Noname manuscript No.
(will be inserted by the editor)

Supervised Functional Principal Component Analysis

Yunlong Nie · Liangliang Wang · Baisen Liu · Jiguo Cao

Received: date / Accepted: date

Abstract In functional linear regression, one conventional approach is to first perform functional principal component analysis (FPCA) on the functional predictor and then use the first few leading functional principal component (FPC) scores to predict the response variable. The leading FPCs estimated by the conventional FPCA stand for the major source of variation of the functional predictor, but these leading FPCs may not be mostly correlated with the response variable, so the prediction accuracy of the functional linear regression model may not be optimal. In this paper, we propose a supervised version of FPCA by considering the correlation of the functional predictor and response variable. It can automatically estimate leading FPCs, which represent the major source of variation of the functional predictor and are simultaneously correlated with the response variable. Our supervised FPCA method is demonstrated to have a better prediction accuracy than the conventional FPCA method by using one real application on electroencephalography (EEG) data and three carefully-designed simulation studies.

Keywords Classification · Functional Data Analysis · Functional Linear Model · Functional Logistic Regression

1 Introduction

In this paper, we study the problem of predicting a scalar response Y using the following functional linear model

$$E(Y|X(t)) = g\left(\beta_0 + \int_{\mathscr{T}} \beta(t)\{X(t) - \mu(t)\}dt\right) \qquad (1)$$

where $\beta_0 \in \mathbb{R}$ is the intercept, $X(t)$ is the functional predictor process with the mean function $\mu(t)$, $\beta(t)$ is the slope function, and both $\beta(t)$ and $X(t)$ are assumed to be smooth and square integrable on the domain $\mathscr{T}$. The link function g is assumed to be monotonic and invertible. The parametric form of g is chosen based on the distribution assumption on Y. For instance, $g(\cdot)$ is usually chosen as the inverse logit function if Y is a binary variable.

The above functional linear model has been widely used to link a scalar response with an integral form of a functional predictor. Compared with the classic regression problem in which only scalar predictors are considered, the main challenge in this functional linear model is that even a single functional predictor can lead to a saturated model due to its high flexibility. A common strategy to address this problem is through the functional principal component analysis (FPCA). The FPCA method estimates the functional linear model (1) in two steps: estimating the functional principal components (FPCs) for the functional predictor; and then using several leading FPCs in the functional linear model. This topic has been extensively studied in the literature such as Ramsay and Silverman (2002); Yao et al. (2005); Ramsay and Silverman (2005); Ramsay et al. (2009) and Huang et al. (2009). Furthermore, functional linear models have been naturally extended to generalized functional linear regression when the response variable is binary or multinomial. For example, Ratcliffe et al. (2002) applied a functional logistic regression to predict the high-risk birth rate based on periodically stimulated foetal heart rate tracings. Müller and Stadtmüller (2005) related the response with the integral form of a functional predictor through a smooth function. Cardot et al. (2003) used a multinomial functional regression model to predict the land usage based on the temporal evolution of coarse resolution remote sensing data.

However, a common limitation of the above methods is that the estimation of FPCs in the first step is totally separated from the regression model used to predict the response variable Y in the second step. In the first step, the leading

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada, V5A1S6. E-mail: jiguo_cao@sfu.ca

FPCs mainly focus on explaining the maximum variation of the functional predictor. Thus, the estimated FPCs may not have the maximum prediction power for Y. Therefore, practitioners usually have to include as many FPCs as possible to fit the functional regression, which introduces excessive variability into the model, especially when the sample size is relatively small. Our goal is to borrow the information from the response variable Y to estimate FPCs in the first step such that the resulting FPCs have a better performance in terms of predicting Y. This strategy is called supervised FPCA in this manuscript.

Bair et al. (2006) introduced a supervised principal component analysis (PCA) method in the context of classic multivariate regression problem, especially when the number of predictors was much larger than the sample size. They proposed a latent variable framework in which the response variable is only associated with a subset of predictors through a latent variable. More specifically, their method consists of three steps: first a pre-screening procedure is employed to select those important predictors; then PCA is performed on those selected predictors to estimate the latent variable; finally a regression model is fitted with those estimated PC scores. Li et al. (2015) proposed another version of supervised PCA, namely, a supervised singular value decomposition (SupSVD) model. Unlike Bair et al. (2006) focusing on predicting the response variable Y, the primary interest of the SupSVD model is to recover the underlying low-rank structure of the predictor matrix with the supervision information from Y. In addition, Li et al. (2015) could incorporate a multi-dimensional response variable whereas Bair et al. (2006) only considered a single scalar response variable.

However, neither of the above work can accommodate functional predictors. The extension from supervised PCA to functional data is nontrivial. Recently, Li et al. (2016) extended the SupSVD model to functional principal component analysis (FPCA) and proposed a method called supervised sparse functional principal component (SupSFPC). They assume that the supervision data drive low-rank structures of the functional data of primary interest. The estimation procedure is based on the penalized likelihood function that imposes a smooth and sparsity penalty on PC loadings. The difference between our work and theirs is that we mainly focus on improving the prediction performance of FPCs, while Li et al. (2016) focused on recovering the true FPCs.

The novelty of the paper is three-fold. Firstly, we propose a framework to utilize the scalar response variable, either continuous or categorical, to boost the prediction performance of the estimated FPCs. Our method is particularly useful dealing with 'Large p, Small n' problem when multiple functional predictors exist. Secondly, unlike Bair et al. (2006) which employs three steps, our approach does not require a pre-screening procedure. Thirdly, our estimation algorithm is based on eigenvalue decomposition which is much easier to implement in comparison with the revised EM algorithm used by SupSFPC. An R package "sFPCA" is developed to implement our proposed supervised FPCA method.

The rest of the paper is organized as follows. A review of conventional FPCA analysis is given in Section 2. Details of our method is described in Section 3. Then we show one real data application on electroencephalography (EEG) data in Section 4. Three carefully-designed simulation studies are used to evaluate the finite sample performance of our proposed method in Section 5. Section 6 provides concluding remarks.

2 Estimating Functional Linear Models using FPCA

We first introduce the conventional FPCA method for estimating the functional linear model (1), which is also called unsupervised FPCA method in this article. Consider a stochastic process $X(t)$ on the domain $\mathscr{T}$ with the mean function $E(X(t)) = \mu(t)$. Using the Karhunen-Loève expansion Fukunaga and Koontz (1970), the stochastic process $X(t)$ can be expressed as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \alpha_j \xi_j(t), \quad i = 1, \ldots, n, \tag{2}$$

where $\xi_j(t), j = 1, \ldots, \infty$, are orthogonal to each other and are also called functional principal components (FPCs), and α_j is called the jth FPC score. The FPC score α_j are uncorrelated random variables with mean 0 and variance λ_j. It can also be calculated as $\alpha_j = \int_{\mathscr{T}} (X(t) - \mu(t))\xi_j(t)dt$. For the rest of this paper, we assume $\mu(t) \equiv 0$ without loss of generality.

