

# Structured functional additive regression in reproducing kernel Hilbert spaces

Hongxiao Zhu

*Duke University, Durham, USA*

Fang Yao

*University of Toronto, Toronto, Canada*

Hao Helen Zhang

*North Carolina State University, Raleigh, USA*

**Summary.** Functional additive models (FAMs) provide a flexible yet simple framework for regressions involving functional predictors. The utilization of data-driven basis in an additive rather than linear structure naturally extends the classical functional linear model. However, the critical issue of selecting nonlinear additive components has been less studied. In this work, we propose a new regularization framework for the structure estimation in the context of Reproducing Kernel Hilbert Spaces. The proposed approach takes advantage of the functional principal components which greatly facilitates the implementation and the theoretical analysis. The selection and estimation are achieved by penalized least squares using a penalty which encourages the sparse structure of the additive components. Theoretical properties such as the rate of convergence are investigated. The empirical performance is demonstrated through simulation studies and a real data application.

**Keywords:** Component selection; Additive models; Functional data analysis; Smoothing spline; Principal components; Reproducing kernel Hilbert space.

## 1. Introduction

Large complex data collected in modern science and technology impose tremendous challenges on traditional statistical methods due to their high-dimensionality, massive volume and complicated structures. Emerged as a promising field, functional data analysis (FDA) employs random functions as model units and is designed to model data distributed over continua such as time, space and wavelength. See Ramsay and Silverman (2005) for a comprehensive introduction. Such data may be viewed as realizations of latent or observed stochastic processes and are commonly encountered in many fields, e.g. longitudinal studies, microarray experiments, brain images.

Regression models involving functional objects play a major role in the FDA literature. The most widely used is the functional linear model, in which a scalar response  $Y$  is regressed on a functional predictor  $X$  through a linear operator

$$E(Y|X) = \int_{\mathcal{T}} X(t)\beta(t)dt, \quad (1)$$

where  $X(t)$  is often assumed to be a smooth and square-integrable random function defined on a compact domain  $\mathcal{T}$ , and  $\beta(t)$  is the regression parameter function which is also assumed to be smooth and square-integrable. A commonly adopted approach for fitting model (1) is

through basis expansion, i.e. representing the functional predictor as the linear combinations of a set of basis  $\{\phi_k\}$ :  $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$ , where  $\mu(t) = EX(t)$ . Model (1) is then transformed to a linear form of the coefficients  $\{\xi_k, k = 1, 2, \dots\}$ :  $E(Y|X) = b_0 + \sum_{k=1}^{\infty} \xi_k b_k$ , where  $b_0 = \int_{\mathcal{T}} \beta(t) \mu(t) dt$  and  $b_k = \int_{\mathcal{T}} \beta(t) \phi_k(t) dt$ . More references on functional linear regression can be found in Cardot et al. (1999, 2003), Fan and Zhang (2000), etc. Extensions to generalized functional linear models were proposed by James (2002), Müller and Stadtmüller (2005) and Li et al. (2010). The basis set  $\{\phi_k\}$  can be either predetermined (e.g. Fourier basis, wavelets, B-splines), or data-driven. One convenient choice for the latter is the eigenbases of the auto-covariance operator of  $X$ , in which case the random coefficients  $\xi_k$  are called functional principal component (FPC) scores. The FPC scores have mean zero and variances equal to the corresponding eigenvalues  $\{\lambda_k, k = 1, 2, \dots\}$ . This isomorphic representation of  $X$  is referred to as the Karhunen-Loève expansion, and the related methods are often called functional principal component analysis (FPCA) (Rice and Silverman, 1991; Yao et al., 2005; Yao, 2007; Hall et al., 2006; Hall and Hosseini-Nasab, 2006). Due to the rapid decay of the eigenvalues, the orthogonal eigenbases provide a more parsimonious and efficient representation compared to other bases. Furthermore, FPC scores are mutually uncorrelated, which can considerably simplify the model fitting and theoretical analysis. We focus on the FPC representation of the functional regression throughout this paper, nevertheless, the proposal is also applicable to other prespecified basis.

Although widely used, the linear relationship can be restrictive for general applications. This linear assumption is then relaxed by Müller and Yao (2008) who proposed the functional additive models (FAM). The FAM provides a flexible yet practical framework that accommodates nonlinear associations and at the same time avoids the curse of dimensionality encountered in high dimensional nonparametric regression problems (Hastie and Tibshirani, 1990). In the case of the scalar response, the linear structure was replaced by the addition of nonlinear functional components, i.e.

$$E(Y|X) = b_0 + \sum_{k=1}^{\infty} f_k(\xi_k), \quad (2)$$

where  $\{f_k(\cdot)\}$  are unknown smooth functions. In Müller and Yao (2008), the FAM was fitted by estimating the  $\{\xi_k\}$  using FPCA (Yao et al., 2005) and estimating  $\{f_k\}$  using local polynomial smoothing.

Apparently regularizing (2) is necessary. In Müller and Yao (2008) the regularization was achieved by truncating the eigen-sequence to the first  $K$  leading components, where  $K$  was chosen to explain the majority of the total variation in the predictor  $X$ . Despite its simplicity, this naive truncation procedure can be inadequate in many complex problems. Firstly, the impact of FPCs on the response does not necessarily coincide with their magnitudes specified by the auto-covariance operator of the predictor process alone. For instance, some higher order FPCs may contribute to the regression significantly more than the leading FPCs. Secondly, although a small number of leading FPCs might be able to capture the major variability in  $X$  due to the rapidly decaying eigenvalues, one often needs to include more components for better regression performance, especially for the prediction purpose as observed in Yao and Müller (2010). On the other hand, retaining more than needed FPCs brings the risk of over-fitting, which is caused by including components that contribute little to the regression but introduce noise. Therefore a desirable strategy is to automatically identify “important” components out of a sufficiently large number of candidates, whereas shrink those “unimportant” ones to zero.

With the above consideration, we seek an entirely new regularization and estimation framework for identifying the sparse structure of the FAM regression. Model selection that encourages sparse structure has gained substantial attention in the last decade mostly due to the rapidly emerging high-dimensional data. In the context of linear regression, the seminal works include Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), SCAD (Fan and Li, 2001) and the references therein, while Lin and Zhang (2006), Meier et al. (2009) and Ravikumar et al. (2009) dealt with traditional additive models. In comparison with these work, the sparse estimation in functional regression is much less explored. To our knowledge, there are only work for functional linear models (James et al., 2009; Zhu et al., 2010). The related research for additive structures is scant in the literature. In this paper, we consider the selection and estimation of the additive components in FAMs that encourages a sparse structure, in the framework of reproducing kernel Hilbert space (RKHS). Unlike in standard additive models, the FPC scores are not directly observed in FAMs. They need to be firstly estimated from the functional covariates and then plugged into the additive model. The estimated scores are random variables, which creates a major challenge to the theoretical exploration. It is necessary to properly take into account the influence of the unobservable FPC scores on the resulting estimator. Furthermore, the functional curve  $X$  is not fully observed either. We typically collect repeated and irregularly spaced sample points, which are subject to measurement errors. The existence of measurement error in data adds extra difficulty for model implementation and inferences. All of these issues are tackled in this paper. We propose a two-step estimation procedure to achieve the desired sparse structure estimation in FAM. For the regularization, we adopt the COSSO (Lin and Zhang, 2006) penalty due to its direct shrinkage effects on functions in the reproducing kernel Hilbert space. On the practical side, the proposed method is simple to implement, by taking advantages of existing algorithms of FPCA.

The rest of the article is organized as follows. In Section 2, we present the proposed approach and algorithm, as well as the theoretical properties of the resulting estimator. Simulation results with comparisons to existing methods are included in Section 3. We apply the proposed method to the Tecator data in Section 4, studying the regression of the protein content on the absorbance spectrum. The concluding remarks are provided in Section 5, while the details of the estimation procedure and technical proofs are deferred to the appendices.

## 2. Structured functional additive model regression

Let  $Y$  be a scalar response associated with a functional predictor  $X(t)$ ,  $t \in \mathcal{T}$ , and let  $\{y_i, x_i(\cdot)\}_{i=1}^n$  be the i.i.d. realizations of the pair  $\{Y, X(\cdot)\}$ . The trajectories  $\{x_i(t) : t \in \mathcal{T}\}$  are often observed intermittently on possibly irregular grids  $\mathbf{t}_i = (t_{i1}, \dots, t_{iN_i})^\top$ . Denote the discretized  $x_i(t)$  in the vector form as  $\mathbf{x}_i = (x_{i1}, \dots, x_{iN_i})^\top$ . To be realistic, we also assume that the trajectories are subject to i.i.d. measurement errors, i.e.  $x_{ij} = x_i(t_{ij}) + e_{ij}$  with  $Ee_{ij} = 0$  and  $\text{Var}(e_{ij}) = \nu^2$ . Following the FPCA of Yao et al. (2005) and Yao (2007), denote  $\boldsymbol{\xi}_{i,\infty} = (\xi_{i1}, \xi_{i2}, \dots)^\top$  as the sequence of FPC scores of  $x_i$ , which is associated with eigenvalues  $\{\lambda_1, \lambda_2, \dots\}$  such that  $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ .

### 2.1. Proposed methodology

As discussed in Section 1, the theory of FPCA enables the isomorphic transformation of random functions to their FPC scores, which brings tremendous convenience to model fit-

ting and theoretical development in functional linear regression. To establish a framework for the nonlinear and nonparametric regression, we consider regressing the scalar responses  $\{y_i\}$  directly on the sequences of FPC scores  $\{\xi_{i,\infty}\}$  of  $\{x_i\}$ . For the convenience of model regularization, we would like to restrict the predictor variables (i.e. FPC scores) to take values on a closed and bounded subset of the real line, e.g.  $[0, 1]$  without loss of generality. This is easy to achieve by taking a transformation of the FPC scores through a monotonic function  $\Psi : \mathbb{R} \rightarrow [0, 1]$ , for all  $k$ . In fact the choice of  $\Psi$  is rather flexible. A wide range of CDF functions can be used (see (a.2) in Section 2.2 for the regularity condition). Additionally one may choose  $\Psi$  so that the transformed variables have similar/same variations. This can be achieved by allowing  $\Psi(\cdot)$  to depend on the eigenvalues  $\{\lambda_k\}$ , where  $\{\lambda_k\}$  serve as scaling variables. For simplicity, in the sequel we use a suitable CDF (e.g. normal), denoted by  $\Psi(\cdot, \lambda_k)$ , from a location-scale family with mean zero and variance  $\lambda_k$ . It is obvious that, if  $\xi_k$ 's are normally distributed, the standard normal CDF leads to uniformly distributed transformed variables on  $[0, 1]$ .

