right\},\ j = 1, \ldots, d,$$

where ψ_k's are the centered spline basis functions defined in Section 2.2.

Proposition 1. Suppose that $f \in \mathscr{F}$ and $\mathrm{E}\, f(\zeta_j) = 0$. Then under Assumptions A and B, there exists an $f_{nj} \in \mathscr{S}_{nj}^0$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \{f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 = O_p(m_n^{-2\rho} + n^{-1}).$$

if $m_n = O(n^{\alpha})$ with $0 < \alpha < 0.5$.

Let $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_{m_n}(x))^T$ for $x \in [0,1]$. Proposition 1 implies that, uniformly over $j \in \{1, \ldots, d\}$, there exists $\boldsymbol{\beta}_j \in \mathbb{R}^{m_n}$, such that $\frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{\beta}_j^T\boldsymbol{\psi}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\}^2 = O_p(m_n^{-2\rho} + n^{-1})$ under Assumptions A and B, provided $m_n = O(n^{\alpha})$. Furthermore, we can take $\boldsymbol{\beta}_j = \mathbf{0}$ for $j \in A_0$. Denote $\{j : \tilde{\boldsymbol{\beta}}_j \neq \mathbf{0}\}$ and $\{j : \tilde{\boldsymbol{\beta}}_j = \mathbf{0}\}$ as $\tilde{A}_1$ and $\tilde{A}_0$, respectively. Theorem 1 establishes the selection consistency and estimation consistency of $\tilde{\boldsymbol{\beta}}_j$'s obtained from the group LASSO step.

Theorem 1. Suppose that Assumptions A, B and C hold and $\lambda_1 \geq C\sqrt{n\log(m_n)}$ for some sufficiently large constant C. Then it follows that

(i) If $m_n \to \infty$ as $n \to \infty$ with rate satisfying $m_n = o(n^{1/6})$ and $(\lambda_1^2 m_n^2)/n^2 \to 0$ as $n \to \infty$, then all the nonzero $\boldsymbol{\beta}_j$, $j \in A_1$, are selected with probability converging to 1.

(ii) If $m_n = o(n^{1/6})$, then $\sum_{j=1}^{d}||\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j||_2^2 = O_p\left(\frac{m_n^2\log m_n}{n}\right) + O_p\left(\frac{m_n^2\lambda_1^2}{n^2}\right)$ $+ O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho-1}}\right)$.

Theorem 2 further illustrates that the estimated functions obtained from the group LASSO step, $\tilde{f}_j$'s, also enjoy selection consistency and estimation consistency.

Theorem 2. Suppose that Assumptions A, B and C hold and $\lambda_1 \geq C\sqrt{n\log(m_n)}$ for some sufficiently large constant C. Then we have

(i) If $m_n \to \infty$ as $n \to \infty$ with rate satisfying $m_n = o(n^{1/6})$ and $(\lambda_1^2 m_n)/n^2 \to 0$ as $n \to \infty$, then in the group LASSO step, all the nonzero additive components f_j's, $j \in A_1$, are selected with probability converging to 1.

(ii) If $m_n = o(n^{1/6})$, then $||\tilde{f}_j - f_j||_2^2 = O_p\left(\frac{m_n \log m_n}{n}\right) + O_p\left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho}}\right) + O_p\left(\frac{m_n \lambda_1^2}{n^2} + \frac{m_n}{n}\right)$, $j \in A_1 \cup \tilde{A}_1$.

For two (positive) sequences $\{a_n\}$ and $\{b_n\}$, if $\frac{a_n}{b_n}$ is bounded away from 0 and ∞, then denote $a_n \sim b_n$. The following corollary can be directly derived from Theorem 2.

Corollary 1. Suppose that Assumptions A, B and C hold. If $m_n \sim n^{1/(2\rho+1)}$ and $\lambda_1 \sim \sqrt{n \log(m_n)}$, then

(i) If $\rho > \frac{5}{2}$, then in the group LASSO step, all the nonzero additive components f_j, $j \in A_1$, are selected with probability converging to 1.