In practice, we usually select the first several leading FPCs to approximate each random curve $X(t)$. Here we denote the number of FPCs chosen as p and we will discuss how to determine p later in this manuscript. Then the representation in (2) reduces to

$$X(t) = \sum_{j=1}^{p} \alpha_j \xi_j(t) = \alpha^T \xi(t), \tag{3}$$

in which $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ and $\xi(t) = (\xi_1(t), \xi_2(t), \ldots, \xi_p(t))^T$. Substituting (3) into (1) gives

$$\begin{aligned} E(Y|X(t)) &= g\left(\beta_0 + \int_{\mathscr{T}} \beta(t) \left(\sum_{j=1}^{p} \alpha_j \xi_j(t) \right) dt \right) \\ &= g\left(\beta_0 + \sum_{j=1}^{p} \alpha_j \int_{\mathscr{T}} \beta(t) \xi_j(t) dt \right). \end{aligned} \tag{4}$$

In the meanwhile, we can also express $\beta(t)$ as a linear combination of the p leading FPCs as

$$\beta(t) = \sum_{j=1}^{p} \gamma_j \xi_j(t) = \gamma^T \xi(t), \tag{5}$$

in which $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ is an unknown coefficient vector to be estimated from the data. Plugging (5) back into (4), we have

$$E(Y|X(t)) = g\Big(\beta_0 + \alpha^T \gamma\Big). \tag{6}$$

We can estimate the unknown coefficient vector γ by regressing Y on the FPC scores α.

Determining an appropriate value of p is a difficult task in practice. One common strategy is first choosing a large value of p such that the leading p FPCs in (2) together explain more than 99% of the total variation. More formally,

$$p = \inf\{k : \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j} \geq 99\%\}.$$

If the resulting p is too large compared to the sample size, one of the popular shrinkage techniques such as LASSO and SCAD can be employed to do the variable selection. However, this procedure's prediction performance is still not satisfying in many complex problems due to three reasons. First, the prediction power of those FPCs might not coincide with the amount of variation they account for. For instance, the response variable might only depends on the 10th FPC instead of any of the first 9 FPCs; Second, given a small sample size, a large value of p introduces excessive variability into the model, making the model selection a very difficult task. Particularly, in practice when multiple functional predictors exist in the model, even with only a small number of FPCs selected for each functional predictors, this large-p-small-n problem is still quite common. Thus, there is necessity to boost the prediction power of the estimated FPCs for each functional predictor.

3 Method

We first consider the scenario when the response variable is continuous and then extend to the case in which the response variable is binary.

3.1 Supervised FPCA

Without loss of generality, we assume $E(X(t)) = 0$ and $E(Y) = 0$ in the following discussion. One can always centralize $X(t)$ and Y to satisfy these two assumptions. We propose to estimate FPCs: $\xi_1(t), \xi_2(t), \ldots$, such that the estimate $\hat{\xi}_k(t)$ maximizes

$$Q(\xi) = \frac{\theta\langle \xi, \hat{\mathscr{C}}\xi \rangle + (1-\theta)\mathsf{cov}^2(Y, \langle X, \xi \rangle)}{||\xi||^2}, \tag{7}$$

subject to $||\xi|| = 1$, $\langle \xi, \hat{\xi}_j \rangle = 0$, for every $j < k$, and $0 \leq \theta \leq 1$. Here the norm $||\xi|| = \sqrt{||\xi||^2} = \sqrt{\langle \xi, \xi \rangle}$ and $\langle f, g \rangle$ denotes the usual $\mathscr{L}^2$ inner product $\langle f, g \rangle = \int_{\mathscr{T}} f(t)g(t)dt$. In addition, $\hat{\mathscr{C}}$ is denoted as the empirical covariance operator:

$$\hat{\mathscr{C}}\xi = \int_{\mathscr{T}} \hat{C}(\cdot, t)\xi(t)dt,$$

where the empirical covariance function $\hat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t)$, and $X_i(t)$ is an independent realization of the stochastic process $X(t)$.

We take a closer look at the formalization of $Q(\xi)$ shown in (7). The first term in the numerator, $\langle \xi, \hat{\mathscr{C}}\xi \rangle$, represents the variation within the functional predictor $X(t)$ that can be explained by $\xi(t)$; the second part in the numerator, $\mathsf{cov}^2(Y, \langle X, \xi \rangle)$, represents the squared covariance between the corresponding FPC score $\langle X, \xi \rangle$ and the response variable Y. The balance between these two terms is governed by the weight parameter θ. Apparently, specifying $\theta = 1$ will give rise to unsupervised FPCA. On the other hand, specifying θ less than 1 will lead to supervised FPCA. The weight parameter θ can be treated as a tuning parameter and can be determined using cross-validation.

It is also worth mentioning the main rationale behind the 'squared' covariance, the second term on the numerator in (7), is two-fold. First, we wish to keep this term, which describes the association between the estimated FPC score and the response variable, of the numerator in equation (8) positive, since the variance of the FPC scores in the first term is always positive. Second, the 'squared covariance' also help to convert the estimation process into an eigenvalue decomposition problem, which will be illustrated in more details in Section 3.3

3.2 Smooth Supervised FPCA

The FPCs obtained using (7) might need to be further smoothed or regularized. We define another type of norm as $||f||_\lambda = \sqrt{||f||^2 + \lambda||\mathscr{D}^2 f||^2}$, in which $\mathscr{D}^2 f = \int_{\mathscr{T}} f''(t)dt$. The smooth estimate for the k-th supervised FPC is obtained by maximizing

$$Q(\xi) = \frac{\theta\langle \xi, \hat{\mathscr{C}}\xi \rangle + (1-\theta)\mathsf{cov}^2(Y, \langle X, \xi \rangle)}{||\xi||_\lambda^2}, \tag{8}$$

subject to $||\xi||_\lambda = 1$, $\langle \xi, \hat{\xi}_j \rangle = 0$, for every $j < k$, and $0 \leq \theta \leq 1$. The smoothing parameter λ controls the degree of smoothness. For instance, when $\lambda = 0$, there is no penalty on the roughness of the estimated component $\hat{\xi}(t)$ and the smooth supervised FPCs will reduce to the regular supervised FPCs discussed in Section 3.1. On the other hand, a vary large value of λ will force the estimated component $\hat{\xi}(t)$ taking a linear form. Moreover, this method is very easy to implement once the smoothing parameter λ is determined. In addition, Silverman et al. (1996) showed that under appropriate conditions the estimated FPCs were consistent. In the rest of this section, we will focus on the details of estimating the smooth supervised FPCs, which can be easily applied to unsmooth supervised FPCs by setting $\lambda = 0$.