Denoting the transformed variable of  $\xi_{ik}$  by  $\zeta_{ik}$ , i.e.  $\zeta_{ik} = \Psi(\xi_{ik}, \lambda_k)$ , and denoting  $\zeta_{i,\infty} = (\zeta_{i1}, \zeta_{i2}, \dots)^T$ , we propose an additive model as follows:

$$y_i = b_0 + \sum_{k=1}^{\infty} f_{0k}(\zeta_{ik}) + \epsilon_i, \quad (3)$$

where  $\{\epsilon_i\}$  are independent errors with zero mean and variance  $\sigma^2$ , and  $f_0(\zeta_{i,\infty}) = b_0 + \sum_{k=1}^{\infty} f_{0k}(\zeta_{ik})$  is a smooth function. For each  $k$ , let  $H^k$  be the  $l$ th order Sobolev Hilbert space on  $[0, 1]$ , defined by

$$H^k([0, 1]) = \{g \mid g^{(\nu)} \text{ is absolutely continuous for } \nu = 0, 1, \dots, l-1; g^{(l)} \in L_2\}.$$

One can show that  $H^k$  is an reproducing kernel Hilbert space (RKHS) equipped with the norm

$$\|g\|^2 = \sum_{\nu=0}^{l-1} \left\{ \int_0^1 g^{(\nu)}(t) dt \right\}^2 + \int_0^1 \{g^{(l)}(t)\}^2 dt.$$

See Wahba (1990) and Lin and Zhang (2006) for more details. Note that  $H^k$  has the orthogonal decomposition  $H^k = \{1\} \oplus \bar{H}^k$ . Then the additive function  $f_0$  corresponds to  $\mathcal{F}$  which is a direct sum of subspaces, i.e.  $\mathcal{F} = \{1\} \oplus \sum_{k=1}^{\infty} \bar{H}^k$  with  $f_{0k} \in \bar{H}^k$ , for all  $k$ . It is easy to check that, for any  $f = b + \sum_k f_k \in \mathcal{F}$ , we have  $\|f\|^2 = b^2 + \sum_{k=1}^{\infty} \|f_k\|^2$ . In this paper, we take  $l = 2$  but the results can be extended to other cases straightforwardly. To distinguish the Sobolev norm from the  $L^2$  norm, we write  $\|\cdot\|$  for the former and  $\|\cdot\|_{L^2}$  for the latter.

As motivated in Section 1, it is desirable to impose some type of regularization conditions on model (3) to select ‘‘important’’ components. An important assumption commonly made in high-dimensional linear regression is the sparse structure of the underlying true model. This assumption is also critical in the context of functional data analysis, which enables us to develop a systematic strategy than the heuristic truncation that retains the leading FPCs. Although widely adopted, retaining the leading FPCs is a strategy guided solely by the covariance operator of the predictor  $X$ , and therefore it fails to take into account the response  $Y$ . To be more flexible, we assume that the number of important functional additive components that contribute to the response is finite, but not necessarily restricted only to the leading terms. In particular, we use the set  $\mathcal{I}$  to denote the indices of these

components and assume that  $|\mathcal{I}| < \infty$ , where  $|\cdot|$  denotes the cardinality of the index set. In other words, there exists a sufficiently large  $s$  such that  $\mathcal{I} \subseteq \{1, \dots, s\}$ , implying that  $f_k \equiv 0$  as long as  $k > s$ . The FAM is thus equivalent to

$$y_i = b + \sum_{k=1}^s f_{0k}(\zeta_{ik}) + \epsilon_i. \quad (4)$$

It is noticed that the initial truncation  $s$  merely controls the total number of additive components to be considered, which is different from the heuristic truncation suggested by Yao et al. (2005) and Müller and Yao (2008) based on model selection criteria such as cross validation, AIC, or the fraction of the variance explained. In practice we suggest to choose  $s$  large so that nearly 100% of the total variabilities are explained. This often leads to more than 10 FPCs in most empirical cases.

With the above assumption, the regression function  $f_0(\zeta) = b + \sum_{k=1}^s f_{0k}(\zeta_k)$  lies in the truncated subspace  $\mathcal{F}^s = \{1\} \oplus \sum_{k=1}^s \bar{H}^k$  of  $\mathcal{F}$ , where  $\zeta$  is the truncated version of  $\zeta_\infty$ , i.e.  $\zeta = (\zeta_1, \dots, \zeta_s)^T$  with the dependence on  $s$  suppressed if no confusion arises. To nonparametrically regularize the unknown smooth functions  $\{f_{0k}\}$ , we employ the COSSO regularization defined for the function estimation in RKHS and estimate  $f_0$  by finding  $f \in \mathcal{F}^s$  that minimizes

$$Q(f | \zeta_i) = \frac{1}{n} \sum_{i=1}^n \{y_i - f(\zeta_i)\}^2 + \tau_n^2 J(f), \quad \text{with } J(f) = \sum_{k=1}^s \|P^k f\|, \quad (5)$$

where  $P^k f$  is the orthogonal projection of  $f$  onto  $\bar{H}^k$ . Here  $\tau_n$  is the only smoothing parameter that requires tuning, whereas the common smoothing spline approach involves multiple smoothing parameters. The penalty  $J(f)$  is a convex functional and is a pseudonorm in  $\mathcal{F}^s$ . One interesting connection between COSSO and LASSO is that, when  $f_{0k}(\zeta_k) = \zeta_k \beta_{0,k}$ , the penalty in (5) reduces to  $\sum_{k=1}^s |\zeta_k \beta_{0,k}|$ , which becomes the adaptive Lasso penalty (Zou, 2006).

Different from the standard additive regression models, the transformed FPC scores  $\{\zeta_i\}$  serving as the predictor variables in (5) cannot be observed. Therefore we need to estimate the FPC scores first before the estimation and structure selection of  $f$ . A simple two-step algorithm is given as follows.

*Algorithm:*

Step 1. Implement FPCA to estimate the FPC scores  $\{\xi_{i1}, \dots, \xi_{is}\}$  of  $x_i$ , and then the transformed variables  $\hat{\zeta}_{ik} = \Psi(\hat{\xi}_{ik}, \hat{\lambda}_k)$ , where  $\hat{\lambda}_k$  is the estimated eigenvalue, and  $s$  is chosen to explain nearly 100% of the total variation.

Step 2. Implement the COSSO algorithm of Lin and Zhang (2006) to solve

$$\min_{f \in \mathcal{F}^s} Q(f | \hat{\zeta}_i) = \min_{f \in \mathcal{F}^s} \frac{1}{n} \sum_{i=1}^n \{y_i - f(\hat{\zeta}_i)\}^2 + \tau_n^2 J(f), \quad \text{with } J(f) = \sum_{k=1}^s \|P^k f\|. \quad (6)$$

We would like to refer to Appendix A for the details in case of densely or sparsely observed predictor trajectories. We call the proposed method the component selection and estimation for functional additive model, abbreviated as CSE-FAM.

## 2.2. Theoretical properties

We focus on the consistency of the resulting estimator of CSE-FAM for the case when  $\{x_i(t)\}$  are densely observed in this subsection, where the rate of convergence is assessed using the empirical norm. In particular, we introduce the empirical norms and the entropy of  $\mathcal{F}^s$  as follows. Let  $g \in \mathcal{F}^s$ , the empirical norm of  $g$  is defined as  $\|g\|_n = \sqrt{1/n \sum_{i=1}^n g(\zeta_i)^2}$ . The empirical inner product of the error term  $\epsilon$  and  $g$  is defined as  $(\epsilon, g)_n = 1/n \sum_{i=1}^n \epsilon_i g(\zeta_i)$ . Similarly, the empirical inner product of  $f$  and  $g$  in  $\mathcal{F}^s$  is  $(f, g)_n = 1/n \sum_{i=1}^n f(\zeta_i)g(\zeta_i)$ .

The assumptions on the regression function  $f$  and the transformation  $\Psi(\cdot, \cdot)$  are listed below in (a.1)–(a.2), while the commonly adopted regularity conditions on the functional predictors  $\{x_i(t)\}$ , the dense design, and the smoothing procedures are deferred to (b.1)–(b.3) in Appendix B.

- (a.1) For any  $f \in \mathcal{F}^s$ , there exist independent  $\{B_i\}_{i=1}^n$  with  $E(B_i^2) < \infty$ , such that with probability 1,

$$\left| \frac{\partial f(\zeta_i)}{\partial \zeta_{ik}} \right| \leq B_i \|f\|_{L^2}.$$

- (a.2) The transformation function  $\Psi(\xi, \lambda)$  is differentiable at  $\xi$  and  $\lambda$ , and satisfies that  $|\frac{\partial}{\partial \xi} \Psi(\xi, \lambda)| \leq C\lambda^\gamma$  and  $|\frac{\partial}{\partial \lambda} \Psi(\xi, \lambda)| \leq C\lambda^\gamma |\xi|$  for some constant  $C$  and  $\gamma$  ( $\gamma < 0$ ).

The assumption (a.1) is a regularization condition that controls the amount of fluctuation in  $f$  relative its  $L^2$  norm. For (a.2), one can easily verify that, if choosing  $\Psi(\cdot, \cdot)$  to be the normal CDF with zero mean and variance  $\lambda$ , then  $C = 1$  and  $\gamma = -1/2$  (when  $\lambda > 1$ ) or  $\gamma = -3/2$  (when  $0 < \lambda < 1$ ). One can also choose the CDF of student- $t$  or other distributions with variances  $\lambda$ .

For brevity of the presentation, the technical lemmas and proofs are deferred in Appendix B. It is noticed that the existence of the minimizer for the criterion (5) is guaranteed in analogy to Theorem 1 of Lin and Zhang (2006), by considering a design conditional on the input  $\{y_i, \zeta_{i1}, \dots, \zeta_{is}\}$ ,  $i = 1, \dots, n$ , where  $s$  is the initial truncation parameter.

*Theorem 1.* Consider the regression model (4) with  $\zeta_{ik} = \Psi(\xi_{ik}, \lambda_k)$ , where  $\{\xi_{ik}\}_{k=1}^s$  are FPC scores of  $x_i(t)$  based on densely observed trajectories, and  $\{\lambda_k\}_{k=1}^s$  are the corresponding eigenvalues. Let  $\hat{f}$  be the minimizer of the target function (6) over  $f \in \mathcal{F}^s$ , and let  $\tau_n$  be the tuning parameter in (6). Assume that the assumptions (a.1)–(a.2) and (b.1)–(b.3) hold. If  $J(f_0) > 0$  and

$$\tau_n^{-1} = n^{2/5} J^{3/10}(f_0), \quad (7)$$

then  $\|\hat{f} - f_0\|_n = O_p(n^{-2/5}) J^{1/5}(f_0)$  and  $J(\hat{f}) = J(f_0) O_p(1)$ . If  $J(f_0) = 0$  and

$$\tau_n^{-1} = n^{1/4}, \quad (8)$$

then  $\|\hat{f} - f_0\|_n = O_p(n^{-1/2})$  and  $J(\hat{f}) = O_p(n^{-1/2})$ .

It is worth mentioning that the technical difficulty arises from the unobserved variables  $\zeta_i$ , and major effort has been devoted to tackle the influence of the estimated quantities  $\hat{\zeta}_i$  on the resulting estimator by utilizing the analytical tools from the spectral decomposition of the auto-covariance operator of  $X$ . Theorem 1 suggests that, if the repeated measures observed for all individuals are sufficiently dense and  $J(f_0)$  is bounded, the resulting estimator  $\hat{f}$  obtained from (6) has the rate of convergence  $n^{-2/5}$ , which is the same as the rate when  $\{\zeta_i\}$  are directly observed.

### 3. Simulation studies

To demonstrate the performance of the proposed CSE-FAM approach, we conduct simulation studies under different settings. We compare the proposed CSE-FAM approach with the FAM-type methods and the multivariate adaptive regression splines (MARS). The FAM-type methods are implemented in three different ways, two of which are the “oracle” methods:  $\text{FAM}_{O1}$  and  $\text{FAM}_{O2}$ , both assuming full knowledge of the underlying model structure. In particular, the  $\text{FAM}_{O1}$  serves as the gold standard, in which both the true values of  $\{\zeta_{ik}\}$  and the true non-vanishing additive components are used. The  $\text{FAM}_{O2}$  is another type of oracle, in which the values of  $\{\zeta_{ik}\}$  are estimated through FPCA, but the true non-vanishing additive components are used.