(ii) If $\rho > \frac{5}{2}$, then $||\tilde{f}_j - f_j||_2^2 = O_p(n^{-2\rho/(2\rho+1)} \log m_n), \qquad j \in A_1 \cup \tilde{A}_1.$

Theorem 3 states that the adaptive group LASSO yields an estimate which is also consistent in both selection and estimation. Furthermore, it illustrates that this estimate compares favorably with that given by the group LASSO with respect to estimation accuracy.

Theorem 3. Suppose that Assumptions A, B and C hold and $m_n \sim n^{1/(2\rho+1)}$, where $\rho > 5/2$. If the tuning parameters satisfy $\lambda_1 \sim \sqrt{n \log(m_n)}$,

$\lambda_2 \leq O(n^{\frac{1}{2}})$, $\frac{\lambda_2}{n^{(8\rho+3)/(8\rho+4)}} = o(1)$ and $\frac{n^{1/(4\rho+2)}\sqrt{\log(m_n)}}{\lambda_2} = o(1)$, then we have

(i) With probability approaching 1, the nonzero components, i.e., $\{f_j, j \in A_1\}$ are selected and $||\hat{f}_j||_2 = 0,\ j \in A_0$.

(ii) $\sum_{j\in A_1} ||\hat{f}_j - f_j||_2^2 = O_p(n^{-2\rho/(2\rho+1)})$.

4. Simulation Studies

In this section we use simulated examples to illustrate several properties of our proposed estimator, and compare our method with several conventional methods commonly used in practice.

We simulate data as follows. In each simulation replicate, we generate n curves and the observations are made at $m = 200$ equally spaced points in [0, 10]. In our simulation studies, we set $n = 100$ or 500. To accommodate measurement errors, the observation at t_j $(j = 1, \ldots, m)$ is generated as $W_{ij} = X_i(t_j) + e_{ij}$, where $\{X_i(t)\}_{i=1}^n$ are i.i.d samples of a stochastic process $X(t)$ and e_{ij} are i.i.d normals with mean 0 and variance 0.1. For $k = 1, \ldots, 20$, let $\lambda_k = 31.5 \times 0.6^k$ denote the k-th eigenvalue of the covariance function of $X(t)$. The corresponding k-th eigenfunction is the k-th Fourier basis function, denoted by $\phi_k(t)$. Then $X_i(t) = m(t) + \sum_{k=1}^{20} \xi_{ik}\phi_k(t)$, where $m(t) = t + \sin t$ denotes the mean function of $X(t)$ and $\{\xi_{ik}\}_{k=1}^{20}$ are independently sampled from $N(0, \lambda_k)$.

The scaled score ζ_{ik} is defined as the uniform score of ξ_{ik}, i.e., $\zeta_{ik} = \Phi(\xi_{ik}/\sqrt{\lambda_k}), k = 1, \ldots, 20,\ i = 1, \ldots, n$, where Φ denotes the cdf of a standard normal distribution. The response variable is generated from Model (2.3): $y_i = a + f_1(\zeta_{i1}) + f_2(\zeta_{i2}) + f_4(\zeta_{i4}) + \epsilon_i$. We set the true intercept to $a = 1.2$, and the true components to $f_1(x) = x\exp(x) - 1$, $f_2(x) = \cos(2\pi x)$ and $f_4(x) = 3\left(x - \frac{1}{4}\right)^2 - \frac{7}{16},\ x \in [0, 1]$. The random errors ϵ_i's are independently sampled from a normal distribution with mean 0 and variance 0.67. The signal-to-noise ratio is defined as $\mathrm{Var}\left\{f_1(\zeta_1) + f_2(\zeta_2) + f_4(\zeta_4)\right\}/\mathrm{Var}\,(\epsilon)$, and we set the signal-to-noise ratio to be approximately 2. We estimate the model by fitting n randomly generated training observations, and evaluate its performance on 200 randomly generated test observations. The simulation is implemented for 100 simulation replicates. Simulation results for $n = 200$ and 300 are presented in the supplementary document.

