Biometrics 73, 551–561
June 2017

DOI: 10.1111/biom.12624

Hypothesis Testing in Functional Linear Models

Yu-Ru Su, Chong-Zhi Di,* and Li Hsu

Biostatistics, Division of Public Health Sciences Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.
**email:* cdi@fredhutch.org

Summary. Functional data arise frequently in biomedical studies, where it is often of interest to investigate the association between functional predictors and a scalar response variable. While functional linear models (FLM) are widely used to address these questions, hypothesis testing for the functional association in the FLM framework remains challenging. A popular approach to testing the functional effects is through dimension reduction by functional principal component (PC) analysis. However, its power performance depends on the choice of the number of PCs, and is not systematically studied. In this article, we first investigate the power performance of the Wald-type test with varying thresholds in selecting the number of PCs for the functional covariates, and show that the power is sensitive to the choice of thresholds. To circumvent the issue, we propose a new method of ordering and selecting principal components to construct test statistics. The proposed method takes into account both the association with the response and the variation along each eigenfunction. We establish its theoretical properties and assess the finite sample properties through simulations. Our simulation results show that the proposed test is more robust against the choice of threshold while being as powerful as, and often more powerful than, the existing method. We then apply the proposed method to the cerebral white matter tracts data obtained from a diffusion tensor imaging tractography study.

Key words: Association-variation index; Functional association; Functional principal component analysis; Power; Wald-type tests.

1. Introduction

With functional data arising increasingly common in many scientific studies, functional data analysis (FDA; Ramsay and Silverman, 2005) has been an important area of research. In these studies, a common question is to quantify the relationship between functional/longitudinal covariates and scalar responses. Functional linear models (FLM; Ramsay and Dalzell, 1991) have been widely used to address such questions, as it allows for a dynamic association between the response and the functional covariate at different points on the support in consideration.

Testing the association of functional covariates with response is of great interest as it provides an overall assessment of the association; however, it remains challenging due to infinite dimensionality of functional covariates. To overcome this issue, a natural strategy is to reduce the dimension. For FLM it is popular to represent functional covariates and the coefficient function by linear combinations of a set of basis functions, such as a pre-specified basis system like B-splines, Fourier, or wavelet bases (James, 2002; Goldsmith et al., 2011, 2013), or data-adaptive basis functions from functional principal component analysis (FPCA; Cardot et al., 1999; Yao et al., 2005b; Goldsmith et al., 2011, 2013). Under such representations, the testing problem in FLM reduces to hypothesis testing under a classical linear model or linear mixed effects model.

We focus on testing procedures based on FPCA, since it provides parsimonious representation of functional data and is widely used. Several approaches have been proposed to test the leading principal components in FPCA for the associations using the covariance-based test (Cardot et al., 2003; Horváth and Kokoszka, 2012), or classical tests such as Wald, score or likelihood ratio (Kong et al., 2013; Swihart et al., 2014). All of these works require pre-specifying a threshold to choose the leading principal components (PCs) for inclusion in the test, where PCs are ranked based on eigenvalues, or equivalently the percentage of variance explained (PVE) for the functional covariates. However, as we will demonstrate, the PVE alone is not an optimal criterion to use for the purpose of testing, because the power is sensitive to the choice of the threshold. Different thresholds often lead to very different p-values and therefore inconsistent conclusions (e.g., the diffusion tensor imaging example in Section 5), making the statistical inference confusing and difficult to interpret for practitioners. In the literature, tests based on pre-specified spline basis representations have also been proposed, for example, a permutation F-test (Ramsay et al., 2009) and restricted likelihood ratio tests (Swihart et al., 2014). The former requires a relatively high computational cost, and the latter are shown to be outperformed by FPCA-based tests in terms of power (Swihart et al., 2014). Thus, we will focus on FPCA-based approaches in this article.

In this article, we propose a novel testing procedure that orders and selects PCs based on an association–variation index (AVI) that combines both the variation and association along each direction. The AVI is directly related to the non-centrality parameter in power functions, so using the AVI to select PCs is more desirable for the testing purpose and can potentially improve power. Compared to existing tests, the proposed procedure is more robust to the choice of tuning parameters (threshold values used to choose the number of PCs), while enjoying power gain with relatively low number

© 2017, The International Biometric Society

of selected PCs in many scenarios. In addition, we provide a comprehensive power study of classical procedures, as, to our knowledge, there is no systematic study that evaluates how the power of classical tests depends on the tuning parameters in FPCA, such as the number of PCs or the threshold of the PVE.

The rest of the article is organized as follows. In Section 2, we review existing FLM works, and discuss potential issues on classical testing procedures. We then introduce our proposed testing procedure and establish the asymptotic properties of the proposed test in Section 3. Simulation studies and results are presented in Section 4, followed by a data example of diffusion tensor imaging study in Section 5. Conclusions and future directions are discussed in Section 6.

2. Classical Testing Procedures and Power Considerations

2.1. *Notation, Models, and Classical Testing Procedures*

We consider the setting with a scalar response Y and a functional covariate $X(t) \in L^2(\mathcal{T})$ defined on a compact support $\mathcal{T} \subseteq \mathbb{R}$. Assuming that the observed sample consists of n independent subjects. Denote $\{Y_i, X_i(\cdot)\}$, $i = 1, \ldots, n$, the i.i.d. realizations of $\{Y, X(\cdot)\}$. We consider the case that $X_i(\cdot)$ is measured on a set of dense and regular grid points over $\{t_j \in \mathcal{T}: j = 1, \ldots, J\}$. The scenario for $X_i(\cdot)$ measured intermittently with error will be discussed in Section 3.

The scalar-on-function FLM model being considered has the following form,

$$Y = \alpha_0 + \int_{\mathcal{T}} \beta(t)X(t)dt + \epsilon, \tag{1}$$

where $\beta(t) \in L^2(\mathcal{T})$ is a smooth square integrable coefficient function, $\alpha_0 \in \mathbb{R}$ is the intercept, and ϵ is an error term with mean 0 and variance σ_ϵ^2. We further assume that ϵ is independent of $X(\cdot)$. Under model (1), $\beta(\cdot)$ describes the effect of the functional covariate $X(\cdot)$ on Y along t in the support $\mathcal{T}$. The statistical problem of interest in this article focuses on hypothesis testing for

$$\begin{aligned} \mathrm{H}_0 &: \beta(t) = 0, \text{for all } t \in \mathcal{T} \text{ versus } \mathrm{H}_\mathrm{a} \\ &: \beta(t) \neq 0 \text{ for } t \text{ in some open subset in} \mathcal{T}. \end{aligned}$$

This corresponds to testing the global null effect of the coefficient function $\beta(\cdot)$.

