To appear in the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–25

Classical Testing in Functional Linear Models

Dehan Kong[a*] Ana-Maria Staicu[b] and Arnab Maity[b]

[a] *Department of Biostatistics, University of North Carolina, Chapel Hill, NC, 27599, U.S.A. ;*
[b] *Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A. ;*

(*v4.0 released June 2015*)

We extend four tests common in classical regression - Wald, score, likelihood ratio and F tests - to functional linear regression, for testing the null hypothesis, that there is no association between a scalar response and a functional covariate. Using functional principal component analysis, we re-express the functional linear model as a standard linear model, where the effect of the functional covariate can be approximated by a finite linear combination of the functional principal component scores. In this setting, we consider application of the four traditional tests. The proposed testing procedures are investigated theoretically for densely observed functional covariates when the number of principal components diverges. Using the theoretical distribution of the tests under the alternative hypothesis, we develop a procedure for sample size calculation in the context of functional linear regression. The four tests are further compared numerically for both densely and sparsely observed noisy functional data in simulation experiments and using two real data applications.

Keywords: Asymptotic distribution, Functional principal component analysis, Functional linear model, Hypothesis Testing

AMS Subject Classification: 62G08; 62G10

*Corresponding author. Email: kongdehanstat@gmail.com

1. Introduction

Functional regression models have become increasingly popular in the field of functional data analysis, with applications in various areas such as biomedical studies, brain imaging, genomics and chemometrics, among many others. We consider the functional linear model (Ramsay and Dalzell 1991) where the response of interest is scalar and the covariate of interest is functional, and the primary goal is to investigate their relationship. In this article, our main focus is to develop hypothesis testing procedures to test for association between the functional covariate and the scalar response when the functional covariate is observed on a dense grid and corrupted with measurement error. We discuss four testing procedures and investigate the theoretical properties for our recommended tests. The approaches are then extended in two directions: 1) first to the case of noisy and sparsely observed covariates, and 2) second to the partial functional linear model (Shin 2009), which accounts for additional covariates using a linear relationship. The finite sample performance for different realistic scenarios is evaluated numerically via a simulation study. The testing procedures are then applied to two data sets: a Diffusion Tensor Imaging tractography data set, portraying a densely and regularly observed functional covariate situation with missingness; and an auction data on eBay of the $Microsoft\ Xbox$ gaming systems, portraying a sparsely observed functional covariate setting.

In functional linear models, the effect of the functional predictor on the scalar response is represented by an inner product of the functional predictor and an unknown, nonparametrically modeled, coefficient function. Typically, such coefficient function is assumed to belong to an infinite dimensional Hilbert space. To estimate the coefficient function, one often projects the functional predictor and the coefficient function onto certain basis systems, such as eigenbasis, or pre-determined basis systems such as spline basis or wavelet basis system to achieve dimension reduction. There is a plethora of literature on estimation of the coefficient function; see for example, Cardot, Ferraty, and Sarda (1999), Yao, Müller, and Wang (2005b). For a detailed review of functional linear model, we refer the readers to Ramsay and Silverman (2005) and the references therein.

Our primarily interest in this article is the problem of testing whether the functional covariate is associated with the scalar response, or equivalently, whether the coefficient function is zero. The problem of testing in the context of functional linear models is important for two main reasons. First, in many real life situations, especially in biomedical studies, evidence for association between a predictor and a response is as valuable as, if not more than, estimation of the actual effect size. In the case when the predictors are functional, estimates of the actual coefficient curves are often hard to interpret and it may not be clear whether the covariate is in fact useful to predict the outcome. Secondly, the tactic of constructing a pre-specified level confidence interval around the estimate and then inverting the interval to construct a test, as is usually done in multivariate situation, is not readily applicable in the functional covariate case. Most of the available literature on functional linear models present point-wise confidence bands of the estimated coefficient functions rather than a simultaneous one. Inverting such a point-wise confidence band to construct a test holds very little meaning. Thus testing for association remains a problem of paramount interest. Unfortunately, the literature in the area of testing for association is relatively sparse and often makes assumptions that are quite strong and impractical.

Cardot, Ferraty, Mas, and Sarda (2003) discussed a testing procedure based on the norm of the cross covariance operator of the functional predictor and the scalar response. Later, Cardot, Goia, and Sarda (2004) proposed two computational approaches by using

a permutation and F tests. Hilgert, Mas, and Verzelen (2013) introduced two minimax adaptive procedures to test the nullity of the slope function in the functional linear model. These two methods built strong theoretical support for their test statistics and have good performance numerically. González-Manteiga, González-Rodríguez, Martínez-Calvo, and García-Portuguéss (2014) proposed a bootstrap independence test to achieve the same goal. A key assumption of these approaches is that the functional covariates are observed on dense regular grids, without measurement error. This assumption is not realistic in many practical situations; for example, in both applications considered, the covariates are observed on irregular grids or with measurement error. Müller and Stadtmüller (2005) proposed the generalized functional linear model and studied the analytical expression of the asymptotic global confidence bands of the coefficient function estimator. A Wald test statistic can be derived from the asymptotic properties of this estimator. However, a crucial assumption in that work is also that the functional covariate is observed fully and without error. Additionally, as we observe in our simulation studies, the Wald test statistic is not very reliable for small sample sizes and exhibits significantly inflated Type I error rate even when the functional covariate is observed on very fine grids and without error. Swihart, Goldsmith, and Crainiceanu (2014) addressed a similar testing problem when the setting involves multiple functional covariates; they discussed the restricted likelihood ratio test and investigated its performance numerically, via simulation studies, but did not present its theoretical properties.