Besides our proposed method CSS-FAM, we also fit the data with three conventional models including MARS (Friedman *et al.*, 2001), two extended functional additive models (FAM) proposed by Müller and Yao (2008) and the component selection and estimation for the functional additive model (CSE-FAM) in Zhu *et al.* (2014). More specifically, MARS is fitted using the function **earth** in the R package **earth** and the variables which enter the final model are examined by the function **evimp**. In the first extended

FAM, denoted by FAM, the response variable y is fitted with a multiple linear regression where the covariates are $f_1(\hat{\zeta}_1)$, $f_2(\hat{\zeta}_2)$ and $f_4(\hat{\zeta}_4)$. In other words, FAM assumes to know the true model structure with three true covariates $f_1(\hat{\zeta}_1)$, $f_2(\hat{\zeta}_2)$ and $f_4(\hat{\zeta}_4)$. The second extended FAM, denoted by S-FAM, considers a saturated model to incorporate the first d FPC scores such that they can explain over 99.9% of the total variability in the smoothed sample curves. The value of d is 15, 16 or 17 in all simulation replicates. We employ the function **gam** in the R package **mgcv** to fit such a model in which the arguments of additive components are $\hat{\zeta}_j$, $j = 1, \ldots, d$. Then p-values of all terms in the model are available from the function **summary.gam**. Only the significant nonparametric components (p-value < 0.05) are retained in computing the true positive (TP) rate and the false positive (FP) rate. We also consider an alternative method for estimating Model (2.3) by only using the two steps of group LASSO and adaptive group LASSO, which is denoted by AGL-FAM.

Table 1 summarizes the comparison between these six methods in 1000 simulation replicates. It suggests that compared with CSE-FAM, CSS-FAM has similar performance in prediction when the sample size $n = 100$ and 500. Both of them outperform the other three methods except FAM in prediction accuracy, and are slightly inferior compared with FAM, which assumes the true components are known. This suggests that the extra adaptive smoothing spline step can increase the prediction accuracy when adaptive group LASSO yields a wiggly estimate. Concerning the quality of estimating nonparametric components, CSS-FAM can rival CSE-FAM as well, since both of them yield estimates which are reasonably close to the true nonparametric components. In addition, the residual sum of squares (RSS) for each component estimated using CSS-FAM is much smaller than that using AGL-FAM, indicating that smoothing spline enables us to obtain a smoother and more accurate estimate of nonparametric components.

Statistics	n	Methods					
		MARS	FAM	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
MSPE	100	1.15 (.25)	.92 (.17)	1.15 (.21)	1.00 (.19)	1.25 (0.23)	1.01 (0.20)
	500	.78 (.09)	.73 (.08)	.77 (.09)	.74 (.08)	.80 (.09)	.74 (.08)
RSS($\hat{f}_1$)	100	-	2.6 (4.9)	3.6 (4.7)	2.6 (4.1)	13.4 (6.9)	3.8 (5.7)
($\times 10^{-2}$)	500	-	0.4 (0.5)	0.6 (0.6)	0.5 (0.4)	2.5 (0.9)	0.6 (0.4)
RSS($\hat{f}_2$)	100	-	6.8 (10.5)	11.2 (10.0)	18.1 (13.6)	12.8 (6.8)	8.1 (13.8)
($\times 10^{-2}$)	500	-	0.5 (0.7)	1.9 (1.3)	2.9 (1.5)	3.1 (1.4)	1.9 (1.3)
RSS($\hat{f}_4$)	100	-	6.7 (10.3)	4.0 (3.3)	4.6 (5.7)	14.3 (7.2)	5.9 (11.2)
($\times 10^{-2}$)	500	-	0.7 (1.1)	0.7 (0.5)	0.5 (0.4)	2.3 (1.0)	0.5 (0.7)
TP%	100	99.1 (.05)	-	98.2 (.08)	95.7 (.12)	94.7 (.17)	94.7 (.10)
	500	100 (.00)	-	100 (.0)	100 (.0)	100 (.0)	100 (.0)
FP%	100	20.4 (.12)	-	13.7 (.11)	3.8 (.07)	0.9 (.03)	0.9 (.03)
	500	29.0 (.14)	-	8.9 (.08)	3.0 (.07)	$<$ 0.01 (.003)	$<$0.01 (.003)
Time	100	.01 (.03)	$<$ 0.01	.39 (.21)	2.87 (.23)	0.48 (.04)	2.40 (.10)
(seconds)	500	.02 (.06)	$<$ 0.01	2.88 (2.28)	117.2 (5.91)	3.57 (0.27)	11.4 (2.77)