To model the coefficient function $\beta(\cdot)$ non-parametrically, a popular approach is to employ FPCA-based expansion. Specifically, the covariance matrix of the random function $X(\cdot)$ has a spectral decomposition, $G(s,t) = \mathrm{cov}[X(s), X(t)] = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where λ_k's are non-negative eigenvalues ranked in non-increasing order with $\sum_{k=1}^{\infty} \lambda_k < \infty$, and $\phi_k(\cdot)$'s are the corresponding orthonormal eigenfunctions. Using the Karhunen–Loève expansion, the random function $X(t)$ can be expressed as $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, where $\mu(\cdot)$ denotes the mean function of $X(\cdot)$, ξ_k's are principal component scores with $E(\xi_k) = 0$, $\mathrm{var}(\xi_k) = \lambda_k$, and $E(\xi_k \xi_{k'}) = 0$, for $k \neq k'$. The scores can be calculated by $\int_{\mathcal{T}} \{X(t) - \mu(t)\}\phi_k(t)\,dt$. Since the eigenfunctions provide a set of basis functions for the functional space, the coefficient function can be represented by $\beta(t) = \sum_{k=1}^{\infty} \beta_k \phi_k(t)$, where $\beta_k = \int_{\mathcal{T}} \beta(t)\phi_k(t)dt$. Applying these expansions to both $X_i(\cdot)$ and $\beta(\cdot)$, the FLM (1) can be rewritten as

$$Y = \alpha_0' + \sum_{k=1}^{\infty} \beta_k \xi_k + \epsilon, \tag{2}$$

where $\alpha_0' = \alpha_0 + \int_{\mathcal{T}} \mu(t)\beta(t)dt$. In practice, a few leading principal components (PC) often suffice to approximate the process $X(\cdot)$ well. Thus, a truncated FLM model using the first K PCs is

$$Y = \alpha_0' + \sum_{k=1}^{K} \beta_k \xi_k + \zeta, \tag{3}$$

where $\zeta = \sum_{k=K+1}^{\infty} \beta_k \xi_k + \epsilon$. Under this model, the null hypothesis becomes $\beta_1 = \cdots = \beta_K = 0$. The error term ζ has mean 0 and a larger variance $\sigma_\zeta^2 = \sum_{K+1}^{\infty} \lambda_k \beta_k^2 + \sigma_\epsilon^2$. A common approach to determine the number of PCs, K, is based on the PVE that directly depends on eigenvalues. For example, given threshold γ (e.g., 95%), K is defined as the smallest integer that satisfies $\sum_{k=1}^{K} \lambda_i / \sum_{k=1}^{\infty} \lambda_k \geq \gamma$, and is finite due to the fact that $\sum_{k=1}^{\infty} \lambda_k = \int_{\mathcal{T}} \{X(t)\}^2\,dt < \infty$.

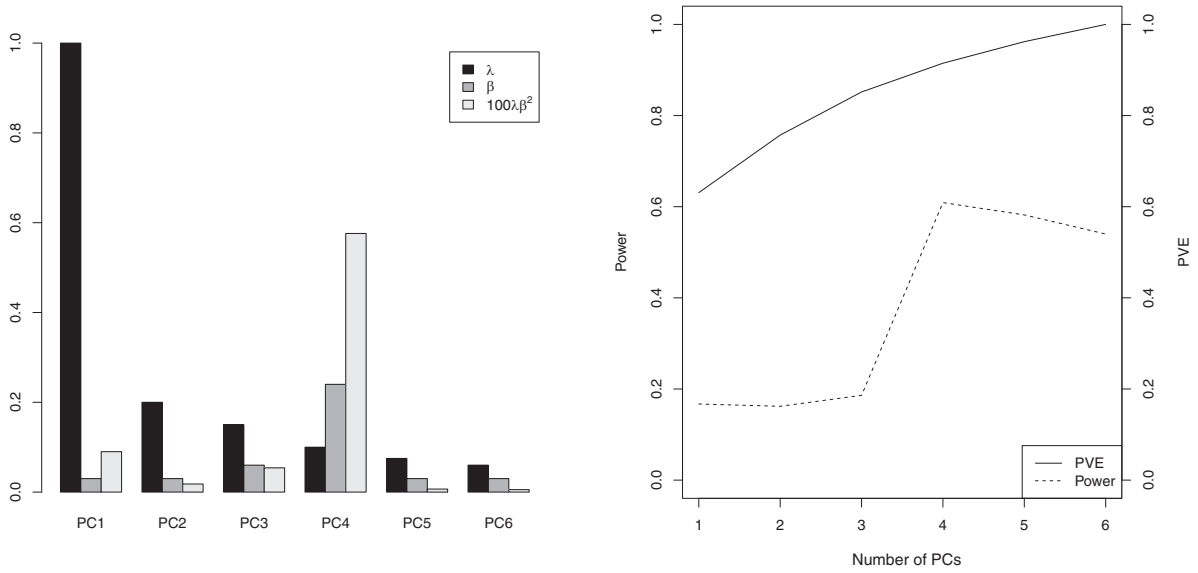
Classical testing procedures based on truncated model (3) are widely used in the literature, with asymptotic properties formally studied by Kong et al. (2013). It has been shown that eigenvalues, eigenfunctions, and PC scores can be consistently estimated via FPCA from the observed data (Hall and Hosseini-Nasab, 2006; Zhu et al., 2014). The least square estimate $\hat{\beta}_k$ obtained from fitting model (3) with ξ_k substituted by $\hat{\xi}_k$ is expressed as $\hat{\beta}_k = (\hat{\boldsymbol{\xi}}_k^T \hat{\boldsymbol{\xi}}_k)^{-1} \hat{\boldsymbol{\xi}}_k^T \mathbf{Y}$. It can be shown that the asymptotic variance of $\hat{\beta}_k$ is $\frac{\sigma_\zeta^2}{n\lambda_k}$, and $\hat{\beta}_k$'s are asymptotically independent with each other. The Wald-type test statistic is thus defined as

$$T^c = \sum_{k=1}^{K_n} \frac{\hat{\beta}_k^2}{\hat{\mathrm{var}}(\hat{\beta}_k)} = \frac{1}{\hat{\sigma}_\zeta^2} \sum_{k=1}^{K_n} \frac{\mathbf{Y}^T \hat{\boldsymbol{\xi}}_k \hat{\boldsymbol{\xi}}_k^T \mathbf{Y}}{\hat{\boldsymbol{\xi}}_k^T \hat{\boldsymbol{\xi}}_k}, \tag{4}$$

where K_n is the number of selected PCs based on estimated eigenvalues, and $\hat{\sigma}_\zeta^2$ is the estimated error variance. As shown in Kong et al. (2013), the null distribution of T^c and a χ^2-distribution with K_n degrees of freedom are asymptotically equivalent under some regularity conditions.

2.2. *Power Considerations*

The truncated model (3) effectively reduces dimensionality and allows for statistical inference using a classical linear model framework. However, the impact of tuning parameter selection (the number of PCs, or equivalently the PVE



(a) Settings on each PC

(b) PVE and powers

Figure 1. An illustrative example: functional regression model on a scalar response with a functional covariate consisting of six Fourier basis. Left: eigenvalues (λ), magnitude of associations (β), and contributions of each PC to the power function. Right: the cumulative percentage of variation explained and the power of corresponding tests versus the number of PCs selected.

threshold) on hypothesis testing, especially statistical power, is not well understood. For example, commonly used thresholds to determine the truncated model by the PVE are 80, 90, 95, and 99%. The leading PCs are selected corresponding to directions that explain the largest variation. From an estimation perspective, a higher PVE threshold generally results in a better approximation of $X(t)$ and $\beta(t)$. However, as will be demonstrated, a higher PVE threshold does not necessarily lead to a higher power in hypothesis testing.