3.3 Computational Details

In this subsection, we give the computational details on how to estimate the smooth supervised FPC $\xi(t)$ given a set of value for (θ, λ). To distinguish them, we call θ and λ as the weight and smoothing parameters, respectively. To ease the computation, we use the same B-spline basis functions $\phi_1(t), \phi_2(t), ..., \phi_M(t)$ to represent both the smooth supervised FPC $\xi_j(t)$ and the functional predictor $X_i(t)$, in which M denotes the total number of basis functions. Note that our method is not restricted to B-spline basis system and can be extended to other basis systems as well. Let $\Phi(t)$ denote the column vector $(\phi_1(t), \phi_2(t), ..., \phi_M(t))^T$, and rewrite $(X_1(t), X_2(t), ..., X_n(t))^T = \mathbf{S}\Phi(t)$, where $\mathbf{S}$ is an $n \times M$ coefficient matrix. In addition, we represent $\xi(t) = \sum_{m=1}^{M} \beta_m \phi_m(t)$ $= \beta^T \Phi(t)$, in which β denotes the coefficient vector $(\beta_1, \beta_2, ..., \beta_M)^T$. Then the empirical covariance function can be expressed as

$$\hat{C}(s,t) = \frac{1}{n} \Phi(t)^T(s) \mathbf{S}^T \mathbf{S} \Phi(t).$$

Thus the first term in the numerator of (7) is given by

$$\langle \xi, \mathscr{C} \xi \rangle = \frac{1}{n} \beta^T \mathbf{W} \mathbf{S}^T \mathbf{S} \mathbf{W} \beta, \tag{9}$$

where $\mathbf{W}$ is an $M \times M$ matrix with elements $w_{ij} = \langle \phi_i(t), \phi_j(t) \rangle$.

As for the second term in the numerator in (7), we first derive the form of the FPC score $\langle X_i, \xi \rangle$. For each $X_i(t)$, the FPC score $\langle X_i, \xi \rangle$ is written as

$$\langle X_i, \xi \rangle = \beta^T \mathbf{W} \mathbf{S}_i = \beta^T \mathbf{W}_i,$$

where $\mathbf{S}_i$ is the i-th row of the coefficient matrix $\mathbf{S}$, and $\mathbf{W}_i = \mathbf{W}\mathbf{S}_i$. Thus, combining all the scores for each $X_i(t)$

$$(\langle X_1, \xi \rangle, \langle X_2, \xi \rangle, \dots, \langle X_n, \xi \rangle)^T = \mathbf{S}\mathbf{W}\beta.$$

Finally the covariance term between Y and the FPC score is written as

$$\mathtt{cov}(Y, \langle X, \xi \rangle) = \frac{1}{n} \beta^T \sum_{i=1}^{n} Y_i W_i = \frac{1}{n} \beta^T \mathbf{W} \mathbf{S}^T \mathbf{Y},$$

in which $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$.

The squared covariance between Y and the FPC score is given as

$$\mathtt{cov}^2(Y, \langle X, \xi \rangle) = \frac{\beta^T \mathbf{M} \mathbf{M}^T \beta}{n^2}, \tag{10}$$

in which $\mathbf{M} = \mathbf{W}\mathbf{S}^T\mathbf{Y}$.

For the denominator part in (7), the norm of $\xi(t)$ is given by

$$||\xi||_\lambda^2 = \beta^T \mathbf{W} \beta + \lambda \beta^T \mathbf{D} \beta = \beta^T \mathbf{G} \beta, \tag{11}$$

where $\mathbf{D}$ denotes a $M \times M$ matrix with element $d_{ij} = \langle \mathscr{D}^2 \phi_i(t), \mathscr{D}^2 \phi_j(t) \rangle$ and $\mathbf{G} = \mathbf{W} + \lambda \mathbf{D}$.

Putting (9), (10) and (11) together, $Q(\xi)$ in (8) is recast into

$$Q(\xi) = \frac{\beta^T \mathbf{U} \beta}{\beta^T \mathbf{G} \beta},$$

where

$$\mathbf{U} = \frac{\theta}{n} \mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W} + \frac{1-\theta}{n^2} \mathbf{M}\mathbf{M}^T.$$

Let $\delta = \mathbf{G}^{\frac{1}{2}} \beta$, maximizing $Q(\xi)$ is equivalent to maximizing $\delta^T (\mathbf{G}^{-1/2})^T \mathbf{U} G^{-1/2} \delta$ subject to $\delta^T \delta = 1$. Then $\delta_1, \dots, \delta_J$ will be the leading J eigenvector of the matrix

$$(\mathbf{G}^{-1/2})^T \mathbf{U} \mathbf{G}^{-1/2}.$$

Consequently, one can derive $\widehat{\beta}_j = (\mathbf{G}^{1/2})^{-1} \delta_j$. The corresponding smooth supervised FPC is $\hat{\xi}_j(t) = \widehat{\beta}_j^T \Phi(t)$ for $j = 1, \dots, J$.

3.4 Binary Response Variable

When the response variable Y is binary, we suggest to replace $\mathtt{cov}^2(Y, \langle X, \xi \rangle)$ in $Q(\xi)$ defined in (8) with the between-group variation of the FPC scores. Formally, let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$, in which $Y_i \in \{0, 1\}, i = 1, \dots, n$, and n_j is the number of Y_i satisfying $Y_i = j$ for $j = 0, 1$. Let $\alpha = (\alpha_1, \alpha_2 \dots, \alpha_n)^T$ denote the vector of FPC scores for one FPC $\xi(t)$, in which $\alpha_i = \langle X_i, \xi \rangle$, and $\bar{\alpha}_j = \frac{1}{n_j} \sum_{\{i: Y_i = j\}} \alpha_i$. Since in this article we assume the mean function of the functional predictor, $E(X(t)) = \mu(t) = 0$, the expectation of the FPC score $E(\alpha) = \langle \mu, \xi \rangle = 0$. The between-group variation of the FPC scores is

$$\begin{aligned} R(\xi) &= \sum_{j=0}^{1} n_j (\bar{\alpha}_j - E(\alpha))^2 = n_1 \bar{\alpha}_1^2 + n_0 \bar{\alpha}_0^2 \\ &= \frac{1}{n_1} (\sum_{i=1}^{n} Y_i \alpha_i)^2 + \frac{1}{n_0} (\sum_{i=1}^{n} ((1 - Y_i) \alpha_i))^2. \end{aligned}$$