Under each experiment setting, we perform 100 Monte Carlo simulations and present the model selection and prediction results for the methods under comparison. We consider the functional data both densely and sparsely observed.

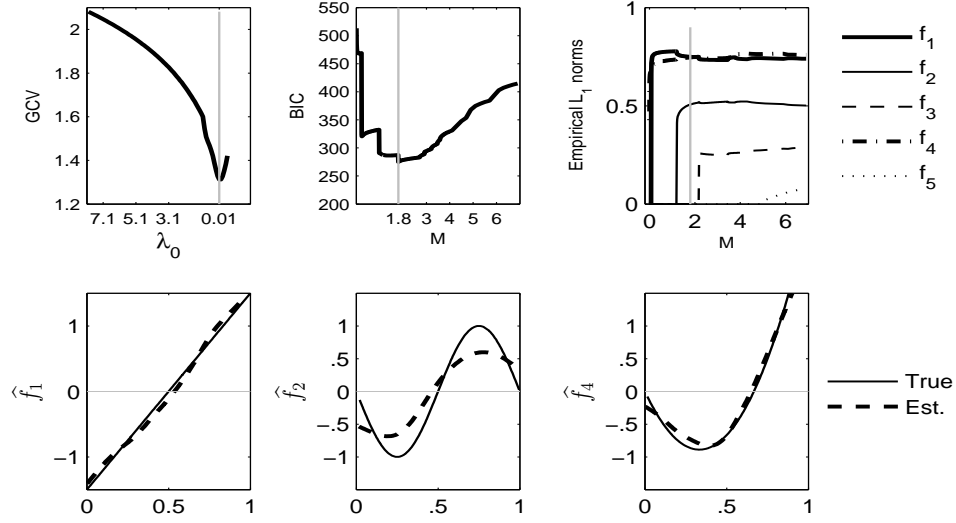
#### 3.1. Dense functional data

We generate 1,000 i.i.d. trajectories using 20 eigenfunctions, among which  $n = 200$  are randomly allocated to the training set and the rest 800 form the test set. The functional predictors  $x_i(t), t \in [0, 10]$ , are measured over a grid with 100 equally spaced points, with independent measurement errors  $e_{ij} \sim N(0, v^2)$ ,  $v^2 = 0.2$ . The eigenvalues of  $x_i(t)$  are generated by  $\lambda_k = ab^{k-1}$  with  $a = 45.25, b = 0.64$ . The true FPC scores  $\{\xi_{ik}\}$  are generated from  $N(0, \lambda_k)$ , and the eigenbasis  $\{\phi_k(\cdot)\}$  are taken to be the first twenty Fourier basis on  $[0, 10]$ , and the mean curve is set to be  $\mu_x(t) = t + \sin(t)$ . We use the normal CDF to obtain the transformed variables:  $\zeta_k = \Psi(\xi_k; 0, \lambda_k), k = 1, \dots, 20$ . The values of  $y_i$  are then generated by  $y_i = f_0(\zeta_i) + \epsilon_i$ , where  $\epsilon_i \sim N(0, \sigma^2)$ . We assume that the  $f_0$  only depends on three nonzero additive components, the first, the second and the fourth, i.e.  $f_0(\zeta_i) = b + f_{01}(\zeta_{i1}) + f_{02}(\zeta_{i2}) + f_{04}(\zeta_{i4}), \mathcal{I} = \{1, 2, 4\}$ . Here we take  $\sigma^2 = 1, b = 1.4, f_{01}(\zeta_1) = 3\zeta_1 - 3/2, f_{02}(\zeta_2) = \sin(2\pi(\zeta_2 - 1/2)), f_{04}(\zeta_4) = 8(\zeta_4 - 1/3)^2 - 8/9$  and  $f_{0k}(\zeta_k) \equiv 0$  for  $k \notin \mathcal{I}$ . This gives the signal-to-noise ratio (SNR) 2.2, where the SNR is defined as  $\text{SNR} = \text{Var}(f_0(\zeta))/\text{Var}(\epsilon)$ , where  $\text{Var}(f_0(\zeta)) = \sum_{k \in \mathcal{I}} \int_0^1 f_{0k}^2(\zeta_k) d\zeta_k = 2.2$  given  $\zeta_k \sim U[0, 1]$ .

We apply the proposed CSE-FAM algorithm to the training data, following the FPCA and COSSO steps described in Section 2.1 and Appendix A. For illustration, we pick one Monte Carlo simulation and display the component selection and estimation results in Figure 1. In FPCA, the initial truncation is  $s = 18$  accounting for nearly 100% of the total variation, and is passed to the COSSO step. The component selection is then achieved by tuning the regularization parameters  $\lambda_0$  in (9) with generalized cross validation (GCV) and  $M$  in (10) with Bayesian information criterion (BIC), illustrated in the top left and top middle panels of Figure 1, while the empirical  $L_1$  norms of  $\hat{f}_k$ , (computed by  $n^{-1} \sum_{i=1}^n |\hat{f}_k(\hat{\zeta}_{ik})|$  at different  $M$ ) are shown on the top right panel. In the bottom panels of Figure 1, the resulting estimates of  $f_k, k = 1, 2, 4$ , are displayed, and  $\{\hat{f}_k, k \neq 1, 2, 4\}$  are shrunk to 0 as desired.

The model selection and prediction results are presented in the top panel of Table 3.1. We implement the FAM procedure in a different manner from that in Müller and Yao (2008). Instead of using local polynomial smoothing for estimating each  $f_k$  separately, we perform a more general additive fitting, the generalized additive model (GAM), on the transformed FPC scores which allows back-fitting and also provides a  $p$ -value for each





**Fig. 1.** The plots of component selection and estimation results for one simulation. Top Left: GCV vs.  $\lambda_0$ . Top middle: BIC vs.  $M$ . Top right: empirical  $L_1$  norms at different  $M$  values. Grey vertical bars on the top panels indicate the tuning parameters chosen. The three bottom panels show the estimated  $f_k$ 's (dashed line) vs. the true (solid line), for  $k = 1, 2, 4$ .

additive components. The only reason for doing so is that the GAM algorithm shows more numerical stability especially when the number of additive components is large. Due to the use of the true model structure, both of the oracle methods  $FAM_{O1}$  and  $FAM_{O2}$  are expected to outperform the rest. Because of the estimation error induced in the FPCA step,  $FAM_{O2}$  is expected to sacrifice certain estimation accuracy and prediction power as compared to  $FAM_{O1}$ . The  $FAM_S$  is the saturated model based on the estimated FPC scores and the leading  $s$  terms used in the CSE-FAM. No model selection is performed in  $FAM_S$ . The  $s$  values vary from 17 to 19 which take into account nearly 100% of the total variabilities of  $\{x_i(t)\}$ . The MARS method is based on Hastie et al. (2001).

It is noticed that the subjective truncation based on the explained variation in  $X$  is sub-optimal for regression purpose (results not reported for conciseness). Therefore, in Table 3.1, we report (under the “model size” column) the counts of selected number of nonvanishing additive components in CSE-FAM, and the counts of the number of significantly nonzero additive components in the case of FAM,  $FAM_{O1}$  and  $FAM_{O2}$ . For display convenience, only the counts for model sizes up to 8 are reported. The “selection frequency” of Table 3.1 records the number of times that each additive component is estimated to be nonzero. For MARS, if the  $j$ th covariate  $\hat{\zeta}_j$  is selected in one or more basis, we counted as 1 and 0 otherwise. Regarding “prediction error”, we used the population estimates from the training set (e.g. mean, covariance and eigenbasis) to get the FPC scores for both training and test set. Then apply the  $\{\hat{f}_k\}$  estimated from the training set to get the prediction for  $\{y_i^*\}$  in the test set. The prediction errors are calculated by  $n^{-1} \sum_{i=1}^n (y_i^* - \hat{y}_i^*)^2$ . From the top panel of Table 3.1, we see that under the dense design, the CSE-FAM choses the correct models (with



**Table 1.** Summary of the model selection and prediction in 100 Monte Carlo simulations under the dense design and sparse design.

Data	Model	Model Size								Selection Frequency								Prediction
		1	2	3	4	5	6	7	8	$\zeta_1$	$\zeta_2$	$\zeta_3$	$\zeta_4$	$\zeta_5$	$\zeta_6$	$\zeta_7$	$\zeta_8$	Error
Dense design	CSE-FAM	0	5	61	29	5	0	0	0	100	94	22	100	7	3	0	1	1.30 (.13)
	FAM <sub>S</sub>	0	0	10	32	21	21	8	4	100	98	51	100	32	14	12	8	1.50 (.17)
	MARS	-	-	-	-	-	-	-	-	100	99	60	100	41	23	25	18	1.46 (.16)
	FAM <sub>O2</sub>	0	1	99	-	-	-	-	-	100	99	-	100	-	-	-	-	1.28 (.12)
	FAM <sub>O1</sub>	0	0	100	-	-	-	-	-	100	100	-	100	-	-	-	-	1.07 (.06)
Sparse design	CSE-FAM	0	22	61	13	4	0	0	0	100	78	10	82	12	9	7	1	2.07 (.16)
	FAM <sub>S</sub>	0	0	14	30	25	20	9	2	100	98	41	96	35	17	9	12	2.17 (.16)
	MARS	-	-	-	-	-	-	-	-	100	98	58	98	56	30	20	23	2.11 (.14)
	FAM <sub>O2</sub>	0	4	96	-	-	-	-	-	100	98	-	98	-	-	-	-	2.01 (.14)
	FAM <sub>O1</sub>	0	0	100	-	-	-	-	-	100	100	-	100	-	-	-	-	1.05 (.05)

model size equals three) 61% of the time whereas the FAM<sub>S</sub> always overselects ( $\alpha = .05$  is used to retain significant additive components). The prediction error of CSE-FAM is the smallest among the three non-oracle models. As compared with the oracle methods, the CSE-FAM has less prediction power than FAM<sub>O2</sub> (slightly) and FAM<sub>O1</sub>, which can be regarded as the price paid by both estimating the  $\zeta$  and selecting the additive components.

To assess the estimation quality, the averaged integrated squared errors (AISE) for the first eight additive components are presented in the top panel of Table 2, where the ISE is defined by  $\text{ISE}(f_k) = E_{\zeta_k} \{\hat{f}_k(\zeta_k) - f_k(\zeta_k)\}^2 = \int_0^1 (\hat{f}_k(t) - f_k(t))^2 dt$ . From the top panel of Table 2, we see that the CSE-FAM provides considerably smaller ISE for the truly zero components ( $f_j$ ,  $j = 1, 3, 6, 7, 8$ ) than the FAM<sub>S</sub>. For the nonzero components, The CSE-FAM, FAM<sub>S</sub> and FAM<sub>O2</sub> have similar AISE values.