We now study how the power of classical tests depends on the choice of the PVE threshold or the number of PCs selected. Consider a specific alternative hypothesis $\mathrm{H_a} : \beta(t) = \beta_a(t)$, where $\beta_a(t) \neq 0$ for $t$in some open subset in $\mathcal{T}$. It can be shown that the distribution of T^c and a non-central χ^2-distribution, $\chi^2_{K_n}(\eta)$, are asymptotically equivalent (Müller and Stadtmüller, 2005, Theorem 4.1), where

$$\eta = \frac{n}{\sigma_\zeta^2}\sum_{k=1}^{K_n}\lambda_k\beta_{ka}^2, \text{ where}\beta_{ka} = \int_{\mathcal{T}}\beta_a(t)\phi_k(t)dt. \tag{5}$$

Consequently, the power function of T^c is approximated by $\Pr\{\chi^2_{K_n}(\eta) \geq q_{\chi^2_{K_n},1-\alpha}\}$, where $q_{\chi^2_{K_n},1-\alpha}$ is the $(1-\alpha)\%$ quantile of $\chi^2_{K_n}$.

The formula above clearly demonstrates that the power contribution from the kth component involves both the eigenvalue λ_k as well as the magnitude of association β_k. Among all PCs, the one that contributes the most power is not the first PC with the largest λ_k, but rather the component with the largest $\lambda_k\beta_k^2$ value. Similarly, to maximize power among all truncated models with K components (i.e., fixing the degree of freedom to be K), the optimal procedure chooses those with the largest values of $\lambda_k\beta_k^2$. In contrast, classical procedures choose the first K components with the largest eigenvalues (λ_k), and thus might not perform well in terms of power, especially when the PCs that are strongly associated with the response have small variations or when leading PCs are not associated with the outcome. The potential issue of using the leading PCs for functional regression was also briefly mentioned by Zhu et al. (2014), but they focused on estimation instead of hypothesis testing.

To illustrate this phenomenon, we show a numerical example below. The eigenfunctions are six Fourier basis functions $\phi_k(t)$, with corresponding eigenvalues 1, 0.2, 0.15, 0.1, 0.075, and 0.06, respectively. The true coefficient function is $\beta(t) = 0.03\phi_1(t) + 0.03\phi_2(t) + 0.06\phi_3(t) + 0.24\phi_4(t) + 0.03\phi_5(t) + 0.03\phi_6(t)$. Figure (1a) visualizes λ_k, β_k, and $\lambda_k\beta_k^2$ along each PC. In this example, the fourth PC is most strongly associated with the response. To explore how the PVE threshold affects the power performance, we considered a set of thresholds 50, 70, 80, 90, 95, and 99%, corresponding to 1–6 selected PCs by truncated models, respectively. Figure (1b) shows the PVE and power as functions of the number of included PCs. For the first three thresholds, the corresponding power are 0.167, 0.162, and 0.186, respectively. The power increases substantially to 0.609 when the threshold increases to 90%, under which the fourth PC is additionally selected in the model. If the threshold further increases to 95 and 99%, however, the power decreases to 0.582 and 0.540, respectively. This demonstrates that the choice of the PVE threshold is critical in power performance and that a higher PVE threshold does not imply a higher power. On the other hand, it is possible that an alternative procedure achieves higher power than any of the six truncated model under consideration by the classical procedure. In fact, if one includes only the first and the fourth ones, the corresponding power is 0.666,

superior to the highest power 0.609 attained under any PVE thresholds. This example shows that eigenvalues alone do not provide the optimal criterion to rank and select the PCs for the testing purpose.

3. A New Selection Criterion and Testing Procedure

3.1. *The Proposed Procedure*

To incorporate the association between the covariate and the response into the selection criterion, we hereby propose an AVI

$$V_k = \lambda_k \beta_k^2, \qquad k = 1, \ldots, \infty,$$

for each principal component. The AVI is motivated by the non-central parameter shown in (5) and is closely related to the coefficient of determination in the linear model. First, the power of the classical method demonstrated in the previous section depends on the non-centrality parameter η. Since λ_k and β_k affect the value of η only through $\lambda_k \beta_k^2$, it is natural to consider the AVI defined above. Second, it can be shown that the AVI $V_k = \frac{\mathrm{cov}^2(\xi_k, Y)}{\lambda_k}$ is proportional to the coefficient of determination $R_k^2 = \mathrm{corr}^2(\xi_k, Y)$, the proportion of the variation of Y explained by ξ_k. By taking the association β_k into account in the selection criterion, the AVI captures the most important directions for hypothesis testing in FLM, yet keeps the amount of information along each direction in consideration.

However, unlike the selection criterion based on the variation λ_k that can be estimated based on the covariate process $X(\cdot)$ only, V_k involves unknown parameter β_k that also depends on the outcome. To obtain $\hat{\beta}_k$, we propose to pre-fit the truncated model (3) with a high threshold of PVE, say, 95–99%. This step is used to control the total number of PCs to be considered, and a high threshold of PVE provides more comprehensive information for the selection procedure as described below, while still being able to yield stable estimates for β_k's (Web Figure S3). Suppose that there are C_n principal components obtained from the pre-fitted truncated model. The AVIs can be estimated by $\hat{V}_k = \hat{\lambda}_k \hat{\beta}_k^2$.

We propose a Wald-type test statistic similar to (4), but construct it based on the (non-increasing) order statistics $\hat{V}_{(k)}$ of $\{\hat{V}_k, k = 1, \ldots, C_n\}$. Before presenting the test statistic, we introduce a new measure, the *percentage of association–variation explained* (PAVE), to describe how well the truncated model approximates the FLM. The idea of PAVE is comparable to PVE except that it depends on the AVIs rather than λ_k only. For a given positive integer $K \le C_n$, the corresponding PAVE is expressed as

$$\sum_{k=1}^{K} \hat{V}_{(k)} \Big/ \sum_{k=1}^{C_n} \hat{V}_{(k)}.$$

This measure determines how many eigenfunctions should be included. Given a threshold of PAVE, say γ, the test statistic based on the estimated AVI is defined as

$$T = \sum_{k=1}^{K_n} \frac{\hat{\beta}_{(k)}^2}{\hat{\mathrm{var}}\left(\hat{\beta}_{(k)}\right)} = \frac{1}{\hat{\sigma}_\xi^2} \sum_{k=1}^{K_n} \frac{\mathbf{Y}^T \hat{\boldsymbol{\xi}}_{(k)} \hat{\boldsymbol{\xi}}_{(k)}^T \mathbf{Y}}{\hat{\boldsymbol{\xi}}_{(k)}^T \hat{\boldsymbol{\xi}}_{(k)}},$$

where

$$K_n = \underset{K \le C_n}{\mathit{argmin}} \sum_{k=1}^{K} \hat{V}_{(k)} \Big/ \sum_{k=1}^{C_n} \hat{V}_{(k)} \ge \gamma. \tag{6}$$

T is proportional to the sum of the K_n largest values among the C_n estimated AVIs. Intuitively, the proposed testing procedure orders the principal components by the AVI and aggregates information along the directions associated with large AVIs.