Note that $\sum_{i=1}^{n} Y_i\alpha_i = \alpha^T\mathbf{Y} = \beta^T\mathbf{W}\mathbf{S}^T\mathbf{Y}$, thus

$$\frac{1}{n_1}(\sum_{i=1}^{n} Y_i\alpha_i)^2 = \frac{1}{n_1}\beta^T\mathbf{M}_1\mathbf{M}_1^T\beta,$$

in which $\mathbf{M}_1 = \mathbf{W}\mathbf{S}^T\mathbf{Y}$. Similarly,

$$\frac{1}{n_0}(\sum_{i=1}^{n}((1-Y_i)\alpha_i))^2 = \frac{1}{n_0}\beta^T\mathbf{M}_2\mathbf{M}_2^T\beta,$$

in which $\mathbf{M}_2 = \mathbf{W}\mathbf{S}^T(\mathbf{I}_n - \mathbf{Y})$. Eventually, the between-group variation $R(\xi)$ can be expressed as a quadratic form of β:

$$\begin{aligned} R(\xi) &= \frac{1}{n_1}\beta^T\mathbf{M}_1\mathbf{M}_1^T\beta + \frac{1}{n_0}\beta^T\mathbf{M}_2\mathbf{M}_2^T\beta \\ &= \beta^T\left(\frac{1}{n_1}\mathbf{M}_1\mathbf{M}_1^T + \frac{1}{n_0}\mathbf{M}_2\mathbf{M}_2^T\right)\beta \end{aligned}$$

Then the smooth estimate for the k-th supervised FPC is obtained by maximizing

$$Q_b(\xi) = \frac{\theta\langle\xi, \mathscr{C}\xi\rangle + (1-\theta)R(\xi)}{||\xi||_\lambda^2} = \frac{\beta^T\mathbf{U}_b\beta}{\beta^T\mathbf{G}\beta}, \; 0 \le \theta \le 1,$$

subject to $||\xi||_\lambda = 1$, $\langle\xi, \hat{\xi}_j\rangle = 0$, for every $j < k$, where

$$\mathbf{U}_b = \frac{\theta}{n}\mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W} + (1-\theta)\left(\frac{1}{n_1}\mathbf{M}_1\mathbf{M}_1^T + \frac{1}{n_0}\mathbf{M}_2\mathbf{M}_2^T\right).$$

Let $\delta = \mathbf{G}^{\frac{1}{2}}\beta$. It is equivalent to maximize $\delta^T(\mathbf{G}^{-1/2})^T\mathbf{U}_b\mathbf{G}^{-1/2}\delta$, subject to $\delta^T\delta = 1$. Then $\delta_1, \ldots, \delta_J$ will be the the leading J eigenvector of the matrix

$$(\mathbf{G}^{-1/2})^T\mathbf{U}_b\mathbf{G}^{-1/2}.$$

Consequently, one can derive the estimate for the vector of basis coefficients $\widehat{\beta}_j = (\mathbf{G}^{1/2})^{-1}\delta_j$. The corresponding estimate for the j-th smooth supervised FPC is $\hat{\xi}_j(t) = \widehat{\beta}_j^T\Phi(t)$ for $j = 1, \ldots, J$.

3.5 Functional Regression

With the estimated first leading p FPCs, i.e., $\hat{\xi}_1(t), \hat{\xi}_2(t), \ldots, \hat{\xi}_p(t)$, one can fit a functional regression model between the functional predictor $X(t)$ and the response Y as discussed in Section 1. More specifically,

$$E(Y|X(t))) = g\left(\beta_0 + \int_{\mathscr{T}}\beta(t)X(t)dt\right), \qquad (12)$$

in which $g(\cdot)$ is the link function. It is usually chosen as the inverse logit function if Y is binary and the identify function if Y is continuous. One can follow the same strategy described in Section 1 to express the unknown coefficient function

$$\beta(t) = \gamma^T\hat{\xi}(t),$$

in which the unknown coefficient vector γ can be estimated by maximizing the likelihood function of Y with the mean expressed in terms of FPCs and FPC scores

$$E(Y|X(t)) = g\left(\beta_0 + \sum_{j=1}^{p}\gamma_j\int_{\mathscr{T}}\hat{\xi}_j(t)\mu(t)dt + \alpha^T\gamma\right).$$

The number of FPCs, denoted by p, used in the functional regression can be considered as a tuning parameter. We recommend to determine the value of p in the following way. We start with the number of FPCs $p = 1$ and obtain the cross-validation error as p increases. Our experience suggests choosing the value of p when the cross-validation error stops decreasing significantly. For example, one can conduct a paired t-test between the cross-validation errors for p and $p+1$. If no significant improvement is observed, we choose p as the optimal value. This rule is valid because the estimated first supervised FPC always has larger prediction ability than the second supervised FPC, and so forth.

Our method can also be extended to accommodate multiple function predictors. Suppose there are Q functional predictors: $X^{(1)}(t), \ldots, X^{(Q)}(t)$, then the multiple functional regression model can be expressed as

$$E(Y) = g\left(\beta_0 + \sum_{q=1}^{Q}\int_{\mathscr{T}}\beta^{(q)}(t)X^{(q)}(t)dt\right).$$

We can conduct the smooth supervised FPCA for each functional predictor $X^{(q)}(t)$, $q = 1, \ldots, Q$, and estimate the FPCs for $X^{(q)}(t)$. We denote the first q_p estimated FPCs for the functional predictor $X^{(q)}(t)$ as $\hat{\xi}^{(q)}(t) = (\hat{\xi}_1^{(q)}(t), \ldots, \hat{\xi}_{q_p}^{(q)}(t))$ with the corresponding score vector $\alpha^{(q)}$ and the coefficient function $\beta^{(q)}(t) = (\gamma^{(q)})^T\hat{\xi}^{(q)}(t)$. Then the unkown coefficient vector $\gamma^{(q)}, q = 1, \ldots, Q$, can also be estimated by maximizing the likelihood function of Y with the mean expressed in terms of FPCs and FPC scores

$$E(Y) = g\left(\beta_0 + \sum_{q=1}^{Q}\sum_{j=1}^{q_p}\gamma_j^{(q)}\int_{\mathscr{T}}\hat{\xi}_j^{(q)}(t)\mu^{(q)}(t)dt + \sum_{q=1}^{Q}(\alpha^{(q)})^T\gamma^{(q)}\right), \qquad (13)$$

in which $\mu^{(q)}(t)$ represents the mean trajectory for $X^q(t)$.

In practice, when multiple functional predictors exist, the number of total FPCs is sometimes close or larger than the sample size. In this case, we recommend to employ one of those popular variable selection tools such as LASSO or SCAD to estimate the model. We will demonstrate this procedure in our real data application with a binary response variable.

4 Application

We apply our method to analyze an electroencephalography (EEG) dataset. The EEG dataset, collected by Zhang et al.

(1995), is used to study the genetic predisposition to alcoholism. The original dataset is available in UCI machine learning repository (`https://archive.ics.uci.edu/ml/datasets/EEG+Database`). In total 122 subjects are separated into two groups: alcoholic and control. Each subject is exposed to two non-matching stimuli, i.e., two different pictures. In addition, 64 electrodes are placed on each subject's scalp to record the brain activities. Each electrode is sampled at 256Hz for 1 second. Our goal here is to classify alcoholic and control subjects based on their brain activities.

The number of trials with two non-matching stimuli ranges from 10 to 30 from subject to subject. Figure 1 displays the measurements averaged over all trials for 15 randomly selected subjects in both alcoholic and control groups at 256 time points at one sensor called the AF1 channel.