### 3.2. Sparse functional data

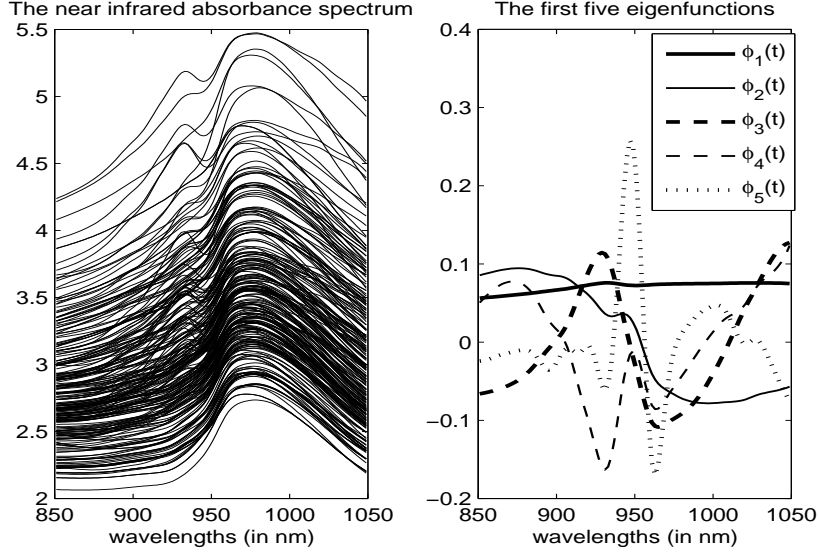
To compare with the dense case, we also conducted a simulation to examine the performance of the CSE-FAM for sparse functional data. We generated 1,200 i.i.d. trajectories, with 300 in training set and 900 in test set. For each trajectory, there are 5 – 10 repeated observations uniformly located on  $[0, 10]$ , with the number of points chosen from 5 to 10 with equal probabilities. The other settings are the same as in the dense design. The summary of the model selection, prediction and estimation results are presented in the bottom panel of Table 3.1 and Table 2. We observe the similar pattern as in the dense design case. Moreover, Table 2 suggests that, for the sparse design, the FAM<sub>S</sub> estimates of  $f_k$  become quite unstable for higher order components (e.g.  $k \geq 7$ ). The AISE increases rapidly due to the influence of outlying estimates. This is not a surprise, because under the sparse design the high-order eigenfunctions and FPC scores are difficult to be estimated accurately due to the sparseness of the data and the moderate sample size, which leads to inaccurate  $f_k$  estimates when the saturated model FAM<sub>S</sub> is used. Under this situation, we see that the proposed CSE-FAM model still performs quite stable, since the COSSO penalty has the effect of automatically down-weighting the “unimportant” components. This provides further support for the proposed CSE-FAM.

## 4. Real data application

We demonstrate the performance of the proposed method through the regression of protein content on the near infrared absorbance spectral measured over 240 meat samples. The dataset is collected by the Tecator company and is publicly available on the StatLib web-

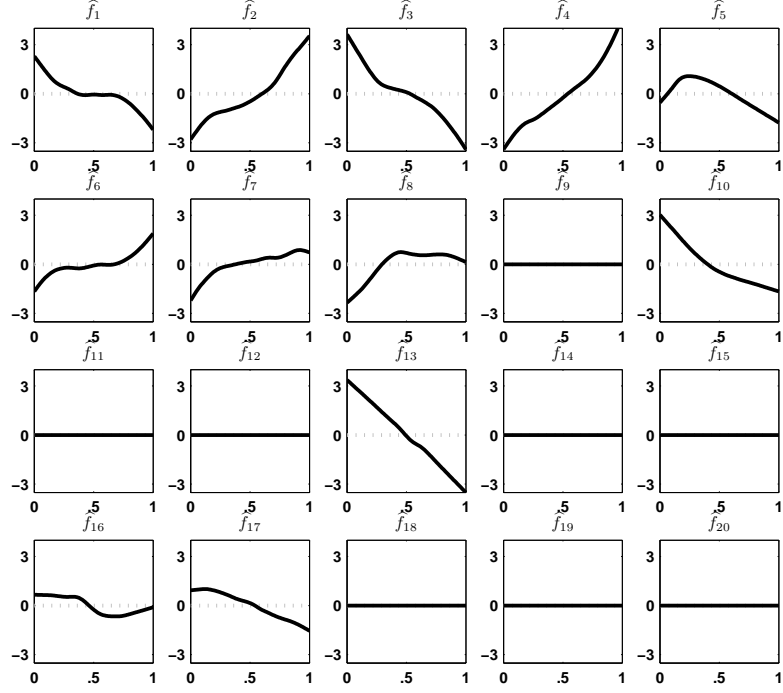
**Table 2.** Averaged ISE for 100 Monte Carlo simulations under the dense and sparse design.

Name		AISE								
		$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$	$f$
Dense design	CSE-FAM	.038	.117	.022	.038	.005	.001	.000	.001	.226
	FAM <sub>S</sub>	.030	.095	.050	.047	.031	.018	.016	.015	.476
	FAM <sub>O2</sub>	.027	.090	-	.042	-	-	-	-	.158
	FAM <sub>O1</sub>	.007	.028	-	.019	-	-	-	-	.054
Sparse design	CSE-FAM	.033	.22	.036	.298	.055	.040	.045	.001	.720
	FAM <sub>S</sub>	.016	.118	.032	.159	.102	.121	.399	2.64	$> 10^3$
	FAM <sub>O2</sub>	.026	.129	-	.220	-	-	-	-	.376
	FAM <sub>O1</sub>	.007	.016	-	.013	-	-	-	-	.036

**Fig. 2.** Plot of the near infrared absorbance spectral curves and the first five estimated eigenfunctions.

site (<http://lib.stat.cmu.edu>). The measurements were made through a spectrometer named Tecator Infratec Food and Feed Analyzer. The spectral curves were recorded at wavelengths ranging from 850nm to 1050nm. For each meat sample the data consist of a 100 channel spectrum of absorbances (100 grid points) as well as the contents of moisture (water), fat and protein. The absorbance is  $-\log_{10}$  of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry. Of primary interest is to predict the protein contents using the spectral trajectories. The 240 meat samples were randomly split into a training set (with 185 samples) and a test set (with 55 samples). We aim to predict the contents of the protein in the test set based on the training data. Figure 2 illustrates the spectral curves and the first five eigenfunctions estimated using FPCA.

We initially retain the first 20 FPCs which take into account nearly 100% of the total variabilities. The proposed CSE-FAM were then applied for component selection and esti-



**Fig. 3.** Plot of the estimated additive components  $\{f_k(\cdot), k = 1, \dots, 5\}$ .

mation. The tuning parameters in the COSSO step are guided by the GCV criterion for  $\lambda_0$  which gives  $\lambda_0 = .0013$ , and by 10-fold cross-validation for  $M$  which gives  $M = 10.0$ . The estimated additive components are plotted in Figure 3, from which the proposed method selects 12 out of the 20 components,  $\{\hat{f}_1, \dots, \hat{f}_8, \hat{f}_{10}, \hat{f}_{13}, \hat{f}_{16}, \hat{f}_{17}\}$ , and the other components are estimated to be zero.

To assess the performance of the proposed method, we report the prediction error (PE) on the test set in Table 3, where the PE is calculated in the same way as in simulations.

We also report the quasi- $R^2$  for the test set, which is defined as  $\hat{R}_Q^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$ . To show the influence of the initial truncation, we also use a small value of  $s$ ,  $s = 10$  in CSE-FAM, which gives suboptimal results. This suggests that we shall use a sufficiently large  $s$  to begin with. The FAM is carried out with the leading 5, 10, 20 FPCs, respectively. An interesting phenomenon is that, though the high-order FPCs (over 10) explain very little variation of the functional predictor (less than 1%), their contribution to the prediction is surprisingly substantial. Such phenomena are also observed for MARS and partial least squares (PLS is a popular approach in chemometrics, see Xu et al. (2007) and the references therein). One more comparison is with the classical functional linear model (FLM) with the estimated leading FPCs served as predictor, where a heuristic AIC is used to choose the first 7 components. From Table 3, we see that, when the initial truncation is set at 10, the proposed CSE-FAM is not obviously advantageous compared to FAM. As

**Table 3.** Prediction results on the test set, as compared with several other methods. Note: PC10 indicates that 10 FPC scores were used. PLD20 indicates the number of PLS directions used were 20. AIC7: 7 FPC scores were used based on the regression AIC criterion.

	CSE-FAM		FAM			MARS	PLS	FLM
	$s = 10$	$s = 20$	PC5	PC10	PC20	PC20	PLD20	AIC7
PE	2.22	0.72	3.98	2.13	0.84	0.77	1.02	1.50
$R_Q^2$	0.82	0.94	0.68	0.83	0.93	0.93	0.92	0.88

the number of FPCs increases to 20, the proposed method provides much smaller PE and higher  $R_Q^2$  than all other methods. A sensible explanation is that, for this data, most of the first 10 FPC components (except the 9th) have nonzero contributions to the response (shown in Figure 3). Therefore penalizing these components does not help to improve the prediction. However, as the number of FPC scores increases, more redundant terms come into play, so the penalized method (CSE-FAM) gains more prediction power.

## 5. Discussion

We proposed a structure estimation method for functional data regression where a scalar response is regressed on a functional predictor. The model is constructed in the framework of FAM, where the additive components are functions of the scaled FPC scores. The selection and estimation of the additive components are performed through penalized least squares using the COSSO penalty in the context of RKHS. The proposed method allows for more general nonparametric relationships between the responses and predictors, therefore serves as an important extension of the functional linear regression. Through the adoption of the additive structures, it avoids the curse of dimensionality caused by the infinite-dimensional predictor process. The proposed method provides a way to select the important features of the predictor processes and simultaneously shrink the unimportant ones to zero. This selection scenario takes into account not only the explained variation of the predictor processes, but also their contributions to the responses. The theoretical result shows that, under the dense design, the nonparametric rate from component selection and estimation will dominate the discrepancy due to the unobservable FPC scores. The COSSO penalty usually requires that  $s < n$ , which does not conflict with the assumption that the initial truncation  $s$  is chosen to be sufficiently large to include all important features, since in practice the number of FPCs accounting for nearly 100% predictor variation is often far less than the sample size  $n$  due to the fast decay of the eigenvalues.

Although we have focused on the FPC-based analysis in this work, the CSE-FAM framework is generally applicable to other basis structure, e.g. splines and wavelets, where the additive components are functions of the corresponding basis coefficients of the predictor processes. It may also work for nonparametric penalties other than COSSO, such as the sparsity-smoothness penalty proposed in Meier et al. (2009). The proposed method may be further extended to accommodate categorical responses, where an appropriate link function can be chosen to associate the mean response with the additive structure. Another possible extension is the regression with multiple functional predictors, where component selection can be performed for selecting significant functional predictors. In this case the additive components associated with each functional predictor need to be selected in a group manner.

## Acknowledgements

This work was conducted through the Analysis of Object Data program at Statistical and Applied Mathematical Sciences Institute, U.S.A.

## Appendix A: Estimation procedures

To estimate  $\zeta_i$ , we assume that the functional predictors are observed with measurement error on a grid of  $\mathcal{T}$ . We adopt two different procedures for functional data that are densely or sparsely observed.