In this paper, we consider the situation where the functional predictor is observed at irregular sets of points and is possibly corrupted with measurement error. We investigate four traditional test statistics, namely, score, Wald, likelihood ratio and F test statistics. To facilitate these testing procedures, we mainly rely on the use of the eigenbasis functions, derived from the functional principal component analysis of the observed functional covariates, to model the coefficient function. This method, commonly known as functional principal component regression has been well researched in literature; see for example Müller and Stadtmüller (2005), and Hall and Horowitz (2007).

We use functional principal component analysis and model the coefficient function using the eigen functions derived from the Karhunen-Loève expansion of the covariance function of the predictor. As a result, we re-express the functional linear model as a multiple regression model, where the effect of the functional covariate can be approximated as a linear combination of the functional principal component scores. Traditional tests such as Wald, score, likelihood ratio and F tests are then formulated using the unknown coefficients in the re-written model. Using functional principal component analysis to model the coefficient function has various advantages. First, one can accommodate sparsely observed functional covariates at the subject level, where smoothing of individual curves is practically impossible. In addition, theoretical properties of the functional principal component scores have been studied in a variety of settings: see for example Hall and Hosseini-Nasab (2006), Hall, Müller, and Wang (2006) and Yao, Müller, and Wang (2005a). Finally, functional principal component analysis provides automatic choices of data adaptive, empirical, basis functions, and as such one can readily choose the number of basis functions to be used in the model by looking at the percent of variance explained by the corresponding number of principal components.

This article makes two major contributions. First, we derive the theoretical properties for our recommended tests, namely F test and score test. In particular, we derive the null distributions and asymptotic theoretical alternative distributions, for dense and noisy observations of the functional covariate. Furthermore we develop the asymptotical rate of our tests: the testing procedures are shown to be asymptotically near optimal. Second, as a consequence of our theoretical results, we develop a procedure for sample

size calculation in the context of functional linear regression. To the best of our knowledge, this is the first such result in the existing literature. Such sample size calculation procedures are immensely useful when one has a fair idea of what the underlying covariance structure of the functional covariates is, from a pilot or preliminary study, and is interested in determining the sample size of a future larger study within the same cohort.

Our theoretical results are asymptotic, in the sense that they are derived assuming that the sample size is diverging to infinity. While such results are of great interest, it is also important to observe the performance of the testing procedures in finite sample sizes. We investigate numerically the performance of the four tests, when the functional covariate is observed either at regular, dense designs as well as sparse, irregularly spaced designs. The results show that, while all the four test statistics behave very similarly in terms of both Type I error rate and power, for very large sample size, they show different behavior for small and moderate sample sizes. In particular, for small and moderate sample sizes, the likelihood ratio and the Wald tests exhibit significantly inflated Type I error rate in all the designs, while the score test shows a conservative Type I error. On the other hand the F test retains close to nominal Type I error rates and provides larger power than the score test; thus F test may be viewed as a robust testing procedure, even for small sample sizes and sparse irregular designs.

The rest of this article is organized as follows. Section 2 describes the proposed methodology including the model setup and testing procedures. Asymptotic properties of our method are studied in Section 3. Sections 4 and 5 discuss the extension to the sparsely and noisy observed functional data and the partially functional linear model. The testing procedures are applied to two real data sets in Section 6, and evaluated numerically in Section 7.

2. Methodology

2.1. *Model specification*

Suppose for $i = 1, \ldots, n$, we observe a scalar response Y_i and covariates $\{W_{i1}, \ldots, W_{im_i}\}$ corresponding to points $\{t_{i1}, \ldots, t_{im_i}\}$ in a closed interval $\mathcal{T}$. Assume that W_{ij} is a proxy observation of the true underlying process $X_i(\cdot)$, such that $W_{ij} = X_i(t_{ij}) + e_{ij}$, where $\eta(\cdot)$ is the mean function, and e_{ij}'s are independent and identically distributed Gaussian variables with zero mean variance σ_e^2. Furthermore, it is assumed that the true process $X_i(\cdot) \in L^2(\mathcal{T})$ has zero mean, for simplicity, and covariance kernel $K(\cdot, \cdot)$. We also assume that the true relationship between the response and the functional covariate is given by a functional linear model (Ramsay and Silverman 2005)

$$Y_i \quad = \quad \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i, \tag{1}$$

where ϵ_i are independently and identically distributed normal random variable with mean 0 and variance σ^2, α is an unknown intercept and $\beta(\cdot)$ is an unknown coefficient function quantifying the effect of the functional predictor across the domain $\mathcal{T}$ and represents the main focus of our paper. Recently McLean, Hooker, and Ruppert (2014) proposed a restricted likelihood ratio test for testing for linear dependence between a scalar response and a functional covariate, in the class of functional generalized additive models (Mclean, Hooker, Staicu, Scheipl, and Ruppert 2014; Müller, Wu, and Yao 2013). In what follows, we write $\int X_i(t)\beta(t)dt$ instead of $\int_{\mathcal{T}} X_i(t)\beta(t)dt$ for notational convenience.

Our goal is to test the null hypothesis that there is no relationship between the covariate $X(\cdot)$ and the response Y. Formally, the null and the alternative hypotheses can be stated as

$$H_0 : \beta(t) = 0 \text{ for any } t \in \mathcal{T} \text{ vs } H_a : \beta(t) \neq 0 \text{ for some } t \in \mathcal{T}. \tag{2}$$

To the best of our knowledge most of the existing methods, for example Müller and Stadtmüller (2005), Cardot et al. (2003) and Cardot et al. (2004), assume that the functional covariates are observed fully and without noise. In this paper, we consider the case where the functional covariate may be observed densely with measurement error. We develop four testing procedures to test H_0, study their theoretical properties, and compare their performances numerically.