For each subject, we randomly select 2/3 of the total trials as the training trials and the rest 1/3 of the trials as the test trials. Then, the training and test observations are computed as the average of all training and test trials, respectively. We then apply our smooth supervised FPCA method to estimate the first leading p FPCs for each sensor.

Next, we fit a multiple functional logistic regression

$$\operatorname{logit}\{P(Y=1)\} = \beta_0 + \sum_{q=1}^{Q} \int_{\mathscr{T}} \beta^{(q)}(t) X^{(q)}(t) dt, \tag{14}$$

where $Y = 0, 1$ correspond to the control and alcoholic subject, respectively, and $X^{(q)}(t)$ is the brain activity for the q-th sensor. Following the method outlined in Subsection 3.5, we add an L_1 penalty on the coefficients for the slope function $\beta^{(q)}(t)$. In the context of a binary response, equation (13) becomes

$$\operatorname{logit}\{P(Y=1)\} = \beta_0' + \sum_{q=1}^{Q} (\alpha^{(q)})^T \gamma^{(q)},$$

where $\beta_0' = \beta_0 + \sum_{q=1}^{Q} \sum_{j=1}^{p} \gamma_j^{(q)} \int_{\mathscr{T}} \xi_j^{(q)}(t) \mu^{(q)}(t) dt$. Then the penalized log likelihood function is written as

$$\begin{aligned} l(\beta_0', \gamma_1, \ldots, \gamma_q) = & \frac{1}{n}\left(n_1 \log(p_i) + n_0 \log(1 - p_i) \right) \\ & + \lambda_L \left(|\beta_0'| + \sum_{q=1}^{Q} \sum_{j=1}^{p} |\gamma_j^{(q)}| \right) \end{aligned}$$

where $p_i = \Pr(Y_i = 1 | X_i^{(1)}(t), \ldots, X_i^{(Q)}(t)) = \text{inv-logit}(\beta_0' + \sum_{q=1}^{Q} (\alpha^{(q)})^T \gamma^{(q)})$ and $n_1 = \sum Y_i$. The tuning parameter, λ_L, in the LASSO penalty is chosen using a five-fold cross validation. For supervised FPCA, the weight parameter θ and the smoothing parameter λ are selected from a 9-by-6 meshgrid, i.e., $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] \times [1, 10, 10^2, 10^3, 10^4, 10^5]$. Both of them are determined by a five-fold cross validation using the training data set simultaneously with the sparsity parameter, λ_L, in the LASSO penalty. We also apply the unsupervised FPCA method, in which the weight parameter θ is always set to be 1, and the smoothing parameter λ is selected from $[1, 10, 10^2, 10^3, 10^4, 10^5]$ by a five-fold cross validation simultaneously with the sparsity parameter, λ_L, in the LASSO penalty. After obtaining the estimate $\hat{\beta}_0$ and $\hat{\beta}^{(q)}(t)$ for the multiple functional logistic regression (14), we classify the subjects on the test data, and obtain the corresponding classification error.

Table 1 The means and standard deviations of the classification error on testing set in 100 random data splitting using both supervised FPCA and unsupervised FPCA in the EEG data application. Here sFPCA and FPCA stands for supervised FPCA and unsupervised FPCA respectively.

		# of FPCs			
Method	classification error	1	2	3	4
sFPCA	mean	0.208	0.200	0.180	0.166
	sd	0.018	0.020	0.031	0.024
FPCA	mean	0.266	0.217	0.207	0.212
	sd	0.024	0.024	0.032	0.026

We repeat the above process for 100 replicates of random data splittings and summarize the test classification errors. Table 1 shows the mean and standard deviation of the classification errors when the number of FPCs is selected for each sensor varies from 1 to 4 for supervised FPCA and unsupervised FPCA. It shows that supervised FPCA has a higher classification accuracy than unsupervised FPCA. For instance, supervised FPCA improves the classification accuracy by about 20%, when just using one FPC, in comparison with unsupervised FPCA. As one reviewer points out, it is not clear whether the difference of the misclassification rate between FPCA and sFPCA is statistically significant.

5 Simulation Studies

Three different simulations are conducted to evaluate the proposed method. We first briefly introduce the generation mechanism for the functional predictor $X(t)$ in the beginning of this section, since this generation mechanism stays the same across different simulations. Then we discuss each simulation in details. We also do two more simulation studies to compare our proposed supervised FPCA method with three alternative methods including supervised PCA proposed by Bair et al. (2006), SupSVD (Li et al., 2015) and SupSFPC (Li et al., 2016). The results for these two additional simulation studies are provided in the supplementary files.

In order to make the simulation setting similar to real data, we use four FPCs, shown in Figure S1 in the supplementary document, to generate sample functional predictors. They are the first four leading FPCs estimated from the Canadian weather data (Ramsay et al., 2009), which consist

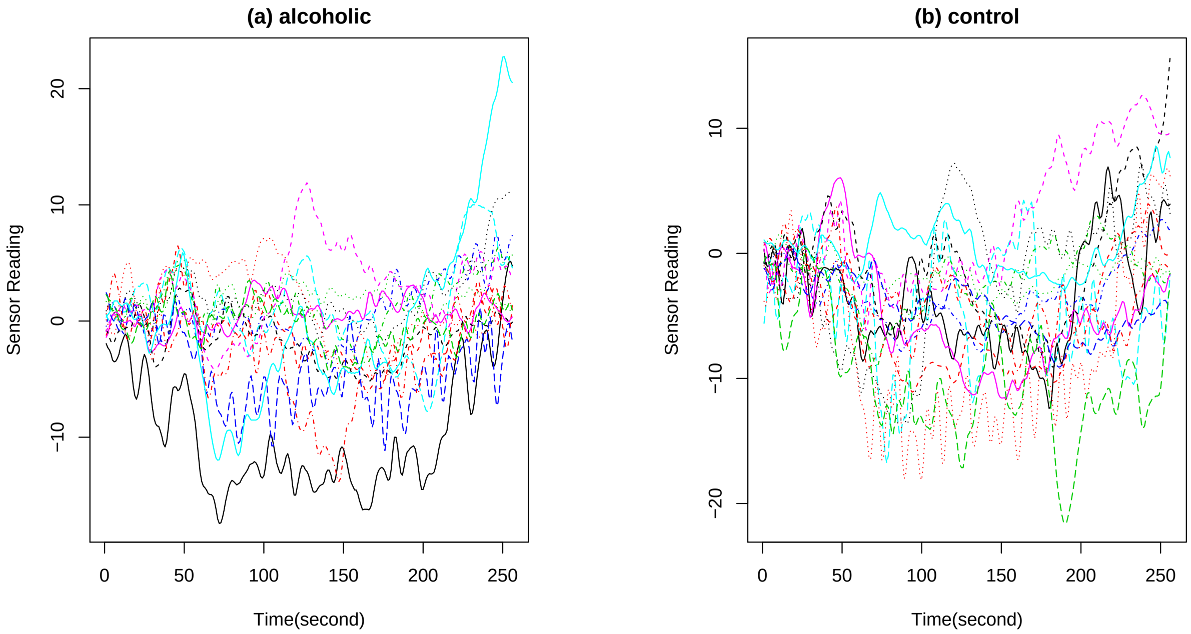

Fig. 1 The readings of the brain activities at the AF1 channel for 15 randomly selected alcoholic subjects (panel a) and 15 randomly selected control subjects (panel b). All of them are exposed to two non-matching stimuli in an EEG case study on genetic predisposition to alcoholism.