- *Obtain  $\hat{\zeta}_i$  in the dense design.* If  $\{x_i(t)\}$  are observed on a sufficiently dense grid for each subject, we apply the local linear smoothing to the data  $\{t_{ij}, x_{ij}\}_{j=1, \dots, N_i}$  individually, which gives the smooth approximations  $\hat{x}_i(t)$ . The mean and covariance functions are obtained by  $\hat{\mu}(t) = 1/n \sum_{i=1}^n \hat{x}_i(t)$ , and  $\hat{G}(s, t) = 1/n \sum_{i=1}^n \{\hat{x}_i(s) - \hat{\mu}(s)\}\{\hat{x}_i(t) - \hat{\mu}(t)\}$ , respectively. The eigenvalues and eigenfunctions are estimated by solving the equation  $\int_{\mathcal{T}} \hat{G}(s, t) \phi_k(s) ds = \lambda_k \phi_k(t)$  for  $\lambda_k$  and  $\phi_k(\cdot)$ , subject to  $\int_{\mathcal{T}} \phi_k^2(t) dt = 1$  and  $\int_{\mathcal{T}} \phi_m(t) \phi_k(t) dt = 0$  for  $m \neq k$ ,  $k, m = 1, \dots, s$ . The FPC scores are obtained by  $\hat{\xi}_{ik} = \int_{\mathcal{T}} (\hat{x}_i(t) - \hat{\mu}(t)) \hat{\phi}_k(t) dt$ . Finally CDF transformation yields  $\hat{\zeta}_{ik} = \Psi(\hat{\xi}_{ik}; 0, \hat{\lambda}_k)$ .
- *Obtain  $\hat{\zeta}_i$  in the sparse design.* We adopt the principal component analysis through the conditional expectation (PACE) proposed by Yao et al. (2005), where the mean estimate  $\hat{\mu}(t)$  is obtained using local linear smoothers based on the pooled data of all individuals. In particular,  $\hat{\mu}(t) = \sum_{i=1}^n \sum_{j=1}^{N_i} K\{(t_{ij} - t)/b\} \{x_{ij} - \beta_0 - \beta_1(t - t_{ij})\}^2$  with  $K(\cdot)$  a kernel function and  $b$  a bandwidth. For the covariance estimation, denote  $G_{ijl} = \{x_{ij} - \hat{\mu}(t_{ij})\}\{x_{il} - \hat{\mu}(t_{il})\}$  and let  $K_h^*(\cdot, \cdot)$  be a bivariate kernel function with a bandwidth  $h$ , one minimizes  $\sum_{i=1}^n \sum_{j \neq l} K^*\{(t_{ij} - s)/h, (t_{il} - t)/h\} \{G_{ijl} - \beta_{00} - \beta_{11}(s - t_{ij}) - \beta_{12}(t - t_{il})\}^2$ . One may estimate the noise variance  $\nu^2$  by taking the difference between the diagonal of the surface estimate  $\hat{G}(t, t)$  and the local polynomial estimate obtained from the raw variances  $\{(t_{ij}, G_{ijj}) : j = 1, \dots, N_i; i = 1, \dots, n\}$ . The eigenvalues/functions are obtained as in the dense case. To estimate the FPC scores, denote  $\mathbf{x}_i = (x_{i1}, \dots, x_{iN_i})^\top$ , the PACE estimates are given by  $\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\phi}_{ik}^\top \hat{\Sigma}_{\mathbf{x}_i}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i)$ , which leads to  $\hat{\zeta}_{ik} = \Psi(\hat{\xi}_{ik}; 0, \hat{\lambda}_k)$ ,  $k = 1, \dots, s$ . Here  $\phi_{ik} = (\phi_k(t_{i1}), \dots, \phi_k(t_{iN_i}))^\top$ ,  $\boldsymbol{\mu}_i = (\mu_k(t_{i1}), \dots, \mu_k(t_{iN_i}))^\top$ , and the  $(j, l)$ th element  $(\boldsymbol{\Sigma}_{\mathbf{x}_i})_{j,l} = G(t_{ij}, t_{il}) + \nu^2 \delta_{jl}$  with  $\delta_{jl} = 1$  if  $j = l$  and 0 otherwise, and “ $\hat{\cdot}$ ” is the generic notation for their estimates.

We next estimate  $f_0 \in \mathcal{F}^s$  by minimizing (6), following the COSSO procedure conditional on the estimated values  $\hat{\zeta}_i$ . It is important to notice that the target function (6) is equivalent to  $1/n \sum_{i=1}^n \{y_i - f(\hat{\zeta}_i)\}^2 + \lambda_0 \sum_{k=1}^s \theta_k^{-1} \|P^k f\|^2 + \lambda \sum_{k=1}^s \theta_k$ , subject to  $\theta_k \geq 0$  (Lin and Zhang, 2006), which enables a two-step iterative algorithm. Specifically, one first find  $\mathbf{c} \in \mathbb{R}^n$  and  $\mathbf{b} \in \mathbb{R}$  by minimizing

$$(\mathbf{y} - \mathbf{R}_\theta \mathbf{c} - b \mathbf{1}_n)^\top (\mathbf{y} - \mathbf{R}_\theta \mathbf{c} - b \mathbf{1}_n) + n \lambda_0 \mathbf{c}^\top \mathbf{R}_\theta \mathbf{c}, \quad (9)$$

with fixed  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_s)^\top$ , where  $\mathbf{y} = (y_1, \dots, y_n)^\top$ ,  $\lambda_0$  is the smoothing parameter,  $\mathbf{1}_n$  is the  $n \times 1$  vector of 1's,  $\mathbf{R}_\theta = \sum_{k=1}^s \theta_k \mathbf{R}_k$ , and  $\mathbf{R}_k$  is the reproducing kernel of  $\bar{H}^k$ , i.e.

$\mathbf{R}_k = \{R_k(\hat{\zeta}_{ik}, \hat{\zeta}_{jk})\}_{1 \leq i, j \leq n}$ . This optimization is exactly a smoothing spline problem. We then fix  $\mathbf{c}$  and  $b$ , and find  $\boldsymbol{\theta}$  by minimizing

$$(\mathbf{z} - Q \boldsymbol{\theta})^\top (\mathbf{z} - Q \boldsymbol{\theta}) \quad \text{subject to} \quad \theta_k \geq 0; \quad \sum_{k=1}^s \theta_k \leq M, \quad (10)$$

where  $\mathbf{z} = \mathbf{y} - (1/2)n\lambda_0\mathbf{c} - b\mathbf{1}_n$  and  $Q$  is an  $n \times s$  matrix with the  $k$ th column being  $R_k\mathbf{c}$ . This step is the same as calculating the non-negative garrote estimate using  $M$  as the tuning parameter. Upon convergence, the final estimation of  $f$  is then given by  $\hat{f}(\boldsymbol{\zeta}) = \sum_{i=1}^n \hat{c}_i R_{\hat{\boldsymbol{\theta}}}(\hat{\zeta}_i, \boldsymbol{\zeta}) + \hat{b}$ .

Regarding the choices for tuning parameters involved, besides the sufficiently large initial truncation  $s$ , the most relevant are the  $\lambda_0$  and  $M$  in the COSSO step, while the bandwidths in smoothing steps of FPCA are chosen by traditional cross-validation or its generalized approximations. For more details, see Fan and Gijbels (1996) for the dense case and Yao et al. (2005) for the sparse case. We suggest to select  $\lambda_0$  using the generalized cross validation (GCV), i.e.  $\text{GCV}(\lambda_0) = (\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) / \{n^{-1} \text{tr}(I - A)\}^2$  with  $\hat{\mathbf{y}} = A\mathbf{y}$ . For choosing  $M$ , we adopt the BIC criterion i.e.  $\text{BIC}(M) = (\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y}) / \hat{\sigma}^2 + \log(n) \cdot df$  where  $df$  is the degree of freedom in (10), while an alternative is the cross-validation which requires more computation.

## Appendix B: Technical assumptions and proofs

We first lay out the commonly adopted regularity conditions on the functional predictor process  $X$  for the dense design. Recall that  $\{t_{ij}, j = 1, \dots, N_i; i = 1, \dots, n\}$  is the grid on the support  $\mathcal{T}$  over which the functional predictor  $x_i(t)$  is observed. Without loss of generality, let  $\mathcal{T} = [0, a]$ . Denote  $t_{i0} = 0$ ,  $t_{iN_i} = a$  and  $\mathcal{T}_d = [-d, a + d]$  for some  $d > 0$ . Denote the bandwidth used for individually smoothing the  $i$ th trajectory as  $b_i$ .

- (b.1) Assume that the second derivative  $X^{(2)}(t)$  is continuous on  $\mathcal{T}_d$  with probability 1 (w.p.1.), and  $\int E[\{X^{(k)}(t)\}^4]dt < \infty$  w.p.1. for  $k = 0, 2$ . Also assume  $E(e_{ij}^4) < \infty$ , where  $e_{ij}$  is the i.i.d. measurement error of the observed trajectory  $\mathbf{x}_i$ .
- (b.2) Assume that there exists  $m \equiv m(n) \rightarrow \infty$ , such that  $\min_i N_i \geq m$  as  $n \rightarrow \infty$ . Denoting  $\Delta_i = \max\{t_{ij} - t_{i,j-1} : j = 1, \dots, N_i + 1\}$ , assume that  $\max_i \Delta_i = O(m^{-1})$ .
- (b.3) Assume that there exists a sequence  $b = b(n)$ , such that  $cb \leq \min_i b_i \leq \max_i b_i \leq Cb$  for some  $C \geq c > 0$ . Furthermore,  $b \rightarrow 0$  and  $m \rightarrow \infty$  as  $n \rightarrow \infty$  in rates such that  $(mb)^{-1} + b^4 + m^{-2} = O(n^{-1})$ , e.g.  $b = O(n^{-1/2})$ ,  $m = O(n^{3/2})$ . Also assume that the kernel function  $K(\cdot)$  is compactly supported and Lipschitz continuous.

Denote the operator associated with the covariance function  $G(s, t)$  by  $G$ , and define  $\|G\|_S^2 = \int_{\mathcal{T}} \int_{\mathcal{T}} G^2(s, t) ds dt$ . Denote the smoothed trajectories of  $X_i(t)$  using local linear smoothing with the bandwidth  $b_i$  by  $\hat{X}_i$  and the estimated eigenvalues/functions and FPC scores in the dense design by  $\hat{\lambda}_k, \hat{\phi}_k, \hat{\xi}_{ik}$ , respectively. Since the decay of eigenvalues plays an important role, define  $\delta_1 = \lambda_1 - \lambda_2$  and  $\delta_k = \min_{j \leq k} (\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$  for  $k \geq 2$ .

*Lemma 1.* Under the assumptions (b.1)–(b.3), we have

$$E(\|\widehat{X}_i - X_i\|_{L^2}^2) = O(n^{-1}), \|\widehat{\mu} - \mu\|_{L^2} = O_p(n^{-1/2}), \|\widehat{G} - G\|_S = O_p(n^{-1/2}) \quad (11)$$

$$|\widehat{\lambda}_k - \lambda_k| \leq \|\widehat{G} - G\|_S, \quad \|\widehat{\phi}_k - \phi_k\|_{L^2} \leq 2\sqrt{2}\delta_k^{-1}\|\widehat{G} - G\|_S, \quad (12)$$

$$|\widehat{\xi}_{ik} - \xi_{ik}| = O_p(\|\widehat{X}_i - X_i\|_{L^2} + \delta_k^{-1}\|X_i\|_{L^2} \cdot \|\widehat{G} - G\|_S), \quad (13)$$

where  $O(\cdot)$  and  $O_p(\cdot)$  are uniform over  $1 \leq i \leq n$ .