Table 1: Summary statistics for evaluating six methods. MSPE refers to the mean squared prediction error on the test data; the residual sum of squares (RSS) for each estimated component $\hat{f}_j$ is defined as: $\text{RSS}(\hat{f}_j) = \int_0^1 (\hat{f}_j(x) - f_j(x))^2 dx$; TP% and FP% stand for the true positive and false positive rates in percentage, respectively. The point estimate for each measure is averaged over 100 simulation replicates, and the corresponding estimated standard error is given in parenthesis.

Table 1 also compares these methods from the perspective of variable selection, where the true positive (TP) rate and the false positive (FP) rate are employed for assessment. Combined with Table S1 in the supplementary document, we find that although CSS-FAM and AGL-FAM perform slightly worse than the other models in correctly selecting nonzero variables for a relatively small sample ($n = 100$ or 200), this tiny gap vanishes when the sample size increases ($n = 300$ or larger). Furthermore, CSS-FAM and AGL-FAM dominate the other methods in not selecting irrelevant components, regardless of how large or small the sample size is. The other methods mistakenly select irrelevant variables substantially more often than CSS-FAM or AGL-FAM, especially when the sample size is relatively small.

The computational time for each method is recorded in Table 1 as well. Obviously CSE-FAM is the most computationally intensive method if a full basis is employed. This is a serious issue in implementations, particularly when the sample size is large, as mentioned in Section 1. In comparison, the proposed method, CSS-FAM, can still be implemented within 12 seconds even when the training data set consists of 500 curves.

Figure S1 in the supplementary document illustrates the estimation details for one randomly selected simulation replicate when the number of curves is $n = 500$. After estimating the scaled FPC score, we fit group

LASSO on the training data, as shown in (2.5). The top left panel in Figure S1 describes how the 5-fold cross-validation error changes with λ_1. The optimal λ_1 is chosen to minimize the 5-fold cross-validation error. Like the top left one, the top middle panel explains how to choose the optimal λ_2 for the adaptive group LASSO step in (2.6) based on 5-fold cross validation. The top right panel shows how to choose the optimal smoothing parameter (λ_3) by minimizing GCV in the smoothing spline step. The bottom three panels in Figure S1 illustrate the effects of the extra smoothing spline step on the estimation of the nonparametric components after using adaptive group LASSO. The adaptive group LASSO method may lead to an excessively wiggly estimate for each nonzero nonparametric component. Smoothing spline can control the roughness appropriately and hence yield a smoother and more accurate estimate.

5. Applications

In this section, we fit the sparse functional additive model (2.3) using our proposed method (CSS-FAM), together with several conventional models considered in the simulation studies, to analyze two real data sets. Application to the air pollution data is introduced in the supplementary document. Besides the models considered in the simulation, we fit a multiple linear model to investigate whether a functional linear model can adequately characterize the relationship between the scalar response and the functional predictor in these two examples. The covariates in the multiple linear model are the first d FPC scores. We choose the truncation level d in the same way as for Model (2.2). This multiple linear model is actually a special case of Model (2.2): each additive component taking a linear form. LASSO (Tibshirani, 1996) is implemented when fitting the mulitple linear model in these two examples to obtain a more parsimonious model and reduce variability. This estimating method is called LAF in this paper. In the air pollution data, the trajectories of the functional predictor for some subjects are sparsely observed. In contrast, in the Tecator data, the functional predictor is regularly spaced and densely observed across all subjects. In each example, we randomly divide the whole data set into a training set and a test set, and the training set is used to fit each model while the test set is used for evaluation. All these models are compared with respect to the mean squared prediction errors calculated on the test set.