3.2. *Asymptotic Distributions under the Null and Alternatives*

The randomness of $\hat{\beta}_k$ in the proposed selection criterion induces complications in deriving the null distribution of T. As a result, the classical χ^2 distribution does not hold. We investigate the asymptotic null distribution of T herein. The required regularity conditions are listed in Assumptions C1–C5 in Web Appendix A. There are two approaches to determining the number of eigenfunctions K_n in the testing procedure. One is to pre-specify a positive integer ($\le C_n$), and the other is to set a tuning parameter for the PAVE as shown in (6). We present the asymptotic null distributions corresponding to both approaches.

THEOREM 1. *Denote C_n the number of selected PCs from a pre-fitted model, and $Z_{(1)}, \ldots, Z_{(C_n)}$ as order statistics (in a decreasing order) of C_n i.i.d. χ_1^2 random variables $Z_1, \ldots, Z_{C_n}$. Assume that regularity conditions C1–C4 and the null hypothesis hold.*

(1) Given $K_n = K$, the distributions of T and $\sum_{k=1}^{K} Z_{(k)}$ are asymptotically equivalent.

(2) Let $\gamma \in (0, 1)$ be a pre-specified value of the PAVE. The distribution of T and that of $\sum_{k=1}^{K^*} Z_{(k)}$ are asymptotically equivalent, where K^*, a random quantity, satisfies $\sum_{k=1}^{K^*-1} Z_{(k)} \Big/ \sum_{k=1}^{C_n} Z_{(k)} < \gamma$ and $\sum_{k=1}^{K^*} Z_{(k)} \Big/ \sum_{k=1}^{C_n} Z_{(k)} \ge \gamma$.

The proof of Theorem 1 is included in Web Appendix B. From Theorem 1, the asymptotic null distribution of T is affected by not only the values but also the order of the estimated AVIs. It is perceivable that the typical usage of a χ_K^2 as the null distribution of the Wald-type T shall inflate the type I error since T has a heavier right tail. Although Theorem 1 explicitly gives the asymptotic null distribution of T, unfortunately there is no closed form since it involves order statistics. However, the quantiles and tail probabilities of the null distribution of T can be obtained by fast Monte Carlo approximations. Remark 1 below shows a special case under which an analytical distribution exists, and the quan-

tiles and tail probabilities can be easily obtained by existing softwares.

REMARK 1. *When K_n is pre-specified as 1, the CDF of T under the null hypothesis can be approximated by $\left[F_{\chi_1^2}(t)\right]^{C_n}$, where $F_{\chi_1^2}$ stands for the CDF of a χ_1^2 random variable.*

To comprehensively understand the performance of the proposed method, we also study the asymptotic distribution under an alternative hypothesis, and investigate the power of the testing procedure. For a given threshold $\gamma \in (0,1)$ of the PAVE, we denote by K_n^0 a positive integer which satisfies

$$\sum_{k=1}^{K_n^0-1} V_{(k)} \Big/ \sum_{k=1}^{C_n} V_k < \gamma \text{ and} \sum_{k=1}^{K_n^0} V_{(k)} \Big/ \sum_{k=1}^{C_n} V_k \geq \gamma.$$

Below we consider an alternative hypothesis $\mathrm{H_a} : \beta(\cdot) \neq 0$, for a certain known $\beta(\cdot) = \sum_{k=1}^{\infty} \beta_k \phi_k(\cdot) \in \mathcal{H}$. The asymptotic distribution of T under the alternative hypothesis, as $n \to \infty$, is presented in the following theorem and a sketch of the proof is provided in Web Appendix C.

THEOREM 2. *Assume that regularity conditions C1–C5 hold. Denote $\{V_{(k)}\}$ the ordered statistics of $V_1, \ldots, V_{C_n}$ and $\eta_n = \frac{n}{\sigma_\zeta^2} \sum_{k=1}^{K_n^0} V_{(k)}$. Under $\mathrm{H_a}$, the proposed test statistic T follows the following asymptotic distribution.*

$$\frac{T - (K_n + \eta_n)}{\sqrt{2(K_n + 2\eta_n)}} \underset{\mathcal{D}}{\to} N(0,1), \text{ as } n \to \infty. \tag{7}$$

Theorem 2 can be used for calculating sample size when a desired level of power p^* is specified. For a given sample size n, significance level α, threshold of PVE for the pre-fit model, and specified alternative $\beta(\cdot)$ with associated V_k, the power can be approximated by $1 - \Phi\left(\frac{q_{(\alpha,K_n,C_n)} - (K_n+\eta_n)}{\sqrt{2(K_n+2\eta_n)}}\right)$, where $q_{(\alpha,K_n,C_n)}$ is the $100(1-\alpha)$-percent quantile of the null distributions shown in Theorem 1. By incorporating the estimate of the covariance matrix from some pilot study and solving $1 - \Phi\left(\frac{q_{(\alpha,K_n,C_n)} - (K_n+\eta_n)}{\sqrt{2(K_n+2\eta_n)}}\right) \geq p^*$, researchers can determine an appropriate sample size for desired power p^*.

3.3. *Other Extensions*

In practice, the functional covariate observed from the ith subject is often recorded intermittently at grid points $\vec{t}_i = (t_{i1}, \ldots, t_{in_i})$, and may also be subject to measurement error. Specifically, the observed functional covariate $\vec{W}_i = (W_{i1}, \ldots, W_{in_i})$ can be expressed as $W_{ij} = X_i(t_{ij}) + e_{ij}$, where $e_{ij} \sim (0, \sigma_e^2)$ is an error term. The measurement error introduces additional variation in the observations and hence results in a covariance matrix $cov(\vec{W}_i)$ with the (j_1, j_2)th-element as $G(t_{ij_1,ij_2}) + \sigma_e^2 I(j_1 = j_2)$. Yao et al. (2005a) extended FPCA to adapt the situations with measurement error. They proposed an approach called PACE (principle component analysis based on conditional expectations) to consistently estimate eigenvalues, eigenfunctions, and the error variance. If one uses the standard approach to estimate the PC scores, $\int \{X(t) - \mu(t)\}\phi_k(t)\,dt$, the estimates will be contaminated by noise, and measurement error correction is needed in the FLM regression. Regression calibration is a popular and effective approach for measurement error models (Carroll et al., 2006). It provides consistent estimators for regression coefficients in linear models. It requires substituting ξ_k by its conditional expectation $\hat{\xi}_{ik} = \hat{E}(\xi_{ik}|\vec{W}_i)$, which is exactly the PACE estimator for the PC scores. Plugging in $\hat{\xi}_k$ to the regression model is in essence equivalent to implementing a regression calibration correction and thus yields consistent estimators for β_k. Thus, the proposed testing procedure is still valid in the presence of measurement error with the PACE estimator for PC scores.