Note that the measurement error  $e_{ij}$  is independent of the process  $X_i$ , which makes it possible to factor the probability space  $\Omega = \Omega_X \times \Omega_e$  and characterize the individual smoothing and cross-sectional averaging separately. Then (11) can be shown using standard techniques with local polynomial smoothing (not elaborated for conciseness), see Hall et al. (2006) for more details of this type of arguments. Consequently (12) and (13) follow immediately by the classical perturbation result provided in Lemma 4.3 of Bosq (2000). We see from Lemma 1 that, when the measurements are sufficiently dense for each subject satisfying (b.3), the impact due to individual smoothing on the estimated population quantities (e.g. mean, covariance, eigenvalues/functions) are negligible.

The following lemma characterizes the discrepancy between the underlying and estimated transformed variables  $\zeta_{ik}$ , as well as the boundedness of the derivative of the resulting estimate  $\widehat{f}$ .

*Lemma 2.* Under the assumptions (a.2) and (b.1)–(b.3), we have

$$|\widehat{\zeta}_{ik} - \zeta_{ik}| = O_p\left(\lambda_k^\gamma\{\|\widehat{X}_i - X_i\|_{L^2} + (\delta_k^{-1}\|X_i\|_{L^2} + |\xi_{ik}|)\|\widehat{G} - G\|_S\}\right), \quad (14)$$

$$\frac{1}{n} \sum_{i=1}^n \left( \sum_{k=1}^s |\widehat{\zeta}_{ik} - \zeta_{ik}| \right)^2 = O_p(n^{-1}), \quad (15)$$

Additionally, if the assumption (a.1) holds, let  $\widehat{f}$  be the estimate of  $f_0$  obtained by minimizing (6). Then there exists a constant  $\rho > 0$ , such that

$$\left| \frac{\partial \widehat{f}(\zeta_i)}{\partial \zeta_{ik}} \right| \leq \rho, \quad (16)$$

uniformly over  $1 \leq k \leq s$  and  $1 \leq i \leq n$ .

*Proof of Lemma 2.* From Lemma 1 and (a.2), one has in probability,

$$\begin{aligned} |\widehat{\zeta}_{ik} - \zeta_{ik}| &= \left| (\widehat{\xi}_{ik} - \xi_{ik}) \frac{\partial}{\partial \xi_{ik}} \Psi(\xi_{ik}, \lambda_k) + (\widehat{\lambda}_k - \lambda_k) \frac{\partial}{\partial \lambda_k} \Psi(\xi_{ik}, \lambda_k) + o_p(|\widehat{\xi}_{ik} - \xi_{ik}| + |\widehat{\lambda}_k - \lambda_k|) \right| \\ &\leq |\widehat{\xi}_{ik} - \xi_{ik}| \cdot \left| \frac{\partial}{\partial \xi_{ik}} \Psi(\xi_{ik}, \lambda_k) \right| + |\widehat{\lambda}_k - \lambda_k| \cdot \left| \frac{\partial}{\partial \lambda_k} \Psi(\xi_{ik}, \lambda_k) \right| + o_p(|\widehat{\xi}_{ik} - \xi_{ik}| + |\widehat{\lambda}_k - \lambda_k|) \\ &= O_p\left(\lambda_k^\gamma\{\|\widehat{X}_i - X_i\|_{L^2} + (\delta_k^{-1}\|X_i\|_{L^2} + |\xi_{ik}|)\|\widehat{G} - G\|_S\}\right). \end{aligned}$$

Abbreviate  $\sum_{i=1}^n$  to  $\sum_i$ ,  $\sum_{k=1}^s$  to  $\sum_k$  and  $O_p(\cdot)$  to  $\sim$ . Since  $E\|\widehat{X}_i - X_i\|_{L^2} \leq \{E(\|\widehat{X}_i - X_i\|_{L^2}^2)\}^{1/2} = O(n^{-1/2})$ , it easy to see that  $E(n^{-1} \sum_i \|\widehat{X}_i - X_i\|_{L^2}) = E\|\widehat{X}_i -$



$X_i\|_{L^2} = O(n^{-1/2})$ , To show (15) for any fixed  $s$ , note  $n^{-1} \sum_i \left( \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}| \right)^2 \leq s n^{-1} \sum_i \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|^2$ . Then

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^s (\hat{\zeta}_{ik} - \zeta_{ik})^2 &\sim \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^s \lambda_k^{2\gamma} \left\{ \|\hat{X}_i - X_i\|_{L^2} + (\delta_k^{-1} \|X_i\|_{L^2} + |\xi_{ik}|) \|\hat{G} - G\|_S \right\}^2 \\ &\sim \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} \|\hat{X}_i - X_i\|_{L^2}^2 + \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} \delta_k^{-2} \|X_i\|_{L^2}^2 \|\hat{G} - G\|_S^2 + \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} |\xi_{ik}|^2 \|\hat{G} - G\|_S^2 \\ &\quad + \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} \|\hat{X}_i - X_i\|_{L^2} \delta_k^{-1} \|X_i\|_{L^2} \|\hat{G} - G\|_S + \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} \|\hat{X}_i - X_i\|_{L^2} |\xi_{ik}| \|\hat{G} - G\|_S \\ &\quad + \frac{1}{n} \sum_i \sum_k \lambda_k^{2\gamma} \delta_k^{-1} |\xi_{ik}| \|X_i\|_{L^2} \|\hat{G} - G\|_S^2, \end{aligned}$$

Denote the additive terms in above formula  $E_1$  through  $E_6$ , we have  $E_1 = (\sum_k \lambda_k^{2\gamma})(n^{-1} \sum_i \|\hat{X}_i - X_i\|_{L^2}^2) = O_p(n^{-1})$ ,  $E_2 = \|\hat{G} - G\|_S^2 (\sum_k \lambda_k^{2\gamma} \delta_k^{-2})(n^{-1} \sum_i \|X_i\|_{L^2}^2) = O_p(n^{-1})$ ,  $E_3 = \|\hat{G} - G\|_S^2 (1/n \sum_i \sum_k \lambda_k^{2\gamma} |\xi_{ik}|^2) = O_p(n^{-1})$ , as  $E(n^{-1} \sum_{i=1}^n \sum_{k=1}^s \lambda_k^{2\gamma} |\xi_{ik}|^2) = \sum_k \lambda_k^{2\gamma+1} = O(1)$ . For  $E_4$ , applying Cauchy-Schwarz inequality,

$$\begin{aligned} E_4 &\sim \|\hat{G} - G\|_S \left( \sum_{k=1}^s \lambda_k^{2\gamma} \delta_k^{-1} \right) \left( \frac{1}{n} \sum_{i=1}^n \|\hat{X}_i - X_i\|_{L^2} \|X_i\|_{L^2} \right) \\ &\leq 2C \|\hat{G} - G\|_S \left( \sum_{k=1}^s \lambda_k^{2\gamma} \delta_k^{-1} \right) \sqrt{\left( \frac{1}{n} \sum_{i=1}^n \|\hat{X}_i - X_i\|_{L^2}^2 \right) \left( \frac{1}{n} \sum_{i=1}^n \|X_i\|_{L^2}^2 \right)} \\ &= O_p(n^{-1/2}) O(1) O_p(n^{-1/2}) O_p(1) = O_p(n^{-1}). \end{aligned}$$

Similarly, one has  $E_5 = O_p(n^{-1})$  and  $E_6 = O_p(n^{-1})$ , using the facts that  $E\{(\sum_{k=1}^s \lambda_k^{2\gamma} |\xi_{ik}|)^2\} \leq s \sum_{k=1}^s \lambda_k^{4\gamma+1} = O(1)$ .  $E(\sum_{k=1}^s \lambda_k^{2\gamma} \delta_k^{-1} |\xi_{ik}|)^2 \leq s \sum_{k=1}^s \lambda_k^{4\gamma+1} \delta_k^{-2} = O(1)$ . This proves (15).

We now turn to (16). For any  $f \in \mathcal{F}^s$ , one has

$$f(\zeta_i) = \langle f(\cdot), R(\zeta_i, \cdot) \rangle_{\mathcal{F}^s} \leq \|f\| \langle R(\zeta_i, \cdot), R(\zeta_i, \cdot) \rangle_{\mathcal{F}^s}^{1/2} = \|f\| R^{1/2}(\zeta_i, \zeta_i),$$

where  $R(\cdot, \cdot)$  is the reproducing kernel of space  $\mathcal{F}^s$  and  $\langle \cdot, \cdot \rangle_{\mathcal{F}^s}$  is the corresponding inner product. Therefore,

$$\frac{\partial f(\zeta_i)}{\partial \zeta_{ik}} = \left\langle f(\cdot), \frac{\partial R(\zeta_i, \cdot)}{\partial \zeta_{ik}} \right\rangle_{\mathcal{F}^s} \leq \|f\| \left\langle \frac{\partial R(\zeta_i, \cdot)}{\partial \zeta_{ik}}, \frac{\partial R(\zeta_i, \cdot)}{\partial \zeta_{ik}} \right\rangle_{\mathcal{F}^s}^{1/2}.$$

Since  $J(f)$  is a convex functional and a pseudonorm, we have

$$\sum_{k=1}^s \|P^k f\|^2 \leq J^2(f) \leq s \sum_{k=1}^s \|P^k f\|^2. \quad (17)$$

We first claim that  $\|f\| \leq J(f)$ , due to  $\|f\|^2 = b^2 + \sum_{k=1}^s \|P^k f\|^2$ . If  $b = 0$ , the inequality in (17) implies that  $\|f\| \leq J(f)$ . If  $b \neq 0$ , one can write  $\tilde{J}(f) = b + J(f) = b + \sum_{k=1}^s \|P^k f\|^2$ . For minimizing (5), it is equivalent to substitute  $J(f)$  with  $\tilde{J}(f)$ , and (17) implies  $\|f\|^2 =$

$b^2 + \sum_{k=1}^s \|P^k f\|^2 \leq b^2 + J^2(f) \leq \tilde{J}^2(f)$ . Therefore we have  $\|f\| \leq J(f)$  in general. Secondly, due to the orthogonality of  $\{\bar{H}^k\}$ , we can write  $R(\mathbf{u}, \mathbf{v}) = R_1(u_1, v_1) + R_2(u_2, v_2) + \dots + R_s(u_s, v_s)$  by Theorem 5 in Berlinet and Thomas-agnan (2004), where  $R_k(\cdot, \cdot)$  is the reproducing kernel of the subspace  $\bar{H}^k$ . For  $\bar{H}^k$  being a second order Sobolev Hilbert space, we have  $R_k(s, t) = h_1(s)h_1(t) + h_2(s)h_2(t) - h_4(|s - t|)$ , with  $h_1(t) = t - 1/2$ ,  $h_2(t) = [h_1^2(t) - 1/12]/2$  and  $h_4(t) = [h_1^4(t) - h_1^2(t)/2 + 7/240]/24$ . Therefore  $R_k(s, t)$  is continuous and differentiable over  $[0, 1]^2$  and we can find constants  $a_k$  and  $b_k$  such that

$$\langle R_k(u, \cdot), R_k(u, \cdot) \rangle_{\mathcal{F}^s} < a_k, \quad \left\langle \frac{\partial R_k(u, \cdot)}{\partial u}, \frac{\partial R_k(u, \cdot)}{\partial u} \right\rangle_{\mathcal{F}^s} \leq b_k,$$

for  $k = 1, \dots, s$ . One can find a uniform bound  $c$  with  $\langle \partial/\partial \zeta_{ik} R(\zeta_i, \cdot), \partial/\partial \zeta_{ik} R(\zeta_i, \cdot) \rangle_{\mathcal{F}^s} \leq c$ . On the other hand, a  $\hat{f}$  minimizing (6) is equivalent to minimizing  $n^{-1} \sum_i \{y_i - f(\hat{\zeta}_i)\}^2$  under the constraint that  $J(f) \leq \tilde{c}$  for some  $\tilde{c} > 0$ . Therefore let  $\rho = c^{1/2} \cdot \tilde{c}$ , we have

$$\left| \frac{\partial \hat{f}(\zeta_i)}{\partial \zeta_{ik}} \right| \leq \|\hat{f}\| \left\langle \frac{\partial R(\zeta_i, \cdot)}{\partial \zeta_{ik}}, \frac{\partial R(\zeta_i, \cdot)}{\partial \zeta_{ik}} \right\rangle_{\mathcal{F}^s}^{1/2} \leq J(\hat{f}) c^{1/2} \leq \tilde{c} c^{1/2} = \rho. \quad \square$$