2.2. *Testing procedure*

The idea behind developing the testing procedures is to use an orthogonal basis function expansion for both $X(\cdot)$ and $\beta(\cdot)$ and then reduce the infinite dimensional hypothesis testing to the testing for the finite number of parameters by using an appropriate finite truncation of this basis. In this paper, we consider the eigenbasis functions obtained from the covariance operator of $X(\cdot)$. Specifically, let the spectral decomposition of the covariance function $K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$, where $\{\lambda_j, j \geq 1\}$ are the eigenvalues in strictly decreasing order with $\sum_{j=1}^{\infty} \lambda_j < \infty$ and $\{\phi_j(\cdot), j \geq 1)$ are the corresponding eigenfunctions. Then $X_i(\cdot)$ can be represented using Karhunen-Loève expansion as $X_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t)$, where the functional principal component scores are $\xi_{ij} = \int X_i(t)\phi_j(t)dt$, have mean zero, variance λ_j, and are uncorrelated over j. Using the eigenfunctions ϕ_j, the coefficient function $\beta(t)$ can be expanded as $\beta(t) = \sum_{j=1}^{\infty} \beta_j \phi_j(t)$, where β_j's denote the unknown basis coefficients. Thus the functional regression model (1) can be equivalently written as $Y_i = \alpha + \sum_{j=1}^{\infty} \xi_{ij}\beta_j + \epsilon_i$, for $1 \leq i \leq n$, and testing (2) is equivalent to testing $\beta_j = 0$ for all $j \geq 1$.

However, such a model is impractical as it involves an infinite sum. Instead, we approximate the model with a series of models where the number of predictors $\{\xi_{ij}\}_{j=1}^{\infty}$ is truncated to a finite number s_n, which increases with the number of subjects n. Conditional on the truncation point s_n, the model can be approximated by

$$Y_i \quad = \quad \alpha + \textstyle\sum_{j=1}^{s_n} \xi_{ij}\beta_j + \epsilon_i, \tag{3}$$

and the hypothesis testing problem can be reduced to

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{s_n} = 0 \text{ vs } H_a : \beta_j \neq 0 \text{ for at least one } j, 1 \leq j \leq s_n. \tag{4}$$

Our model specification allows the coefficient function $\beta(\cdot)$ to be identifiable only within the eigenspace of X's; nevertheless for testing purposes, it is only required that $\beta(\cdot)$ does not lie in the orthogonal complement of this space. The truncation value s_n is selected with the intention of recovering the full space of X's. A different truncation level s_n from the optimal one does not affect the performance of the Type I error rates of the proposed testing procedures. However, selecting an unnecessarily large number of components may result in a loss of power of the testing procedure. In our numerical investigation, we estimated s_n by the percentage of explained variance; for example our simulations use 95 percent explained variation and show that the tests have very good

size and power performance.

We consider four classical testing procedures, namely Wald, Score, likelihood ratio and F test and examine their application in the context of (3). Define $Y = (Y_1, \ldots, Y_n)^\top$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$. With a slight abuse of notation, define $\beta = (\beta_1, \ldots, \beta_{s_n})^\top$ and $\theta = (\sigma^2, \alpha, \beta^\top)^\top$. Given the truncation s_n and the true scores $\{\xi_{ij}, 1 \le i \le n, 1 \le j \le s_n\}$, the log likelihood function from (3) can be written as

$$L_n(\theta) \;=\; -(n/2)\log(2\pi\sigma^2) - (Y - \alpha 1_n - M\beta)^\top (Y - \alpha 1_n - M\beta)/(2\sigma^2), \qquad (5)$$

where 1_n is a vector of ones of length n, and M is $n \times s_n$ matrix with the (i,j)-th element being $M_{ij} = \xi_{ij}$. We use the likelihood function (5) to develop the tests for testing $H_0 : \beta = 0$.

Let $B = [1_n, M]$, and define the projection matrices $P_1 = 1_n 1_n^\top/n$ and $P_B = B(B^\top B)^{-1}B^\top$. The score function corresponding to (5) is $S_n(\theta) = \partial L_n(\theta)/\partial\theta$ and equals

$$S_n(\theta) = \{-n/2\sigma^2 + (Y - \alpha 1_n - M\beta)^\top (Y - \alpha 1_n - M\beta)/2\sigma^4, (Y - \alpha 1_n - M\beta)^\top B/2\sigma^2\}^\top;$$

the corresponding information matrix $\mathcal{I}_n(\theta)$ is a block-diagonal matrix with two blocks, where the first block is the scalar $\mathcal{I}_{11} = 2n/\sigma^4$ and the second block is the matrix $\mathcal{I}_{22} = B^\top B/\sigma^2$. Define $I_{n\times n}$ as the $n \times n$ identity matrix and let $\widetilde{\theta} = (\widetilde{\sigma}^2, \widetilde{\alpha}, 0_{s_n}^\top)^\top$, where $\widetilde{\sigma}^2 = Y^\top (I_{n\times n} - 1_n 1_n^\top/n)Y/n$ and $\widetilde{\alpha} = \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ are the constrained maximum likelihood estimators for σ^2 and α, respectively, under the null hypothesis. The efficient score test (Rao 1948) is then

$$T_S = S_n(\widetilde{\theta})^\top \{\mathcal{I}_n(\widetilde{\theta})\}^{-1} S_n(\widetilde{\theta}) = Y^\top (P_B - P_1)Y/\widetilde{\sigma}^2.$$