Another extension is to accommodate cases with additional baseline covariates. In the presence of d-dimensional baseline covariates $\mathbf{Z}$, model (1) can be extended to

$$Y = \beta_0 + \mathbf{Z}^T \boldsymbol{\alpha}_1 + \int_{\mathcal{T}} \beta(t) X(t) dt + \epsilon,$$

where $\boldsymbol{\alpha} \in \mathbb{R}^d$ are the regression coefficients for baseline covariates, and the error term ϵ is independent of $\mathbf{Z}$ and $X(\cdot)$. A truncated version analogous to (3) is

$$Y = \beta_0 + \mathbf{Z}^T \boldsymbol{\alpha}_1 + \sum_{k=1}^{K} \beta_k \xi_k + \zeta.$$

The proposed testing procedure can be extended to this setting straightforwardly.

4. Simulation Studies

In this section, we study the finite sample performance of both the classical and proposed testing procedures under the FLM. The cases without and with measurement errors are investigated in the following two sections, respectively.

4.1. *Functional Covariates without Measurement Errors*

For simplicity of presentation, we first consider a scenario where functional covariate $X_i(\cdot)$, $i = 1, \ldots, n$, was generated by three Fourier basis as

$$X_i(t) = \sum_{k=1}^{3} \xi_{ik} \phi_k(t),$$

where $\xi_{ik} \overset{i.i.d}{\sim} N(0, \lambda_k)$ with $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.25$. The three orthonormal Fourier basis functions are, $\phi_1(t) = \sin(2\pi t)/\sqrt{0.5}$, $\phi_2(t) = \cos(4\pi t)/\sqrt{0.5}$, and $\phi_3(t) = \sin(4\pi t)/\sqrt{0.5}$. Besides $X_i(t)$, a binary baseline covariate $Z_i \sim B(1, 0.05)$ was included in the model as a potential confounder. The outcome Y_i was generated based on a model

$$Y_i = -3 + 0.5 Z_i + \int_0^1 \beta(t) X_i(t) dt + \epsilon_i, \tag{8}$$

Table 1

Simulation results of the classical (T^c) and the proposed (T) methods with fixed number of principal components based on 2000 Monte Carlo simulations. The rows with $a = 0$ stand for type I error rates of both methods under different scenarios.

		Setting 1		Setting 2		Setting 3		Setting 4		Setting 5		Setting 6	
a	NPC	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T
0	1	0.050	0.044	0.050	0.044	0.050	0.044	0.050	0.044	0.050	0.044	0.050	0.044
0	2	0.047	0.048	0.047	0.048	0.047	0.048	0.047	0.048	0.047	0.048	0.047	0.048
0	3	0.045	0.049	0.045	0.049	0.045	0.049	0.045	0.049	0.045	0.049	0.045	0.049
0.04	1	0.252	0.252	0.252	0.288	0.254	0.413	0.106	0.164	0.053	0.220	0.050	0.069
0.04	2	0.286	0.276	0.370	0.311	0.522	0.464	0.166	0.180	0.120	0.245	0.048	0.078
0.04	3	0.284	0.291	0.321	0.322	0.454	0.466	0.175	0.180	0.240	0.244	0.076	0.081
0.08	1	0.736	0.736	0.734	0.821	0.730	0.956	0.250	0.491	0.053	0.700	0.052	0.164
0.08	2	0.815	0.810	0.914	0.873	0.990	0.975	0.522	0.568	0.353	0.756	0.049	0.166
0.08	3	0.822	0.829	0.876	0.876	0.974	0.974	0.583	0.589	0.752	0.756	0.160	0.166
0.12	1	0.970	0.978	0.970	0.996	0.966	1.000	0.480	0.826	0.051	0.970	0.053	0.344
0.12	2	0.991	0.990	1.000	0.999	1.000	1.000	0.857	0.896	0.690	0.987	0.048	0.344
0.12	3	0.994	0.994	0.998	0.999	1.000	1.000	0.916	0.915	0.986	0.986	0.336	0.336

with $\epsilon_i \sim N(0, 1)$. The association function $\beta(\cdot)$ is formed by the three same Fourier basis functions for $X_i(t)$, $\beta(t) = a\sum_{k=1}^{3}\beta_k\phi_k(t)$, where $a = 0$, 0.04, 0.08, and 0.12 controls the magnitude of $\beta(\cdot)$ across the support. To explore the performance of the testing methods under different scenarios, we consider the true association parameters $(\beta_1, \beta_2, \beta_3)$ to be $(1, 1, 1)$, $(1, \sqrt{2}, 0)$, $(1, 2, 0)$, $(0.5, 1, 1)$ and $(0, 1, 2)$, for settings 1–5, respectively. Setting 1 corresponds to the scenario that the functional effect is same across three directions, while settings 2 and 3 represent the cases that the direction with the smallest variation has null association. Settings 4 and 5 consider the situations that the directions with smaller variation are strongly associated with the response. These five settings cover a wide spectrum of combinations of variation and association, so that we can compare the performance of both the classical and proposed approaches comprehensively. In addition, we consider an extreme setting 6 with eigenvalues $(\lambda_1, \lambda_2, \lambda_3) = (1, 0.5, 0.01)$, and the magnitudes of association along the three directions were set to be $(\beta_1, \beta_2, \beta_3) = (0, 0, 5)$. Under this setting the third PC explains less than 1% of variation and is the only direction associated with the response.

A total of 2000 datasets each with $n = 1000$ were generated for each setting. We study the power performance under various choices of the number of PCs (NPCs) and the threshold γ of PVE or PAVE. We consider NPCs = 1, 2, and 3 for fixed NPCs, and γ ranging from 0.5 to 0.99 with an increment 0.05 for fixed threshold levels. To our knowledge, the performance of the classical method has not been studied with a wide range of threshold of PVE in the literature. Our simulation provides not only a comparison between the classical and the proposed approaches, but also a thorough investigation on the power of the classical method against the choice of the threshold. The simulation results are shown in Table 1 for fixed NPCs and Table 2 for fixed threshold levels of PVE/PAVE. For concise presentation, we choose to show results corresponding to four selected threshold levels (0.5, 0.7, 0.85, and 0.99) in Table 2, while power curves on the refined grid points of threshold levels are shown in Web Figure S1.