Before stating Lemma 3, we define the entropy of  $\mathcal{F}^s$  with respect to the  $\|\cdot\|_n$  metric. For each  $\omega > 0$ , one can find a collection of functions  $\{g_1, g_2, \dots, g_N\}$  in  $\mathcal{F}^s$  such that for each  $g \in \mathcal{F}^s$ , there is a  $j = j(g) \in \{1, 2, \dots, N\}$  satisfying  $\|g - g_j\|_n \leq \omega$ . Let  $\mathbb{N}(\omega, \mathcal{F}^s, \|\cdot\|_n)$  be the smallest value of  $N$  for which such a cover of balls with radius  $\omega$  and centers  $g_1, g_2, \dots, g_N$  exists. Then  $H(\omega, \mathcal{F}^s, \|\cdot\|_n) = \log \mathbb{N}(\omega, \mathcal{F}^s, \|\cdot\|_n)$  is called the  $\omega$ -entropy of  $\mathcal{F}^s$ .

*Lemma 3.* Assume that  $\mathcal{F}^s = \{1\} \oplus \sum_{k=1}^s \bar{H}^k$ , where  $\bar{H}^k$  is the second order Sobolev space. Denote the  $\omega$ -entropy of  $\{f \in \mathcal{F}^s : J(f) \leq 1\}$  by  $H(\omega, \{f \in \mathcal{F}^s : J(f) \leq 1\}, \|\cdot\|_n)$ . Then

$$H(\omega, \{f \in \mathcal{F}^s : J(f) \leq 1\}, \|\cdot\|_n) \leq A\omega^{-1/2}, \quad (18)$$

for all  $\omega > 0, n \geq 1$ , and for some constants  $A > 0$ . Furthermore, for  $\{\epsilon_i\}_{i=1}^n$  independent with finite variance and  $J(f_0) > 0$ ,

$$\sup_{f \in \mathcal{F}^s} \frac{|\langle \epsilon, f - f_0 \rangle_n|}{\|f - f_0\|_n^{3/4} (J(f) + J(f_0))^{1/4}} = O_p(n^{-1/2}). \quad (19)$$

The inequality (18) is implied by Lemma A.1. of Lin and Zhang (2006). As the  $\{\epsilon_i\}$  satisfy the sub-Gaussian error assumption, the same argument as in Van de Geer (2000) (pg. 168) leads to (19). We are now ready to present the proof of the main theorem.

*Proof of Theorem 1.* We first center the functions as in the proof of theorem 2 in Lin and Zhang (2006) so that (18) and (19) holds. Write  $f(\hat{\zeta}) = c + f_1(\hat{\zeta}_1) + \dots + f_s(\hat{\zeta}_s) = c + \tilde{f}(\hat{\zeta})$ , such that  $\sum_{i=1}^n f_k(\hat{\zeta}_{i,k}) = 0$ , and write  $f_0(\zeta) = c_0 + f_{01}(\zeta_1) + \dots + f_{0s}(\zeta_s) = c_0 + \tilde{f}_0(\zeta)$  such that  $\sum_{i=1}^n f_{0k}(\zeta_{i,k}) = 0$  and  $\hat{f}(\zeta) = \hat{c} + \hat{f}_1(\zeta_1) + \dots + \hat{f}_s(\zeta_s)$ . Since the target function can be written as

$$\begin{aligned} Q(f | \hat{\zeta}_i) &= \frac{1}{n} \sum_{i=1}^n \{y_i - f(\hat{\zeta}_i)\}^2 + \tau_n^2 J(f) = \frac{1}{n} \sum_{i=1}^n \{c_0 + \tilde{f}_0(\zeta_i) + \epsilon_i - c - \tilde{f}(\hat{\zeta}_i)\}^2 + \tau_n^2 J(f) \\ &= (c_0 - c)^2 + \frac{2}{n} (c_0 - c) \sum_i \epsilon_i + \frac{1}{n} \sum_{i=1}^n \{\tilde{f}_0(\zeta_i) + \epsilon_i - \tilde{f}(\hat{\zeta}_i)\}^2 + \tau_n^2 J(f), \end{aligned}$$

one must have that  $\hat{c}$  minimizes  $\{(c_0 - c)^2 + 2n^{-1}(c_0 - c)\sum_i \epsilon_i\}$  and the additive parts of  $\hat{f}$  minimizes the rest. Therefore we have  $\hat{c} - c_0 = n^{-1}\sum_i \epsilon_i$ , implying  $|\hat{c} - c_0| = O_p(n^{-1/2})$ . Denote

$$\tilde{Q}(\tilde{f} \mid \{\hat{\zeta}_i\}) = 1/n \sum_{i=1}^n \{\tilde{f}_0(\zeta_i) + \epsilon_i - \tilde{f}(\hat{\zeta}_i)\}^2 + \tau_n^2 J(\tilde{f}). \quad (20)$$

One can substitute  $\tau_n^2 J(f)$  with  $\tau_n^2 J(\tilde{f})$  in (20). In the rest of the proof, we suppress the tilde notation of  $\tilde{f}_0$  and  $\tilde{f}$  for convenience. Since  $\hat{f} = \underset{f \in \mathcal{F}^s}{\operatorname{argmin}} \tilde{Q}(f \mid \{\hat{\zeta}_i\})$ , one has

$\tilde{Q}(\hat{f} \mid \{\hat{\zeta}_i\}) \leq \tilde{Q}(f_0 \mid \{\hat{\zeta}_i\})$ , which implies

$$\frac{1}{n} \sum_{i=1}^n \{f_0(\zeta_i) + \epsilon_i - \hat{f}(\hat{\zeta}_i)\}^2 + \tau_n^2 J(\hat{f}) \leq \frac{1}{n} \sum_{i=1}^n \{f_0(\zeta_i) + \epsilon_i - f_0(\hat{\zeta}_i)\}^2 + \tau_n^2 J(f_0).$$

Simplification of above inequality gives

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( f_0(\zeta_i) - \hat{f}(\hat{\zeta}_i) \right)^2 + \tau_n^2 J(\hat{f}) \\ & \leq \frac{2}{n} \sum_{i=1}^n \epsilon_i \left( \hat{f}(\hat{\zeta}_i) - f_0(\hat{\zeta}_i) \right) + \frac{1}{n} \sum_{i=1}^n \left( f_0(\zeta_i) - f_0(\hat{\zeta}_i) \right)^2 + \tau_n^2 J(f_0). \end{aligned} \quad (21)$$

Let  $g(\cdot) = \hat{f}(\cdot) - f_0(\cdot)$ . Since both  $\hat{f}$  and  $f_0$  are in  $\mathcal{F}^s$ ,  $g \in \mathcal{F}^s$ . Taylor expansion of  $g(\cdot)$  gives  $g(\hat{\zeta}) = g(\zeta) + Dg(\zeta)(\hat{\zeta} - \zeta) + o_p\left(\sum_{k=1}^s |\hat{\zeta}_k - \zeta_k|\right)$ , for all  $\zeta \in (0, 1)^s$ , where  $Dg(\zeta)(\hat{\zeta} - \zeta) = \sum_{k=1}^s (\partial g(\zeta)/\partial \zeta_k)(\hat{\zeta}_k - \zeta_k)$ . Then we have

$$\frac{2}{n} \sum_{i=1}^n \epsilon_i g(\hat{\zeta}_i) = \frac{2}{n} \sum_{i=1}^n \epsilon_i g(\zeta_i) + \frac{2}{n} \sum_{i=1}^n \epsilon_i \left\{ Dg(\zeta_i)(\hat{\zeta}_i - \zeta_i) + o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) \right\},$$

and plug it into the right hand side (r.h.s.) of (21), leading to the following upper bound,

$$\begin{aligned} & \frac{2}{n} \sum_{i=1}^n \epsilon_i \left( \hat{f}(\zeta_i) - f_0(\zeta_i) \right) + \frac{2}{n} \sum_{i=1}^n \epsilon_i \left\{ (D\hat{f}(\zeta_i) - Df_0(\zeta_i))(\hat{\zeta}_i - \zeta_i) + o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) \right\} \\ & + \frac{1}{n} \sum_{i=1}^n \left( f_0(\zeta_i) - f_0(\hat{\zeta}_i) \right)^2 + \tau_n^2 J(f_0). \end{aligned} \quad (22)$$

Applying Lemma 3, one can bound the first term in (22) as follow,

$$\frac{2}{n} \sum_{i=1}^n \epsilon_i \left( \hat{f}(\zeta_i) - f_0(\zeta_i) \right) = 2(\epsilon, \hat{f} - f_0)_n \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} (J(\hat{f}) + J(f_0))^{1/4}.$$

For the left hand side (l.f.s.) of (21), applying the Taylor expansion,  $\hat{f}(\hat{\zeta}_i) = \hat{f}(\zeta_i) + D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) + o_p(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|)$ , to the first term

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( f_0(\zeta_i) - \hat{f}(\hat{\zeta}_i) \right)^2 = \frac{1}{n} \sum_{i=1}^n \left\{ f_0(\zeta_i) - \hat{f}(\zeta_i) - D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) - o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) \right\}^2 \\ & = \frac{1}{n} \sum_{i=1}^n \left\{ \left( f_0(\zeta_i) - \hat{f}(\zeta_i) \right)^2 + \left( D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \right)^2 - 2 \left( f_0(\zeta_i) - \hat{f}(\zeta_i) \right) D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) + R_i \right\}, \end{aligned}$$

where  $R_i = \{o_p(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|)\}^2 - o_p(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|)\{f_0(\zeta_i) - \hat{f}(\zeta_i) - D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i)\}$ . Substituting the terms on both sides of (21), we obtain

$$\begin{aligned} & \|\hat{f} - f_0\|_n^2 + \frac{1}{n} \sum_{i=1}^n \left\{ \left( D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \right)^2 + 2 \left( \hat{f}(\zeta_i) - f_0(\zeta_i) \right) D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) + R_i \right\} + \tau_n^2 J(\hat{f}) \\ & \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} (J(\hat{f}) + J(f_0))^{1/4} + \frac{1}{n} \sum_{i=1}^n \left( f_0(\zeta_i) - f_0(\hat{\zeta}_i) \right)^2 + \tau_n^2 J(f_0) \\ & \quad + \frac{2}{n} \sum_{i=1}^n \epsilon_i \left\{ (D\hat{f}(\zeta_i) - Df_0(\zeta_i))(\hat{\zeta}_i - \zeta_i) + o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) \right\}. \end{aligned}$$

Dropping the positive term  $n^{-1} \sum_i \left( D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \right)^2$  on the l.h.s. and rearranging the terms,

$$\begin{aligned} \|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) & \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} (J(\hat{f}) + J(f_0))^{1/4} + \tau_n^2 J(f_0) \\ & \quad + T_1 + T_2 + T_3 + \frac{2}{n} \sum_{i=1}^n \epsilon_i \tilde{R}_{2i} + \frac{1}{n} \sum_{i=1}^n \tilde{R}_{1i} \end{aligned} \quad (23)$$

where  $T_1 = -2n^{-1} \sum_i \{(\hat{f}(\zeta_i) - f_0(\zeta_i)) D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i)\}$ ,  $T_2 = 2n^{-1} \sum_i \epsilon_i \{ (D\hat{f}(\zeta_i) - Df_0(\zeta_i))(\hat{\zeta}_i - \zeta_i) \}$ ,  $T_3 = n^{-1} \sum_i \{ f_0(\zeta_i) - f_0(\hat{\zeta}_i) \}^2$ ,  $\tilde{R}_{1i} = o_p(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|) \{ f_0(\zeta_i) - \hat{f}(\zeta_i) - D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \}$  and  $\tilde{R}_{2i} = o_p(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|)$ .