The advantage of the score test is that this statistic only depends on the estimated parameters under the model specified by the null hypothesis, and thus it requires fitting only the null model. In contrast to the score test, the advantage of the Wald test is that it only requires to fit the full model. In particular, let $\widehat{\theta} = (\widehat{\sigma}^2, \widehat{\alpha}, \widehat{\beta}^\top)^\top$ denote the maximum likelihood estimate of θ under the full model. Define $V(\widehat{\beta})$ to be the variance-covariance matrix of $\widehat{\beta}$ evaluated at $\widehat{\theta}$, that is, the $s_n \times s_n$ submatrix of $I_n^{-1}(\widehat{\theta})$ corresponding to β. The Wald test statistic is then defined as

$$T_W = \widehat{\beta}^\top \{V(\widehat{\beta})\}^{-1} \widehat{\beta}.$$

In this work, we consider a slightly modified version of this statistic, where $\widehat{\sigma}^2$ is replaced by the restricted maximum likelihood estimate $\widehat{\sigma}^2_{REML} = Y^\top (I_{n\times n} - P_B)Y/(n - s_n - 1)$, rather than the usually used maximum likelihood estimate. In our simulation study, we found that Wald test with the restricted maximum likelihood estimate for σ^2 yields considerably improved results in terms of Type I error when the sample size is small; even with this adjustment the Type I error is significantly inflated. For large sample sizes, the performance of the Wald test is similar for the two types of estimates for σ^2.

Next we consider the likelihood ratio test statistic. Usually, this statistic is defined as $-2\{L_n(\widetilde{\eta}, \widetilde{\sigma}^2) - L_n(\widehat{\eta}, \widehat{\sigma}^2)\}$ which simplifies to $n\log(\widetilde{\sigma}^2/\widehat{\sigma}^2)$. This test is similar to the 'restricted' likelihood ratio test when there is a single functional covariate, discussed in Swihart et al. (2014); in this scenario, their proposed restricted likelihood ratio test becomes a likelihood ratio test. Using the same argument as in Wald test, in this case also, we consider the restricted maximum likelihood estimate for σ^2 for both the null and

the full model, and define a slightly modified likelihood ratio statistic

$$T_L = s_n + n \log(\widetilde{\sigma}^2_{REML}/\widehat{\sigma}^2_{REML}),$$

where $\widetilde{\sigma}^2_{REML} = Y^\top(I_{n\times n} - P_1)Y/(n-1)$ is the restricted maximum likelihood estimate for σ^2 under the null model. Notice that one needs to fit both the full and the null model to compute this test statistic.

Finally, we define the F test in terms of the residual sum of squares under the full and the null models. In particular, define $RSS_{\text{full}} = Y^\top(I_{n\times n} - P_B)Y$, and $RSS_{\text{red}} = Y^\top(I_{n\times n} - P_1)Y$. to be the residual sum of squares under the full and the null models, respectively. The F test statistic is then defined as

$$T_F = \frac{(RSS_{\text{red}} - RSS_{\text{full}})/s_n}{RSS_{\text{full}}/(n - s_n - 1)} = \frac{Y^\top(P_1 - P_B)Y/s_n}{Y^\top(I_{n\times n} - P_B)Y/(n - s_n - 1)}.$$

Similar to the modified likelihood ratio test, computation of the F test statistic also requires fitting of both the full and the null models.

The test statistics discussed above are based on the true functional principal component scores. In practice, these scores are unknown and need to be estimated. Estimation of the functional principal component scores has been previously discussed in the literature; for example Yao et al. (2005a), and Zhu, Yao, and Zhang (2014). For completeness, we summarize the common approaches in the Supplementary Material. There are various approaches to estimate the number of functional principal component scores, s_n. A very popular approach in practice is based on the cumulative percentage of explained variance of the functional covariates; commonly used threshold values are 90%, 95%, and 99%. The choice of s_n does not depend on the scalar data Y. Thus, one does not have to choose s_n by a data-driven method such as the AIC criterion. From a practical perspective, there are several packages that provide estimation of the functional principal components scores. For example, `refund` package (Crainiceanu et al., 2012), `fda` package (Ramsay et al., 2011), or PACE package in MATLAB (Müller and Wang 2012). In this paper, we consider two approaches for densely and noisy observed functional data. The first one is to apply a local polynomial smoothing to each individual curve and then employ functional principal component analysis to the smooth curves; see Zhang and Chen (2007) for detail. The second one is to apply the conditional expectation method (Yao, Müller, and Wang 2005a) which was originally developed for sparse and noisy functional observations. Empirical studies, and also our preliminary numerical investigations, have shown that both methods have similar performance numerically for the dense design. Moreover, the conditional expectation approach is applicable to sparse designs. For these reasons, as well as for computational and theoretical simplicity, we consider the first approach to develop the theoretical reasoning and the second approach in practical situations.

Once the truncation level s_n and the functional principal component scores are estimated, the testing procedures are obtained by substituting them with their corresponding estimates. Specifically, let $\widehat{M}$ be matrix of the estimated functional principal component scores, $\widehat{\xi}_{ij}$ defined analogously to M. The expressions of the four tests are obtained by replacing M with $\widehat{M}$. For the hypothesis testing, we not only need the test statistics, but also the null distributions of the test statistics. Similar to testing in linear model, we use chi-square with degree of freedom of s_n as the null distribution for T_W, T_S and T_L and use F with degrees of freedom s_n and $n - s_n - 1$ as the null distribution for T_F.