The type I error under all settings are close to the nominal significance level 0.05. The power performance of the classical method depends highly on the NPCs and the choice of γ. For example, under setting 2 with $a = 0.08$ the power increases from 0.734 to 0.914, then drops to 0.876 with increasing γ. A higher threshold of PVE does not guarantee a better power. This is a result from a trade-off between the information included in the selected PCs and the degrees of freedom. In contrast, the power of the proposed method is more stable with respect to the NPCs and the threshold choice across different scenarios, compared to the classical method. In addition, we have the following observations: (1) the proposed method has comparable power to the classical method when the PCs with large variation have stronger association although somewhat power loss is observed in some cases. For example, when two PCs are specified to be selected under setting 2 or 3, the power of the proposed method is slightly lower than the classical one (0.311 vs. 0.370, and 0.464 vs. 0.522, respectively). This phenomenon is not surprising since these situations favor the traditional method as the first two PCs are the optimal PCs to be included in the testing procedure, and the power loss of the proposed one is caused by accounting for the extra randomness in the selecting procedure; (2) when the PCs with smaller variation dominate the association function, such as settings 5 and 6, the proposed method demonstrates substantial power advantage, as it reaches a high power even with a low NPC or a low threshold of PAVE (0.7), while the classical method has much lower power with similar threshold values, even at a threshold value as high as 0.85. In an extreme setting like setting 6, the classical method cannot detect the association effectively even with a PVE threshold of 0.99, while the proposed method performs well in terms of power.

We also conducted simulation studies under more complex scenarios, where there are more PCs and the shape of $\beta(t)$ is more complex. As they demonstrate similar patterns to those

Table 2

Type I error and power of the classical (T^c) and proposed (T) methods, based on 2000 Monte Carlo simulations. The number of PCs were selected based on various threshold choices, γ, for PVE in the former and PAVE in the latter. The rows with $a = 0$ correspond to Type I error rates, while those with $a > 0$ correspond to power under alternative hypotheses.

		Setting 1		Setting 2		Setting 3		Setting 4		Setting 5		Setting 6	
a	γ	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T
0	0.5	0.050	0.043	0.050	0.043	0.050	0.043	0.050	0.043	0.050	0.043	0.045	0.050
0	0.7	0.047	0.048	0.047	0.048	0.047	0.048	0.047	0.048	0.047	0.048	0.046	0.049
0	0.85	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.046	0.053
0	0.99	0.045	0.049	0.045	0.049	0.045	0.049	0.045	0.049	0.045	0.049	0.046	0.055
0.04	0.5	0.252	0.266	0.252	0.296	0.254	0.432	0.106	0.172	0.053	0.227	0.046	0.102
0.04	0.7	0.286	0.276	0.370	0.306	0.522	0.456	0.166	0.175	0.120	0.244	0.046	0.106
0.04	0.85	0.286	0.284	0.370	0.318	0.522	0.460	0.166	0.180	0.120	0.244	0.046	0.100
0.04	0.99	0.284	0.291	0.321	0.321	0.454	0.465	0.175	0.179	0.240	0.244	0.046	0.100
0.08	0.5	0.736	0.771	0.734	0.835	0.730	0.962	0.250	0.532	0.053	0.712	0.045	0.310
0.08	0.7	0.815	0.810	0.914	0.869	0.990	0.971	0.522	0.566	0.353	0.752	0.045	0.297
0.08	0.85	0.815	0.824	0.914	0.878	0.990	0.973	0.522	0.582	0.353	0.758	0.045	0.298
0.08	0.99	0.822	0.829	0.876	0.876	0.974	0.974	0.583	0.589	0.752	0.755	0.045	0.298
0.12	0.5	0.970	0.987	0.970	0.996	0.966	1.000	0.480	0.868	0.051	0.969	0.044	0.634
0.12	0.7	0.991	0.992	1.000	0.998	1.000	1.000	0.857	0.906	0.690	0.984	0.044	0.605
0.12	0.85	0.991	0.993	1.000	0.999	1.000	1.000	0.857	0.912	0.690	0.986	0.044	0.610
0.12	0.99	0.994	0.994	0.998	0.999	1.000	1.000	0.916	0.915	0.986	0.986	0.044	0.620

discussed above, we reported these results in Web Tables S1 and S2 in the supplementary materials.

4.2. *Functional Covariates with Measurement Error*

The observed functional covariate from the ith subject at time t_{ij} was generated by

$$W_i(t_{ij}) = \sum_{k=1}^{3} \xi_{ik}\phi_k(t_{ij}) + e_{ij},$$

where $e_{ij} \sim N(0, \sigma_e^2)$, with $\sigma_e^2 = 0.1$ and 1, corresponding to low and high noise levels. Here, a dense grid of 100 equally spaced points between $[0, 1]$ is considered. The response was generated by (8). The simulation results are presented in Table 3 for two measurement error settings, $\sigma_e^2 = 0.1$ and $\sigma_e^2 = 1$, respectively. The thresholds of PVE for the pre-fitted model of the proposed approach are chosen to be 99%.

The type I error rates of both methods are controlled under 0.05. The choice of threshold of PVE has a large impact on the power performance of the classical method especially under settings 3 and 5. In contrast, the proposed method is relatively robust against the choice of threshold values and tends to include less PCs to achieve comparable power than the classical method (average number of PC included are shown in Web Table S2). Under these scenarios, all threshold values above 0.7 perform well in terms of statistical power. This is appealing in practice in two ways. First, it reduces the dimension of predictors because a lower threshold is sufficient to achieve good power. Second, the choice of the threshold does not affect the analysis result greatly, whereas using different thresholds of PVE for the classical method could lead to different conclusions.

5. Data Example: A Diffusion Tensor Imaging Study

We consider a diffusion tensor imaging (DTI) study conducted on 100 multiple sclerosis (MS) patients in the Johns Hopkins Hospital with multiple clinical visits. This cerebral dataset has been considered in Goldsmith et al. (2011) and available in R package "refund." The scientific question of interest is testing the association between the diffusivity along white matter tracts and the cognitive disability in multiple sclerosis patients. The diffusion of water molecules at each voxel is measured by fractional anisotropy (FA) obtained by magnetic resonance imaging technique. FA profiles along two well-defined white matter tracts, corpus callosum (CC) and right corticospinal tracts (RCST), are considered in the analysis. There are 93 and 55 locations along CC and RCST, respectively. Since the value of FA varies at different location along white matter tracts, it can be treated as a function of location on the white matter tracts. To quantify the cognitive impairment, each patient received a paced auditory serial addition test (PASAT) at every clinical visit, and obtained a score ranging from 0 to 60. The goal in our analysis is to quantify the impact of the FA profile along CC and RCST on the PASAT score, and test for the significance of the associations.

To explore the relationship between FA profiles and PASAT scores, we group the MS patients into four groups according to their PASAT scores separated by the three quartiles (25, 50, and 75%). The estimated mean FA profiles in the four groups along the two white matter tracts are presented in Figure (2a) and b. Figure (2a) shows that the mean FA profile along CC in the group with PASAT scores below 25%-quartile is lower than the mean FA profiles from the other three groups in general, while the other three mean curves tangle with each other except at the right tail. On the other hand, the

Table 3

Type I error and power of the classical (T^c) and proposed (T) methods when the functional covariate is recorded with small ($\sigma_e^2 = 0.1$) and moderate measurement error ($\sigma_e^2 = 1$), based on 2000 Monte Carlo simulations. The number of PCs were selected based on various threshold choices, γ, for PVE in the former and PAVE in the latter. The rows with $a = 0$ correspond to Type I error rates, while those with $a > 0$ correspond to power under alternative hypotheses.