For  $T_1$ , by Cauchy-Schwarz inequality and Lemma 2, we have  $T_1 \leq 2\sqrt{\|\hat{f} - f_0\|_n^2 A}$ , where

$$A = \frac{1}{n} \sum_{i=1}^n \left( \sum_{k=1}^s \frac{\partial \hat{f}(\zeta_i)}{\partial \zeta_{ik}} (\hat{\zeta}_{ik} - \zeta_{ik}) \right)^2 \leq \frac{\rho^2}{n} \sum_{i=1}^n \left( \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}| \right)^2 = O_p(n^{-1}),$$

i.e.  $T_1 = \|\hat{f} - f_0\|_n O_p(n^{-1/2})$ . From (A.1) and (16) of Lemma 2, there exists independent r.v.  $\{B_i\}$  with  $E(B_i^2) < \infty$  such that  $\max_k \{ |\partial \hat{f}(\zeta_i) / \partial \zeta_{ik} - \partial f_0(\zeta_i) / \partial \zeta_{ik}| \} \leq B_i \|\hat{f} - f_0\|_{L^2}$ . Also note that  $\|g\|_n \rightarrow \|g\|_{L_2}$  a.s. by the strong law of large numbers. Therefore we have, for some constant  $c$ ,

$$\begin{aligned} T_2 & \leq \frac{2}{n} \sum_{i=1}^n |\epsilon_i| \sum_{k=1}^s B_i \|\hat{f} - f_0\|_{L^2} |\hat{\zeta}_{ik} - \zeta_{ik}| = 2 \|\hat{f} - f_0\|_{L^2} \left( \frac{1}{n} \sum_{i=1}^n |\epsilon_i B_i| \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}| \right) \\ & \leq c \|\hat{f} - f_0\|_n \sqrt{\left( \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 B_i^2 \right) \left( \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|^2 \right)} = \|\hat{f} - f_0\|_n O_p(n^{-1/2}). \\ T_3 & = \frac{1}{n} \sum_{i=1}^n \left( \sum_{k=1}^s \frac{\partial f_0(\zeta_i)}{\partial \zeta_{ik}} \left\{ (\hat{\zeta}_{ik} - \zeta_{ik}) + o_p(|\hat{\zeta}_{ik} - \zeta_{ik}|) \right\} \right)^2 \leq \frac{c}{n} \sum_{i=1}^n \left( \sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}| \right)^2 = O_p(n^{-1}). \end{aligned}$$

For the remaining terms,  $n^{-1} \sum_{i=1}^n \epsilon_i \tilde{R}_{2i} = o_p(T_2)$ , and

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \tilde{R}_{1i} & = \frac{1}{n} \sum_{i=1}^n o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) (f_0(\zeta_i) - \hat{f}(\zeta_i)) - \frac{1}{n} \sum_{i=1}^n o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \\ & \leq o_p(T_1) + \left\{ \frac{1}{n} \sum_{i=1}^n \left( o_p\left(\sum_{k=1}^s |\hat{\zeta}_{ik} - \zeta_{ik}|\right) \right)^2 \frac{1}{n} \sum_{i=1}^n \left( D\hat{f}(\zeta_i)(\hat{\zeta}_i - \zeta_i) \right)^2 \right\}^{-1/2} = o_p(T_1) + o_p(n^{-1}). \end{aligned}$$

We can now simplify (23) as follows:

$$\begin{aligned} \|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) &\leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} \left( J(\hat{f}) + J(f_0) \right)^{1/4} \\ &\quad + \|\hat{f} - f_0\|_n O_p(n^{-1/2}) + O_p(n^{-1}) + \tau_n^2 J(f_0). \end{aligned}$$

If  $O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} \left( J(\hat{f}) + J(f_0) \right)^{1/4} \geq \|\hat{f} - f_0\|_n O_p(n^{-1/2}) + O_p(n^{-1}) + \tau_n^2 J(f_0)$ , we have

$$\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} \left( J(\hat{f}) + J(f_0) \right)^{1/4}, \quad (24)$$

otherwise,

$$\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq \|\hat{f} - f_0\|_n O_p(n^{-1/2}) + O_p(n^{-1}) + 2\tau_n^2 J(f_0). \quad (25)$$

The proof will be complete by solving them separately. For the case of (24), there are two possibilities.

(i) If  $J(\hat{f}) \geq J(f_0)$ , (24) implies  $\tau_n^2 J^{3/4}(\hat{f}) \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4}$ , and

$$\begin{aligned} J^{1/4}(\hat{f}) &\leq \left( \tau_n^{-2} O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} \right)^{1/3} = O_p(n^{-1/6}) \|\hat{f} - f_0\|_n^{1/4} \tau_n^{-2/3}. \text{ Therefore,} \\ \|\hat{f} - f_0\|_n^2 &\leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} J^{1/4}(\hat{f}) \leq O_p(n^{-2/3}) \|\hat{f} - f_0\|_n \tau_n^{-2/3}, \text{ i.e.} \\ \|\hat{f} - f_0\|_n &= O_p(n^{-2/3}) O_p(\tau_n^{-2/3}), \quad J(\hat{f}) = O_p(n^{-4/3}) O_p(\tau_n^{-10/3}). \end{aligned} \quad (26)$$

(ii). If  $J(\hat{f}) < J(f_0)$ , then  $J(\hat{f}) = O_p(J(f_0)) O_p(1)$ , and (24) implies that  $\|\hat{f} - f_0\|_n^2 \leq O_p(n^{-1/2}) \|\hat{f} - f_0\|_n^{3/4} J^{1/4}(f_0)$ , which leads to

$$\|\hat{f} - f_0\|_n = O_p(n^{-2/5}) J^{1/5}(f_0), \quad J(\hat{f}) = J(f_0) O_p(1). \quad (27)$$

Note that the results (26) and (27) are equivalent under the condition (7),.

For the case of (25), if  $\|\hat{f} - f_0\|_n O_p(n^{-1/2}) > O_p(n^{-1}) + 2\tau_n^2 J(f_0)$ , we have  $\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq \|\hat{f} - f_0\|_n O_p(n^{-1/2})$ , otherwise  $\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq O_p(n^{-1}) + 4\tau_n^2 J(f_0)$ . The first inequality implies that

$$\|\hat{f} - f_0\|_n = O_p(n^{-1/2}), \quad J(\hat{f}) = O_p(n^{-1}) O_p(\tau_n^{-2}). \quad (28)$$

For the second inequality, if  $O_p(n^{-1}) < 4\tau_n^2 J(f_0)$ , we have  $\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq 8\tau_n^2 J(f_0)$ , implying

$$\|\hat{f} - f_0\|_n = O_p(\tau_n) J^{1/2}(f_0), \quad J(\hat{f}) = J(f_0) O_p(1). \quad (29)$$

If  $O_p(n^{-1}) \geq 4\tau_n^2 J(f_0)$  and  $\|\hat{f} - f_0\|_n^2 + \tau_n^2 J(\hat{f}) \leq O_p(n^{-1})$ , then

$$\|\hat{f} - f_0\|_n = O_p(n^{-1/2}) \quad J(\hat{f}) = O_p(n^{-1}) O_p(\tau_n^{-2}). \quad (30)$$

When  $J(f_0) > 0$ , given the condition (7), the rates of  $\|\hat{f} - f_0\|_n$  and  $J(\hat{f})$  from (29), (26) and (27) are the same, and dominate those of (28) and (30). Therefore we have  $\|\hat{f} - f_0\|_n = O_p(n^{-2/5}) J^{1/5}(f_0)$  and  $J(\hat{f}) = J(f_0) O_p(1)$ . When  $J(f_0) = 0$ , then (24) implies (26), while (25) implies (28) and (30). The possibility (ii) of (24) does not exist, nor does the result in (29). Under condition (8), the result of (26) is the same as those of (28) and (30). Therefore  $\|\hat{f} - f_0\|_n = O_p(n^{-1/2})$  and  $J(\hat{f}) = O_p(n^{-1/2})$ .  $\square$

## References

- Berlinet, A. and Thomas-agnan, C. (2004). *Reproducing Kernel Hilbert Space in Probability and Statistics*. Norwell, Massachusetts USA: Kluwer Academic Publishers.
- Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, Volume 149. New York: Springer-Verlag Inc.
- Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Stat.* 30, 241–255.
- Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Stat. Probabil. Lett.* 45, 11–22.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.* 96, 1348–1360.
- Fan, J. and Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J.R. Statist. Soc. B* 62, 303–322.
- Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J.R. Statist. Soc. B* 68, 109–126.
- Hall, P., Müller, H., and Wang, J. (2006). Properties of principle component methods for functional and longitudinal data analysis. *Ann. Statist.* 34, 1493–1517.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.
- Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall/CRC.
- James, G. M. (2002). Generalized linear models with functional predictors. *J.R. Statist. Soc. B* 64(3), 411–432.
- James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that’s interpretable. *Ann. Statist.* 37, 2083–2108.
- Li, Y., Wang, N., and Carroll, R. (2010). Generalized functional linear models with semi-parametric single-index interactions. *J. Am. Statist. Ass.* 105, 621–633.
- Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate non-parametric regression. *Ann. Statist.* 34, 2272–2297.
- Meier, L., Van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* 37, 3779–3821.
- Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* 33(2), 774–805.

- Müller, H. and Yao, F. (2008). Functional additive models. *J. Am. Statist. Ass.* 103(484), 1534–1544.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, Section Edition*. New York: Springer.
- Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *J. R. Statist. Soc. B* 71, 1009–1030.
- Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B* 53, 233–243.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58, 267–288.
- Van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.
- Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.
- Xu, L., Jiang, J., Wu, H., Shen, G., and Yu, R. (2007). Variable-weighted PLS. *Chemometr. Intell. Lab.* 85, 140–143.
- Yao, F. (2007). Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *J. Multivariate Anal.* 98, 40–56.
- Yao, F. and Müller, H. G. (2010). Functional quadratic regression. *Biometrika* 97, 49–64.
- Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.* 100, 577–590.
- Zhu, H., Vannucci, M., and Cox, D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* 66, 463–473.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.* 101, 1418–1429.