In this paper, we focus on the eigenbasis function, and expand both functional predictor and coefficient function on the eigenbasis. Actually, one may also use pre-determined basis

systems such as spline or wavelet. These are interesting topics for future research.

3. Theoretical results

As discussed in Section 2, the tests considered - Wald, score, likelihood ratio, and F - resemble their analogue for multivariate covariates, with a few important differences: 1) the number of true functional principal components, s_n, is not known and thus is approximated, and 2) the functional principal component scores ξ_{ij} are not directly observable. In this section, we develop the asymptotic distribution of the tests, when the truncation s_n diverges with the sample size n and the functional principal component scores are estimated using the methods discussed in Section 2. The results are presented for the score and F tests only, which are our recommended tests. Our numerical study showed that both Wald and the modified likelihood ratio tests exhibit significantly inflated Type I error rates, especially when sample size is small, thus we do not recommend these two tests.

First, we present the results of the asymptotic distribution of the test statistics under H_0; all the proofs are included in the Supplementary Material. For the distributions discussed in this section, we refer to the distributions conditional on the original curve $X_i(\cdot)$ and the observed data points $\{W_{i1}, \ldots, W_{im_i}\}$ for $i = 1, \ldots n$. We begin with introducing some notation. In the following, we use T_S for the score statistic and T_F for the F test statistic.

THEOREM 3.1 *Assume that $X_i(\cdot) \in L^2(\mathcal{T})$ for every $1 \leq i \leq n$ and $s_n = o(n)$. Then, if the null hypothesis, that $\beta(t) = 0$ for all t, is true, we have that: (i) $(T_S - s_n)/\sqrt{2s_n} \to^d N(0,1)$, (ii) $(s_n T_F - s_n)/\sqrt{2s_n} \to^d N(0,1)$.*

Under the null hypothesis and conditioning on the number of functional principal components, the distributions of these test statistics are similar to their counterparts in multiple regression. No matter whether measurement error exists, the null distribution would be exactly the same. In particular, for fixed truncation value s_n, the null distribution of the F test statistic behaves like $F_{s_n, n-s_n-1}$ and the null distribution of the score test behaves like $\chi^2_{s_n}$. Similarly, the null distributions of the likelihood ratio and Wald tests behave like $\chi^2_{s_n}$.

Next, we consider the distribution of the tests under the alternative distribution H_a : $\beta(\cdot) = \beta_a(\cdot)$ for some known real-valued function $\beta_a(\cdot)$ defined on $\mathcal{T}$. When the sampling design is dense, we show that the asymptotic results from classical regression continue to hold, and thus estimating the functional principal component scores adds negligible error. Intuitively, this can be explained by the accurate estimation of the functional principal component scores: in the dense design, the score estimators have convergence rate of order $O_P(n^{-1/2})$ (Hall and Hosseini-Nasab 2006).

We begin with describing the assumptions required by our theoretical developments. With a slight abuse of notation, let C denote a generic constant term. Recall that $\{\lambda_j, j \geq 1\}$ are the eigenvalues in strictly decreasing order with $\sum_{j=1}^{\infty} \lambda_j < \infty$, we define $\delta_j = \min_{1 \leq k \leq j} (\lambda_k - \lambda_{k+1})$.

(A) The number of principal components selected, s_n, satisfies the condition $s_n \to \infty$ and $\delta_{s_n}^{-1} s_n = o(n^{1/2})$.

Condition (A) concerns the divergence of the number of functional principal compo-

nents with n. Specifically, it is assumed that this divergence depends on the spacing between adjacent eigenvalues. Our assumption allows s_n to be diverging, but at a much slower rate than n. In fact, by requiring that the spacing between adjacent eigenvalues is not too small, for example $\lambda_j - \lambda_{j+1} \geq j^{-\alpha-1}$ for $j \geq 1$ and some $\alpha > 1$ (Hall and Horowitz 2007), then condition (A) holds if we assume that $s_n^{2\alpha+4} = o(n)$. An example when the latter condition is met is $s_n = O(\log(n))$.

(B1) For all $C > 0$ and some $\epsilon > 0$,

$$\begin{aligned} \sup_{t\in\mathcal{T}}\{E \mid X_i(t) \mid^C\} &< \infty \\ \sup_{t_1,t_2\in\mathcal{T}}(E[\{\mid t_1 - t_2 \mid^{-\epsilon} \mid X_i(t_1) - X_i(t_2) \mid\}^C]) &< \infty. \end{aligned}$$

(B2) For all integers $r \geq 1$, $\lambda_j^{-r} E(\int_{\mathcal{T}}[X_i(t) - E\{X_i(t)\}]\phi_j(t)dt)^{2r}$ is bounded uniformly in j.
Assumptions (B1)-(B2) are common in functional data analysis; see Hall and Hosseini-Nasab (2006). For example, (B1) and (B2) are met when we have a Gaussian process with Hölder continuous sample paths; see Hall and Hosseini-Nasab (2006) for detail.