5.1 Tecator Data

The Tecator data are recorded for 240 meat samples on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. Each of them consists of a

100-channel spectrum of absorbance, and the percentages of three components of the meat: moisture (water), fat and protein. The spectrum records the negative base 10 logarithm of the transmittance measured by the spectrometer. The percentage of three meat components are determined by analytic chemistry. As demonstrated by a large body of research (see Vila *et al.*, 2000, Goldsmith and Scheipl, 2014, Zhu *et al.*, 2014), the spectrum of absorbance is highly predictive of the percentage of these three meat components. Figure S2 in the supplementary document depicts the trajectories of the spectrum of absorbance of the 240 meat samples. We aim to study the effect of the spectral trajectories of the meat sample on the protein content by using the sparse functional additive model (2.3).

The protein content, denoted by Y, is the response variable of primary interest; the functional predictor $X(t)$ denotes the spectrum of absorbance. FPCA is implemented to estimate FPC scores and then to obtain the scaled FPC scores, denoted by $\hat{\boldsymbol{\zeta}} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_d)$. Zhu *et al.* (2014) suggested that the first $d = 20$ should be retained in order to achieve satisfactory prediction accuracy, even though the first 10 FPCs explain more than 99.9% of total variability in the smoothed sample curves. To compare the performance of various methods with respect to the prediction accuracy, the 240 meat samples are divided into a training sample and a test sample. According to the dataset's original description (`http://lib.stat.cmu.edu/datasets/tecator`), the 240 meat samples have been divided into three parts: the training sample consists of the 172 meat samples, the following 43 meat samples form the test samples and the last 25 meat samples are for extrapolation use and should be ignored. We, however, randomly choose 187 meat samples to train the model; the test set comprises the remaining 53 meat samples.

Methods	MARS	LAF	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
MSPE	0.99	0.66	0.56	0.55	0.92	0.51

Table 2: Mean squared prediction errors (MSPEs) on the test data for six methods.

The comparison among the six models with respect to prediction accuracy is shown in Table 2. Obviously CSS-FAM outperforms the other methods in terms of prediction. In particular, the difference between CSS-FAM and LAF implies that a linear model cannot adequately characterize the relationship between the protein content and the spectrum of absorbance of the meat samples. CSS-FAM can nevertheless achieve a better trade-off between flexibility and simplicity compared with other methods. Additionally, the poor performance of AGL-CSS, especially when compared with

CSS-FAM, suggests that the extra smoothing spline step in the proposed algorithm considerably enhances prediction accuracy.

In AGL-FAM, 10 cubic B-spline basis functions are employed to represent the nonparametric components in the sparse functional additive model (2.3). A 5-fold cross validation suggests that $\lambda_1 = 0.002$ is an optimal choice of the penalty parameter in the group LASSO step and $\lambda_2 = 0.011$ minimizes the 5-fold cross-validation error in the adaptive LASSO step. As a result, 14 non-vanishing components, $\{\hat{f}_1, \ldots, \hat{f}_9, \hat{f}_{11}, \hat{f}_{16}, \hat{f}_{17}, \hat{f}_{19}, \hat{f}_{20}\}$, are selected from the 20 components. This finding is slightly inconsistent with the conclusion drawn in Zhu *et al.* (2014), where they claimed that $\{\hat{f}_1, \ldots, \hat{f}_8, \hat{f}_{10}, \hat{f}_{13}, \hat{f}_{16}, \hat{f}_{17}\}$ are non-vanishing components. To refine these estimated components, smoothing spline is employed and the optimal choice of smoothing parameter, $\lambda_3 = 0.001$, is chosen to minimize the GCV measure.

6. Conclusions and Discussion

Compared with traditional FLR, the sparse functional additive model (2.3) proposed in this article provides a more flexible description of the relationship between a scalar response and a functional predictor. To achieve sparseness, we employ the group LASSO penalty to select and estimate nonzero components in the nonparametric additive model, thereby reducing variability and enhancing interpretability.