		$\sigma_e^2 = 0.1$						$\sigma_e^2 = 1$					
		Setting 1		Setting 3		Setting 5		Setting 1		Setting 3		Setting 5	
a	γ	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T	T^c	T
0	0.5	0.049	0.043	0.049	0.043	0.049	0.043	0.047	0.045	0.047	0.045	0.047	0.045
0	0.7	0.047	0.044	0.047	0.044	0.047	0.044	0.046	0.045	0.046	0.046	0.046	0.046
0	0.85	0.046	0.046	0.046	0.046	0.046	0.046	0.045	0.048	0.045	0.048	0.045	0.048
0	0.99	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.047	0.046	0.047	0.046	0.047
0.04	0.5	0.252	0.268	0.252	0.429	0.053	0.224	0.251	0.261	0.254	0.420	0.052	0.222
0.04	0.7	0.285	0.278	0.522	0.452	0.124	0.234	0.282	0.276	0.512	0.449	0.118	0.227
0.04	0.85	0.285	0.280	0.514	0.458	0.139	0.246	0.281	0.279	0.504	0.450	0.137	0.240
0.04	0.99	0.282	0.285	0.457	0.458	0.236	0.238	0.283	0.284	0.452	0.454	0.236	0.237
0.08	0.5	0.732	0.769	0.722	0.964	0.055	0.710	0.724	0.767	0.711	0.958	0.053	0.686
0.08	0.7	0.816	0.806	0.988	0.972	0.354	0.750	0.812	0.804	0.986	0.970	0.345	0.739
0.08	0.85	0.818	0.822	0.987	0.972	0.404	0.751	0.814	0.818	0.984	0.971	0.400	0.742
0.08	0.99	0.820	0.820	0.974	0.975	0.752	0.754	0.815	0.815	0.972	0.972	0.745	0.746

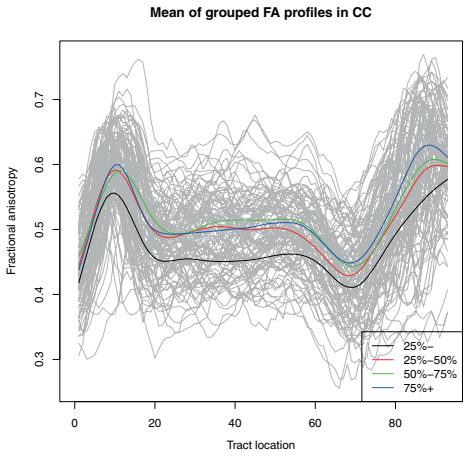
four mean FA profiles along RCST in the four groups stick together except at the peaks and valleys. It is difficult to determine whether the FA profiles have a significant impact on the PASAT scores visually; hence a functional linear regression model is employed with PASAT score as a scalar response and FA profiles as functional covariates.

We applied both the classical and the proposed testing method, and the results are shown in Table 4. We considered three thresholds of PVE/PAVE: 70, 85, and 99%. The PVE threshold for the pre-fitted model in the proposed method is set as 95%. From the results of the classical method, we see that there are inconsistent conclusions with different choices of threshold of PVE. If we use alpha = 0.05 as the significance level, we would conclude that there is a significant association between the FA profiles along CC and the PASAT score if the threshold is 70 or 85%, but the association is not significant if the threshold is 99%. Similarly for the FA profiles along RCST, there is no significant association if the threshold is 70 or 85%, but there is a significant association if the threshold is 99%. For both associations, different thresholds can lead to opposite conclusions. The proposed method overcomes this issue. For CC area, the p-values with 70, 85, and 99% are all smaller than 0.05, which give a consistent conclusion that the FA profile along CC is significantly associated with the PASAT score. Similarly, in the analysis with RSCT, the p-values are all under 0.05 with different choices of the threshold. This suggests a significant impact of the FA profile along RSCT on the PASAT score.

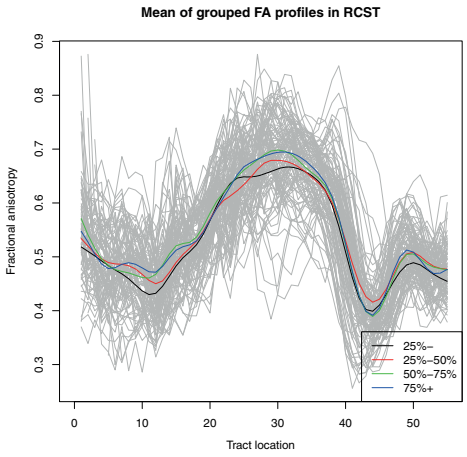
To better explain the inconsistency of results from the PVE based tests, we conduct a simulation study mimicking the DTI data. Specifically, we first estimated eigenvalues and eigenfunctions of the functional covariates as well as the functional regression along both CC and RCST. We then conducted a simulation study, setting the true parameters to be those estimated from the DTI application. The estimated power curves corresponding to a wide range of threshold levels are shown in Figure 2e and f, respectively. For CC area, the power of the classical method decreases with the PVE threshold especially in the range from 0.7 to 0.99. The power loss corresponding to increasing γ explains why the p-value changes from 0.028 ($\gamma = 0.85$) to 0.131 ($\gamma = 0.99$). On the other hand, the power of the proposed method is relatively stable, which explains the consistency in p-values across different threshold choices. For the RCST, the power of the classical method increases substantially with increasing threshold levels, which is why statistical significance is achieved only for $\gamma = 0.99$. Again, the power of the proposed method is relatively robust against the threshold values, leading to the same conclusion in terms of statistical significance across different threshold choices.

6. Conclusions

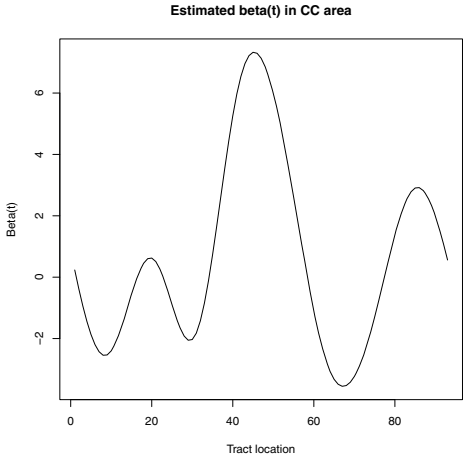
FLM has been a popular tool to describe the dynamic impact of a functional covariate on a scalar response. In the presence of infinite dimensional covariates and regression parameters, applying FPCA can reduce the dimensionality and facilitate statistical inferences. In this work, we investigate the performance of a classical Wald-type test based on the PCs chosen by eigenvalues, and propose a novel association–variance-based selection procedure. The number of PCs can be either pre-specified or determined by the proposed PAVE threshold. We establish the null distributions for both selection criteria and study their asymptotic power. Our numerical studies show that the power performance of the classical approach is sensitive to choice of the number of PCs and including more does not guarantee a higher power. This is rather unappealing in practice because inconsistent conclusions can be drawn based on different thresholds. On the other hand, the proposed PC selection procedure is robust against thresh-