Denote the bandwidth used for each individual smoothing of the ith curve as h_i. Suppose the support of each trajectory $X_i(t)$ is $\mathcal{T} = [a, b]$, and let $\mathcal{T}_d = [a - d, b + d]$ for some $d > 0$.
(C1) Let $X^{(k)}(t)$ be the kth derivative of $X(t)$. Assume that $X^{(2)}(t)$ is continuous on $\mathcal{T}_d$ with probability 1 and $\int_{\mathcal{T}} E[\{X^{(k)}(t)\}^4]dt < \infty$ with probability 1 for $k = 0, 2$. Also assume that $E\{e_{ij}^4\} < \infty$, where e_{ij}'s are independent and identically distributed, and independent of $X_i(\cdot)$.
(C2) Assume there exists $m \equiv m(n) \to \infty$ such that $\min_{1\leq i\leq n} m_i \geq m$ as $n \to \infty$, and $\max_{1\leq i\leq n} \max_{2\leq k\leq m_i} \{t_{ik} - t_{i(k-1)}\} = O(m^{-1})$.
(C3) Assume there exists a sequence $h = h(n)$, such that $ch \leq \min_{1\leq i\leq n} h_i \leq \max_{1\leq i\leq n} h_i \leq Ch$ for some constant $C \geq c > 0$. Furthermore, $h \to 0$ and $m \to \infty$ as $n \to \infty$ in rates that $(mh)^{-1} + h^4 + m^{-2} = O(n^{-1})$. Also assume that the kernel function $K(\cdot, \cdot)$ is compact supported and Lipschitz continuous.
Assumptions (C1)-(C3) are regularity assumptions for the functional predictor process $X(t)$ for the dense design. They are similar to the Conditions 1–3 in Zhu et al. (2014). Under assumption (C3), we obtain $m \geq Cn^\kappa$ with $\kappa \geq 1/2$. For example, if m achieves order $n^{1/2}$, we require that h is between the rate $n^{-1/4}$ and $n^{-1/2}$.

For a function $\beta_a(\cdot)$, denote $\|\beta_a(\cdot)\|_{L_2} = [\int_{\mathcal{T}}\{\beta_a(t)\}^2 dt]^{1/2}$. The following result presents the asymptotic distribution of the score test statistic, T_S, and the F test statistic, T_F, under the alternative hypothesis.

THEOREM 3.2 *Assume the conditions (A),(B1)(B2),(C1)-(C3) are met. Then under the assumption that $H_a : \beta(\cdot) = \beta_a(\cdot)$ is true and $\|\beta_a(\cdot)\|_{L_2} < \infty$, we have:*
(i) $\{(1 + \int \beta_a(t_1)\beta_a(t_2)K(t_1, t_2)dt_1dt_2)T_S - s_n - \Lambda_n\}/\sqrt{2s_n} \to^d N(0, 1)$,
(ii) $\{s_nT_F - s_n - \Lambda_n\}/\sqrt{2s_n} \to^d N(0, 1)$,
where $\Lambda_n = n \int \beta_a(t_1)\beta_a(t_2)K(t_1, t_2)dt_1dt_2(1 + o_P(1))$.

The proof is included in the Supplementary Material. We want to emphasize that $\beta_a(\cdot)$ is some function that is fixed before observing the data; in particular, we exclude

the case $\beta_a(\cdot) = \phi_{s_n+1}(\cdot)$ because neither $\phi_j(\cdot)$'s nor s_n are known before collecting the data. Nevertheless, if X's span a finite dimensional space, and $\beta_a(\cdot)$ is in the orthogonal complement of the space spanned by the X's, then the testing procedures have no power. Meanwhile, this theorem actually shows that when the design is dense enough, with a proper bandwidth h, the measurement error is asymptotically negligible and does not affect the alternative distribution.

Remark 1 The results presented by Theorems 3.1 and 3.2 are asymptotic results and, while they are interesting, they require large sample sizes to ensure the correct Type I error probability. In practice all the testing procedures discussed above behave like the usual χ^2 and F-distributions with appropriate degrees of freedom, which depend on the sample size n.

If the null hypothesis, that $\beta(t) = 0$ for all t, is true, as in Theorem 3.1, we have that: (i) T_S behaves like $\chi^2_{s_n}$, (ii) $T_F \sim F_{s_n,n-s_n-1}$. If the alternative hypothesis that $H_a : \beta(\cdot) = \beta_a(\cdot)$ is true and $|| \beta_a(t) ||< \infty$ as in Theorem 3.2, and the conditions (A),(B1)(B2),(C1)-(C3) are valid, we have (i) $(1 + \int \beta_a(t_1)\beta_a(t_2)K(t_1,t_2)dt_1dt_2)T_S$ behaves like $\chi^2_{s_n}(\Lambda_n)$, and (ii) T_F behaves like $F_{s_n,n-s_n-1}(\Lambda_n)$, where Λ_n is defined above.

Our empirical investigation showed that these approximate null distributions are substantially more accurate, in terms of Type I error probability. Because of this reason, we use these null distributions in our simulation study.