The estimation procedure consists of several important techniques. FPCA is employed to estimate FPC scores and eigenvalues of the covariance function of the functional predictor. Then we use B-spline basis functions to represent the nonparametric additive components in the sparse functional additive model (2.3). The use of the group LASSO penalty enables us to achieve the goal of selecting and estimating nonzero components. To obtain a better estimate of the coefficient vectors, we adopt the idea of the adaptive group LASSO to provide different shrinkages for different components. Considering the fact that the estimated components given by the adaptive LASSO may not be smooth, since a large number of B-spline basis functions are used to represent them, we propose using smoothing splines to further refine the estimated nonzero components obtained from the group LASSO step. Simulation studies demonstrate that this smoothing step can improve both estimation of the additive components and prediction of the response.

In contrast to other methods, we theoretically justify that our proposed estimator enjoys both selection consistency and estimation consistency. These consistency results are also demonstrated by simulation studies. Two real data applications show that the proposed model together with the estimating method provides an appealing tool in predicting a scalar response from a functional predictor, compared with other conventional methods.

Even though this article discusses only regressing a scalar response on a functional covariate, the methodology can be extended to accommodate other scenarios. For example, this framework can be extended to explore the relationship between a scalar response, whose distribution belongs to the exponential family, and a functional predictor. In addition, in this work the truncation level d, such that the first d FPCs can explain over 99.9% of total variability in the functional predictor, is assumed to be fixed. From the theoretical perspective, it is worthwhile to investigate the properties of the corresponding estimator when d is allowed to increase with the sample size in future work.

Supplementary Materials

The supplementary document outlines the procedure of estimating FPC scores introduced in Section 2.1, provides proofs of theoretical results in Section 3 and includes additional simulation results in Section 4 and one application example in Section 5. The R codes for our real data analysis and the simulation studies can be downloaded at `https://github.com/caojiguo/fam`.

Acknowledgements

The authors are very grateful for the constructive comments of the Editor, the AE and two reviewers, which are extremely helpful for us to improve our work. This research was supported by Discovery grants of Wang and Cao from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Ainsworth, L. M., Routledge, R., and Cao, J. (2011). Functional data analysis in ecosystem research: the decline of oweekeno lake sockeye salmon and wannock river flow. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**(2), 282–300.

Chen, D., Hall, P., Müller, H.-G., *et al.* (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, **39**(3), 1720–1747.

De Boor, C. (2001). *A practical guide to splines*. New York: Springer-Verlag.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–102.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Berlin: Springer.

Goldsmith, J. and Scheipl, F. (2014). Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis*, **70**, 362–372.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, **1**(3), 297–310.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. New York: Springer.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**(4), 2282.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**(5), 2272–2297.

Lin, Z., Cao, J., Wang, L., and Wang, H. (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, **26**, 306–318.

Liu, B., Wang, L., and Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Computational Statistics & Data Analysis*, **106**, 153–164.

Luo, W., Cao, J., Gallagher, M., and Wiles, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in Medicine*, **32**, 2681–2694.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, **33**(2), 774–805.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statis-*

tical Association, **103**(484), 1534–1544.

Müller, H.-G., Wu, Y., and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, **100**(3), 607–622.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis 2nd edition*. New York: Springer-Verlag.

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, **85**(2), 228–249.

Sang, P., Lockhart, R. A., and Cao, J. (2018). Sparse estimation for functional semiparametric additive models. *Journal of Multivariate Analysis*. DOI: 10.1016/j.jmva.2018.06.010.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2), 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**(2), 590–606.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Vila, J.-P., Wagner, V., and Neveu, P. (2000). Bayesian nonlinear model selection and neural networks: a conjugate prior approach. *IEEE Transactions on neural networks*, **11**(2), 265–278.

Wong, R. K., Li, Y., and Zhu, Z. (2018). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*. DOI:

10.1080/01621459.2017.1411268.

Wu, H., Lu, T., Xue, H., and Liang, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, **109**(506), 700–716.

Yao, F. and Müller, H.-G. (2010). Functional quadratic regression. *Biometrika*, **97**(1), 49–64.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.

Zhang, H. H. and Lin, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica*, **16**(3), 1021–1041.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3), 581–603.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.