of daily temperature measurements at 35 weather stations across Canada. Each functional predictor $X_i(t), i = 1, \ldots, n$, is simulated as: $X_i(t_k) = \alpha_{1i}\xi_1(t_k) + \alpha_{2i}\xi_2(t_k) + \alpha_{3i}\xi_3(t_k) + \alpha_{4i}\xi_4(t_k), k = 1, 2, \ldots, 365$, where $\xi_j(t_k)$ is the j-th true FPCs, $j = 1, \ldots, 4$. The simulated FPC score is simulated as: $\alpha_i^T = (\alpha_{1i}, \alpha_{2i}, \alpha_{3i}, \alpha_{4i})^T \overset{i.i.d}{\sim} MVN(\mathbf{0}, \Sigma)$, in which $\Sigma = \text{diag}(100, 80, 50, 30)$. Figure S2 in the supplementary document displays 50 random curves simulated under these settings.

5.1 The First Simulation Study

The first simulation study is designed to evaluate the proposed method when the response variable is binary. Here we generate 1000 sample curves, $X_i(t), i = 1, \ldots, 1000$. The response variable Y is generated as:

$$Y_i \sim \text{Bernoulli}(p_i),$$
$$\text{logit}(p_i) = \int_{\mathscr{T}} \beta(t) X_i(t) dt, i = 1, \ldots, 1000,$$

in which $\beta(t) = \xi_4(t)$. In other words, the binary response Y is only related to the fourth FPC $\xi_4(t)$. We randomly select 200 samples as the test set and used the other 800 samples as the training set. For supervised FPCA method, the weight parameter θ and the smoothing parameter λ are selected on a 5-by-3 meshgrid, i.e., $[0.1, 0.3, 0.5, 0.7, 0.9] \times [10, 10^3, 10^5]$, through a five-fold cross validation using those 800 training samples only. For unsupervised FPCA method, the weight parameter θ is fixed to be 1 under different values of λ and the smoothing parameter was selected from 10, 10^3 and 10^5 using a five-fold cross validation as well.

We compare the prediction performance of supervised FPCA with unsupervised FPCA in terms of classification errors on the test data in 100 simulation. Figure 2 summarizes the classification errors. Supervised FPCA yields a much lower classification error than unsupervised FPCA when the number of FPCs used, p, is less than 3. More specifically, the mean test classification error of unsupervised FPCA is slightly less than 50% unless choosing four FPCs, whereas the mean classification error of supervised FPCA is constantly less than 14% even when the number of FPCs is less than 3. This shows that supervised FPCA is able to detect the FPCs that are most related with the response variable in advance and our method can well accommodate the binary response.

To gain some insight, in Figure 3, we compare the first FPC $\hat{\xi}_1(t)$ estimated by supervised and unsupervised FPCA in one simulation run, along with the true FPC used to simulate the response variable. We can see that the first FPC estimated by supervised FPCA method is much closer to the true FPC, in comparison with the first FPC estimated by unsupervised FPCA method.

5.2 The Second Simulation Study

We conduct three simulation scenarios to evaluate the proposed method in different settings when the response vari-

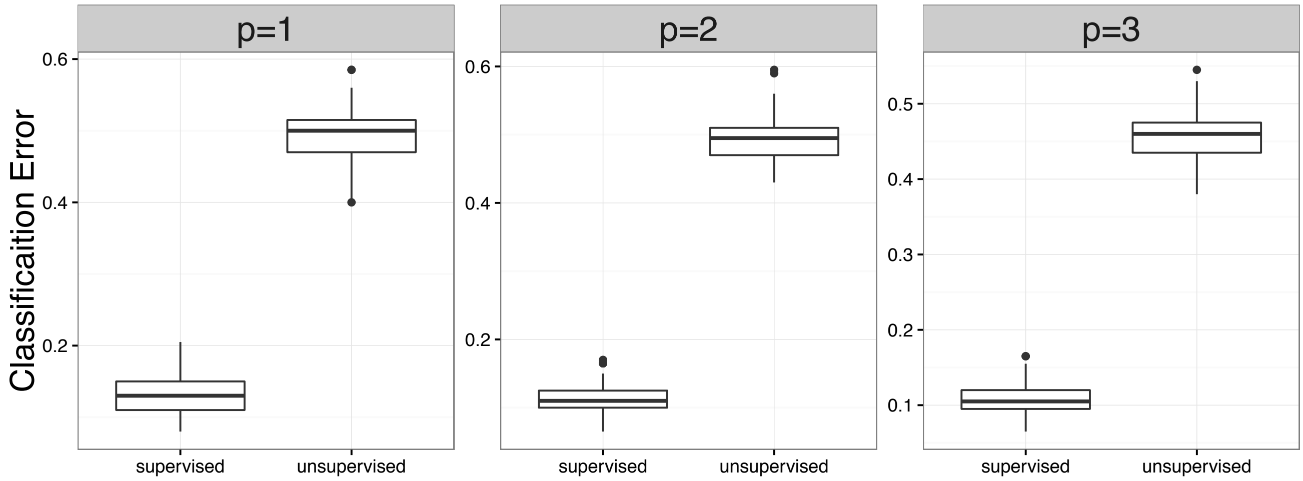

Fig. 2 Boxplots of the classificaition errors for 100 simulation runs when using the first p FPCs estimated by supervised and unsupervised FPCA in the first simulation study when the response variable is binary.