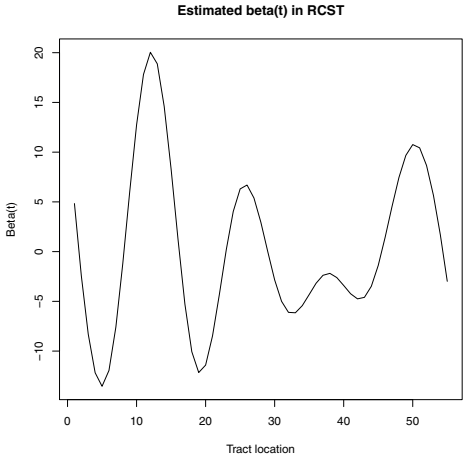
(a) FA profiles along corpus callosum tracts



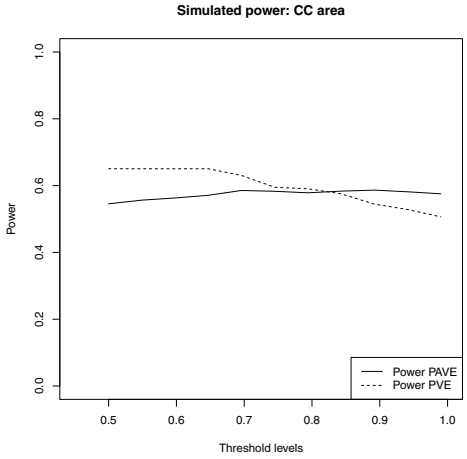
(b) FA profiles along right corticospinal tracts



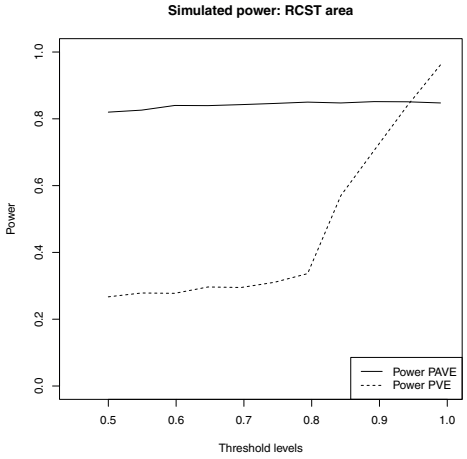
(c) Estimated $\beta(t)$ along corpus callosum tracts



(d) Estimated $\beta(t)$ along right corticospinal tracts



(e) Power curves of PAVE and PVE from data-mimicking simulations on corpus callosum tracts



(f) Power curves of PAVE and PVE from data-mimicking simulations on right corticospinal tracts

Figure 2. Estimated mean FA profiles in the four groups formed by PASAT scores and the corresponding quartiles, estimated $\beta(t)$, and simulated power curves. Left penal: corpus callosum tracts. Right panel: right corticospinal tracts.

Table 4
Testing results of the classical (T^c) and proposed (T) methods on the DTI data with FA profiles in corpus callosum and right corticospinal tracts as functional covariates. The number of PCs were selected based on various threshold choices, γ, for PVE in the former and PAVE in the latter.

	Corpus callosum tracts				Right corticospinal tracts			
	p-value		Number of PCs		p-value		Number of PCs	
γ	T^c	T	T^c	T	T^c	T	T^c	T
0.70	0.007	0.039	2	2	0.111	0.021	3	2
0.85	0.028	0.044	4	3	0.064	0.015	5	3
0.99	0.131	0.042	9	6	0.004	0.028	9	6

old choices. Moreover, it is more powerful than the classical method when the leading PCs are weakly or not associated with the response.

Practitioners might be interested in how to choose an optimal threshold for the proposed PAVE approach. As we illustrated, the choice of threshold is less crucial for our method, due to the robustness of its power across a wide range of threshold levels. This is demonstrated by simulation studies across a wide range of scenarios, including the simple settings in Section 4.1, more complex settings in Web Appendix E, and settings that mimic the DTI data application (power curves in Figure 2e and f). In these scenarios, generally the choice of threshold does not lead to substantial differences in statistical power and changes of conclusions, which makes our method more desirable than the classical approach. Nevertheless, it is always desirable to provide some guidance for practitioners. Based on our empirical experiences, we recommend a threshold level of around 0.8–0.9, which works well across the scenarios that we explored. In particular, simulations mimicking the DTI data suggest that our proposed PAVE-based test with a threshold level of 0.8–0.9 achieves great power (Figure 2e and f) with parsimonious fitting. In case that practitioners want to identify the optimal threshold for highest power, we suggest conducting a small simulation study mimicking their data structure, especially if they have prior knowledge of plausible shapes of the association function $\beta(t)$.

There are several directions for future research. First, we consider functional linear models for continuous outcomes in this article. It will be of interest to extend the method to generalized functional linear models for non-Gaussian outcomes, or time-to-event outcomes. Second, extensions to multilevel functional data will also be useful. Many studies, including the DTI application here, recorded functional data at multiple visits for the same subject, resulting in multilevel functional data. Crainiceanu et al. (2009) and Di et al. (2009) proposed multilevel functional principal component analysis to extract modes of variations at both between and within subject levels for such data. Finally, even though we briefly discussed extending our method to functional linear regression with sparse longitudinal data, the asymptotic theory remains challenging, especially how the smoothing parameter in the FPCA step will affect inference. This will be investigated in a separate article in the future.

7. Supplementary materials

Appendix, Web Figures, and Web Tables referenced in Sections 3 and 4.1 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work was partially supported by the National Institutes of Health grants R01 HL130483, P01 CA53996, R01 CA189532, R01 CA195789, R01 HG006124 and R01 AI121259.

References

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics* **30**, 241–255.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11–22.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: Chapman and Hall/CRC.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association* **104**, 1550–1561.

Di, C.-Z., Crainiceanu, C. M., Caffo, B., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* **3**, 458–488.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penailzed functional regression. *Journal of Computational and Graphical Statistics* **20**, 830–851.

Goldsmith, J., Greven, S., and Crainiceanu, C. M. (2013). Corrected confidence bands for functional data using prinicipal components. *Biometrics* **69**, 41–51.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of Royal Statistical Society Series B* **68**, 109–126.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer-Verlag New York: Springer.

James, G. (2002). Generalized linear models wtih functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.

Kong, D., Staicu, A.-M., and Maity, A. (2013). Classical testing in functional linear models. Technical report, North Carolina State University Department of Statistics Technical Reports, 2647, 1–23.

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

Ramsay, J. and Silverman, B. W. (2005). *Functional data analysis*. Springer-Verlag New York: Springer.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B* **53**, 539–572.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer-Verlag New York: Springer.

Swihart, B. J., Goldsmith, J., and Crainiceanu, C. M. (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics* **56**, 483–493.

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

Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33**, 2873–2903.

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

Received December 2015. *Revised August* 2016.
Accepted October 2016.