Remark 2 The alternative distributions discussed in *Remark 2* can be used for sample size calculation. We briefly illustrate the ideas using the F test, T_F. Let K be the covariance function of the functional covariates X_i determined as $K(t_1,t_2) = \sum_{j\geq 1}\lambda_j\phi_j(t_1)\phi_j(t_2)$ and let s be the leading number of eigenfunctions corresponding to some cummulative explained variance threshold, say 99%. Also, assume the true regression parameter function is $\beta(\cdot) = \beta_a(\cdot)$, for $\beta_a(t) \neq 0$ for some $t \in \mathcal{T}$. Then, the asymptotic distribution of T_F corresponding to a sample size n is approximately F with degrees of freedom s and $n-s-1$ respectively and non-centrality parameter $n\Lambda'_a$, denoted by $F_{s,n-s-1}(n\Lambda'_a)$, where $\Lambda'_a = \int \beta_a(t_1)\beta_a(t_2)K_s(t_1,t_2)dt_1dt_2$, and $K_s(t_1,t_2) = \sum_{j=1}^{s}\lambda_j\phi_j(t_1)\phi_j(t_2)$ is the finite dimensional projection of $K(t_1,t_2)$. It follows that, if $F^*_{\alpha,s,n-s-1}$ denotes the critical value corresponding to right tail probability of α under the F distribution with degrees of freedom s and $n-s-1$ respectively, then for sample size n, the power can be calculated as $P\{F_{s,n-s-1}(n\Lambda'_a) > F^*_{\alpha,s,n-s-1}\}$. Therefore, for a power level equal to p_0 and specified level of significance α, one can find an appropriate sample size to detect the effect $\beta_a(\cdot)$ by solving $P\{F_{s,n-s-1}(n\Lambda'_a) > F^*_{\alpha,s,n-s-1}\} \geq p_0$ for n. In practice, the true coefficient function $\beta_a(\cdot)$ and covariance function $K(\cdot,\cdot)$, or its finite dimensional projection, $K_s(\cdot,\cdot)$ can be estimated from prior studies. We plug in the estimates of these quantities and calculate the sample size needed. Section 7.2 illustrates an excellent performance of the asymptotic power curves for the F test in finite samples, and employs these ideas for the calculation of sample sizes.

Next we present the rate of our testing procedures. The following corollary shows that these tests can detect a local alternative of order $\sqrt{s_n}/n$; thus they achieve the same optimal rate as the goodness-of-fit test for the high dimensional linear model when the number of parameters is s_n (Verzelen and Villers 2010).

COROLLARY 3.3 *Let $\beta_a(\cdot)$ be a nonzero function on the same order of* 1. *Consider the sequence of local alternatives $H_a : \beta(\cdot) = \rho_n\beta_a(\cdot)$. Under the conditions (A),(B1)(B2),(C1)-(C3), our test statistics, T_S and T_F are powerful if $\sqrt{s_n}/(n\rho_n^2) = O(1)$.*

4. Extension to sparsely and noisy observed functional covariate

In practical applications, we often observe sparse realizations of the functional covariate which in addition are corrupted with measurement error; this setting is commonly known as 'sparse design'. Our testing procedures can still be applied to this scenario, with the difference that the estimated functional principal component scores account for the sparse design. In particular, the two step procedure of first curve smoothing and then functional principal component analysis of Zhang and Chen (2007) is no longer applicable. Instead, the conditional expectation method (Yao, Müller, and Wang 2005a) is used; to avoid redundancy, the estimation procedure is detailed in the Supplementary Material. The null distributions of the testing procedures are similar to the dense design case: we use chi-square with degree of freedom of s_n for the null distribution for T_W, T_S and T_L and use F with degrees of freedom s_n and $n - s_n - 1$ for the null distribution for T_F. Similar to the dense design both Wald test, T_W, and the modified likelihood ratio test, T_L, show inflated Type I error. The following corollary gives the asymptotic null distribution of the recommended tests T_S and T_F:

COROLLARY 4.1 *Under the sparse design, assume that $X_i(\cdot) \in L^2(\mathcal{T})$ for every $1 \leq i \leq n$ and $s_n = o(n)$. Then, if the null hypothesis, that $\beta(t) = 0$ for all t, is true, we have that: (i) $(T_S - s_n)/\sqrt{2s_n} \to^d N(0,1)$, (ii) $(s_n T_F - s_n)/\sqrt{2s_n} \to^d N(0,1)$.*

We emphasize that the asymptotic null distribution holds true irrespective of the sampling design (sparse or dense) of the functional covariates, or whether the functional covariate is measured with noise. In particular we can still use the approximate null distributions $F_{s_n, n-s_n-1}$ and $\chi^2_{s_n}$ for T_F and T_S, respectively. This finding is not surprising, since the null distribution of the tests is derived using the true model, i.e. $\beta(\cdot) \equiv 0$, and thus it is not affected by the sampling design of the functional covariate.

5. Extension to partial functional linear regression models

Often, of interest, is to investigate the association between a scalar response and a functional covariate, while accounting for other covariate information that is available. For example, in our tractography study the interest is to test for the association between the cognitive score of multiple sclerosis patients and their fractional anisotropy along the white matter tract by accounting for the patients' sex and age; see Section 6.1 for details. Thus model (1) cannot be used per se; however it can be modified to account for additional covariates.

More generally, we define the following modeling framework. Let the observed data be $[Y_i, \{W_{ij}, t_{ij}, j = 1, \ldots, m_i\}, Z_i]_i$ where Y_i and $W_{ij} = W_i(t_{ij})$ are the response and the noisy functional predictors, respectively, like in Section 2, and Z_i is a vector of covariates for subject i. We consider the partial functional linear model

$$Y_i \quad = \quad Z_i^\top \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i, \tag{6}$$

where $X_i(\cdot)$ is the true functional predictor, $\beta(\cdot)$ is the interest parameter function and α is $(p+1)$-dimensional vector of nuisance parameters. For notation simplicity assume that the first element of Z_i is 1. This model has been studied by Shin (2009) and Li, Wang, and Carroll (2010).