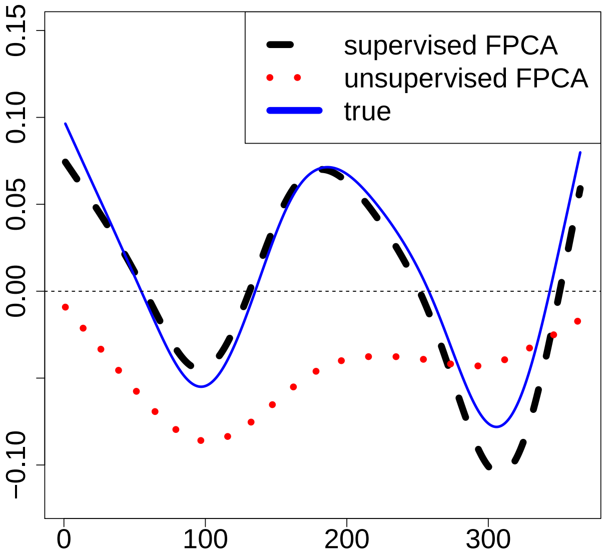

Fig. 3 The first FPC estimated with supervised and unsupervised FPCA in one simulation run of the first simulation study when the response variable is binary.

able is continuous. Here we generate 100 sample curves, $X_i(t), i = 1, \ldots, 100$, in the same way as discussed in the beginning of this section. The response variable Y is generated using the functional linear regression model (1) with $\beta(t)$ being specified as $\beta(t) = \gamma_1\xi_1(t) + \gamma_2\xi_2(t)\xi_j(t) + \gamma_3\xi_3(t) + \gamma_4\xi_4(t) = \gamma^T\xi(t)$, where $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)^T$ and $\xi(t) = (\xi_1(t), \xi_2(t), \xi_3(t), \xi_4(t))^T$. In addition, the link function $g(\cdot)$ is the identity function. Without loss of generality we set $\beta_0 = 0$.

5.2.1 Scenario 1

In the first scenario, we set the true $\gamma = (0,0,0,1)^T$ such that the true slope function $\beta(t) = \xi_4(t)$. In other words, the response variable Y is only related to the fourth leading FPC $\xi_4(t)$. In addition, the noise term ε follows a normal distribution $N(0, 30\rho)$, in which ρ denotes the signal-to-noise ratio. We set $\rho = 5\%$ and 50%. We randomly select 20 samples as the test set and treat the other 80 samples as the training set. Both the smoothing parameter and the weight parameter are chosen via five-fold cross validation using the training samples only on the same meshgrid used in the previous simulation. As for the unsupervised FPCs, the weight parameter θ is set to be 1. The smoothing parameter λ is selected from $\{10, 10^3, 10^5\}$ using a five-fold cross validation. For unsupervised FPCA method, the weight parameter θ is set to be 1. We compare the prediction performance of the supervised FPCs with that of the unsupervised FPCs using 500 simulation runs. The prediction error is evaluated using relative mean squre error (RAMSE) defined as

$$\texttt{RAMSE} = \frac{\sum_{\ell=1}^{n} (\hat{y}_\ell - y_\ell)^2}{\sum_{\ell=1}^{n} (\bar{y} - y_\ell)^2}. \qquad (15)$$

Here y_ℓ and $\hat{y}_\ell$ denote the observed ℓth response in the test set, respectively, and $\bar{y}$ represents the average of theose observed responses the training set.

Figure 4 summarizes the prediction RAMSEs for 100 repeated runs when the noise-to-signal ratio $\rho = 5\%$. As we can see, supervised FPCA method consistently give lower RAMSE compared with unsupervised FPCA when p is less than 3. More specifically, when $p < 4$, the unsupervised FPCs perform no better than simply using the sample mean of the training set as the average prediction error is constantly around 100%. In contrast, the supervised FPCs is able to capture the information of the response variable and improve its prediction performance accordingly. For example, even restricting only one FPC in the functional linear regression, the average RAMSE is less than 45%, only half of the average RAMSE of the unsupervised FPCs.

To gain some insight, Figure 5 displays the first FPC $\hat{\xi}_1(t)$ estimated by supervised and unsupervised FPCA along with the true FPC related to the response variable when the noise-to-signal ratio of the data is $\rho = 5\%$. We can see the first FPC estimated by supervised FPCA is much more closer to the true FPC compared with the first FPC estimated by unsupervised FPCA. This indicates that supervised FPCA is able to detect the FPC that is truly related to the continuous response variable.

Figure S3 in the supplementary document displays the boxplots of the prediction RAMSEs when the simulation data have the noise-to-signal ratio as $\rho = 50\%$. It shows that supervised FPCA yielded a more robust estimator since the mean RAMSE is only increased about 20% when the noise-to-signal ratio of the simulated data is increased from 5% to 50%. The detailed results are available in the supplementary materials.

5.2.2 Scenario 2

The only difference between this scenario and the previous one in section 5.2.1 is that we specify $\gamma = (0.25, 0.73, 0.29, 0.56)^T$, such that the response variable Y is related to a linear combination of all $\xi_i(t), i = 1,2,3,4$. In practice, this case might be more realistic compared to the scenario when the response is only related to a single FPC.

Figure 6 summarizes the prediction errors for 100 simulation runs when the noise-to-signal ratio of the data is $\rho = 5\%$. It shows that supervised FPCA still outperforms unsupervised FPCA when using up to 3 FPCs. More specifically, when just using one FPC, i.e. $p = 1$, unsupervised FPCA only performs slightly better than simply using the sample mean of the training set as the average RAMSE is about 91%, because unsupervised FPCA only successfully recovers the first FPC, while the response variable is correlated with all four FPCs. In contrast, the average RAMSE using supervised FPCA is only 14.6% when just using one FPC, which is quite satisfying.

The two scenarios in the second simulation study show that the prediction performance of supervised FPCA seems quite satisfactory no matter whether the response variable is related to a single FPC or a linear combination of several FPCs.

6 Concluding Remarks

In this paper, we consider the problem of predicting a scalar response variable by using one or several functional predictors. The conventional FPCA method focuses on finding FPCs that maximize the variation of FPC scores and ignores the response variable. We have proposed a one-step supervised FPCA to detect those FPCs whose scores are correlated with the response variable. The resulting FPCs have a better prediction performance compared to the conventional FPCA method.

Through our real data application and simulations, we demonstrate that our method can accommodate both continuous and binary response variable. Even through we only show examples with binary response variable, we believe that our method can be easily extended to predicting multinomial response variable. Lastly, our method is also quite user-friendly. An R package “sFPCA” has been developed to implement supervised FPCA and is available in the supplementary material.

SUPPLEMENTARY MATERIAL

Supplementary Document: This file contains some additional figures for simulation studies in Section 5, two additional simulation studies with an arbitrary coefficient function and a large number of FPCs related to the outcome, respectively. We also include another real data application analyzing the time course yeast gene expression data. (supplementary.pdf, PDF file),

demon_sFPCA.R: This file contains the R codes for demonstration of using the “sFPCA” R package.

Acknowledgements

The authors would like to thank the Editor, the Associated Editor, and two reviewers for their very constructive suggestions and comments on the manuscript. They are extremely helpful for us to improve our work. The authors are also very grateful to Prof. Gen Li and Prof. Xiaoyan Leng for kindly providing us with their data and their computing codes. This research was supported by Nie’s Postgraduate Scholarship-Doctorial (PGS-D) from the Natural Sciences and Engineering Research Council of Canada (NSERC), and the NSERC Discovery grants of Wang and Cao.

References

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006). Prediction by supervised principal components. *Journal of the American Statistical Association 101*(473).

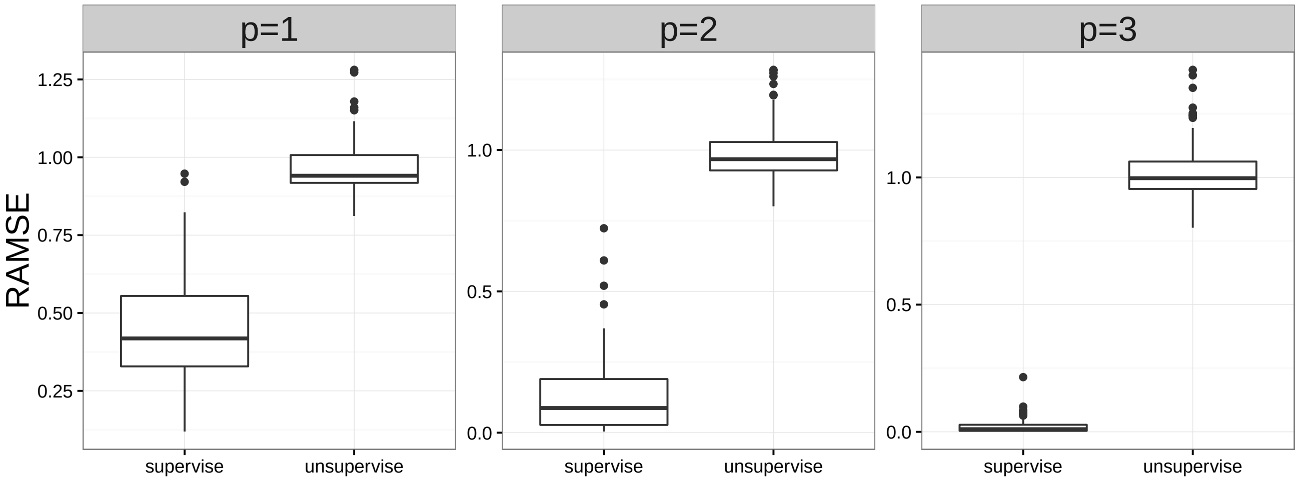

Fig. 4 Boxplots of the prediction `RAMSE`s for 100 simulation runs using the first `p` FPCs estimated by supervised and unsupervised FPCA in Scenario 1 of the second simulation study when the response variable is continuous.

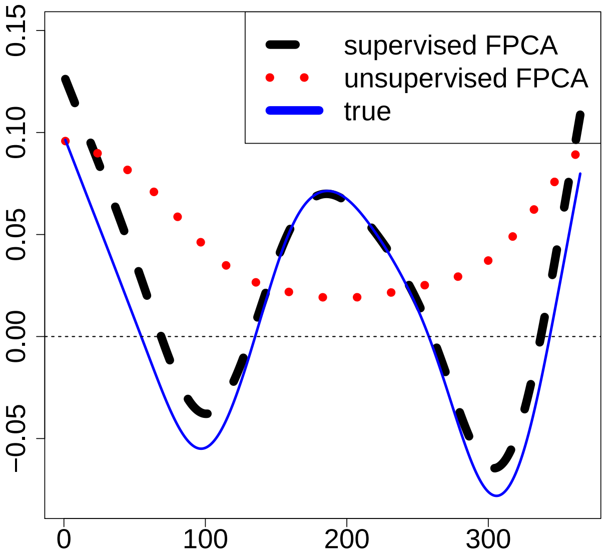

Fig. 5 The first FPC estimated with supervised and unsupervised FPCA at one simulation run in Scenario 1 of the second simulation study when the response variable is continuous.

Cardot, H., R. Faivre, and M. Goulard (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics 30*(10), 1185–1199.

Fukunaga, K. and W. L. Koontz (1970). Representation of random processes using the finite karhunen-loeve expansion. *Information and Control 16*(1), 85–101.

Huang, J. Z., H. Shen, and A. Buja (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association 104*(488).

Li, G., H. Shen, and J. Z. Huang (2016). Supervised sparse and functional principal component analysis. *Journal of Computational and Graphical Statistics 25*(3), 859–878.

Li, G., D. Yang, A. B. Nobel, and H. Shen (2015). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*.

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

Ramsay, J., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer New York.

Ramsay, J. O. and B. W. Silverman (2002). *Applied functional data analysis: methods and case studies*, Volume 77. Springer New York.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). New York: Springer.

Ratcliffe, S. J., G. Z. Heller, and L. R. Leader (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in medicine 21*(8), 1115–1127.

Silverman, B. W. et al. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics 24*(1), 1–24.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Zhang, X. L., H. Begleiter, B. Porjesz, W. Wang, and A. Litke (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin 38*(6), 531–

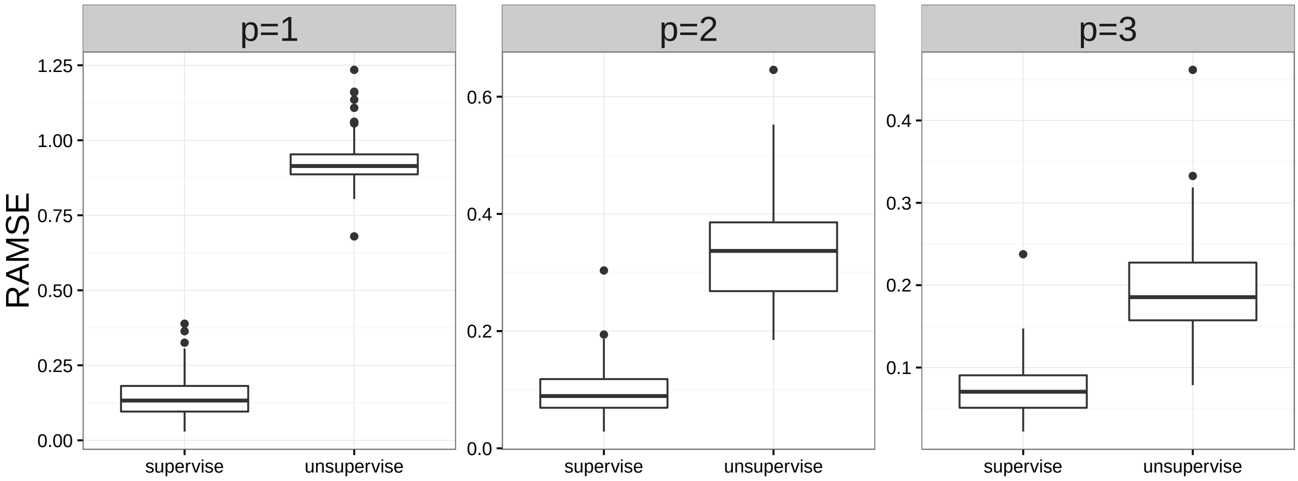

Fig. 6 Boxplots of the prediction `RAMSE`s for 100 simulation runs using the first `p` FPCs estimated by supervised and unsupervised FPCA in Scenario 2 of the second simulation study when the response variable is continuous.

538.