The objective is to test the hypothesis $H_0 : \beta(t) = 0$ for all t, by accommodating

nuisance parameters using the modeling framework (6). The four testing procedures can be easily extended to this setting. As in Section 2.2, the approach is based on using a (3), obtained by approximating the model using a truncated number s_n of the functional principal component scores. Let Z be the $n \times (p+1)$ dimensional matrix obtained by row-stacking Z_i^T, and let M be the $n \times s_n$ dimensional matrix of the functional principal component scores as defined in Section 2.2. Then conditional on the truncation level and the true functional principal component scores, the log likelihood function can be written as $L_n(\sigma^2, \alpha, \beta) = -(n/2)\log(2\pi\sigma^2) - (Y - Z\alpha - M\beta)^\top (Y - Z\alpha - M\beta)/(2\sigma^2)$ which resembles to (5) with the modification that the 1_n vector is replaced by the matrix Z.

The score function and the information matrix can be derived accordingly; the Wald, likelihood ratio and F test statistics follow easily. In particular, the maximum likelihood estimate of σ^2 is $\widetilde{\sigma}^2 = Y^\top (I_{n\times n} - P_Z)Y/n$, and the constrained maximum likelihood estimate of σ^2 is $\widehat{\sigma}^2 = Y^\top (I_{n\times n} - P_B)Y/n$, where $B = [Z, M]$ is defined correspondingly to this setting. Furthermore, the score test statistic is given by $T_S = Y^\top (P_B - P_Z)Y/\widetilde{\sigma}^2$. Here P_B and P_Z denote the projection matrices for B and Z respectively and, for completeness, are included in the Supplementary Material. Likewise, the Wald, likelihood ratio and F test statistics are included in the Supplementary Material.

In practice, the test statistics are calculated based on the estimated functional principal component scores, and thus based on the estimated design matrix $\widehat{M}$, as detailed in Section 2.2. Under the null hypothesis that $\beta(\cdot) \equiv 0$, the null distribution of T_W, T_S and T_L can be approximated by $\chi^2_{s_n}$, while the null distribution of T_F is approximately $F_{s_n, n-s_n-(p+1)}$, where the degrees of freedom are changed from (3.1) to account for the dimension of the nuisance parameter.

A more general extension is the partially functional linear regression model with multiple functional predictors (Kong, Xue, Yao, and Zhang 2016).

$$Y_i \quad = \quad Z_i^\top \alpha + \sum_{\ell=1}^{d} \int_{\mathcal{T}_\ell} X_{i\ell}(t)\beta_\ell(t)dt + \epsilon_i, \tag{7}$$

where $X_{i\ell}(\cdot)$ is the ℓth functional predictor, $\beta_\ell(\cdot)$ is the corresponding regression parameter function. The objective is to test the hypothesis $H_0 : \beta_\ell(t) = 0$ for all $t \in \mathcal{T}_\ell$ and all $\ell \in \mathbb{L}$, where $\mathbb{L}$ is a subset of $\{1, \ldots, d\}$.

The four testing procedures can be easily extended to this setting. Suppose we select $s_{n\ell}$ principal components for the ℓth functional predictor $X_\ell(\cdot)$. Let Z be the $n \times (p+1)$ dimensional matrix obtained by row-stacking Z_i^T, and let M_1 be the $n \times \sum_{\ell\in\mathbb{L}} s_{n\ell}$ dimensional matrix of the functional principal component scores whose corresponding indices are in $\mathbb{L}$, and let M_2 be the $n \times \sum_{\ell\notin\mathbb{L}} s_{n\ell}$ dimensional matrix of the functional principal component scores whose corresponding indices are not in $\mathbb{L}$. Define β_1 to be the coefficient corresponding to M_1 and β_2 to be the coefficient corresponding to M_2. Let $Z_1 = [Z, M_2]$ and $\eta = (\alpha^\top, \beta_2^\top)^\top$. Then conditional on the truncation level and the true functional principal component scores, the log likelihood function can be written as $L_n(\sigma^2, \eta, \beta_1) = -(n/2)\log(2\pi\sigma^2) - (Y - Z_1\eta - M_1\beta_1)^\top (Y - Z_1\eta - M\beta_1)/(2\sigma^2)$. The score function and the information matrix can be derived accordingly; the Wald, likelihood ratio and F test statistics follow easily. We use illustrate these ideas in the tractography data application where we assume a modeling framework as (7).

6. Real data application

6.1. *The Diffusion Tensor Imaging data*

Consider our motivating application, the Diffusion Tensor Imaging tractography study, where we investigate the association between cerebral white matter tracts in multiple sclerosis patients and cognitive impairment. The study has been previously described in Greven, Crainiceanu, Caffo, and Reich (2010); Staicu, Crainiceanu, Ruppert, and Reich (2012); Goldsmith, Feder, Crainiceanu, Caffo, and Reich (2011), and we discuss it briefly here. Multiple sclerosis is a demyelinating autoimmune disease that is associated with lesions in the white matter tracts of affected individual and results in severe disability. Diffusion Tensor Imaging is a magnetic resonance imaging technique that allows the study of white matter tracts by measuring the diffusivity of water in the brain: in white matter tracts, water diffuses anisotropically in the direction of the tract. Using measurements of diffusivity, Diffusion Tensor Imaging can provide relatively detailed images of white matter anatomy in the brain (Basser, Mattiello, and LeBihan 1994; Basser, Pajevic, Pierpaoli, and Duda 2000). Some measures of diffusion are fractional anisotropy, and parallel diffusivity among others. For example, fractional anisotropy is a function of the three eigenvalues of the estimated diffusion process that is equal to zero if water diffuses perfectly isotropically, such as Brownian motion, and to one if water diffuses anisotropically, such as for perfectly organized and synchronized movement of all water molecules in one direction. The measurements of diffusion anisotropy are obtained at every voxel of the white matter tracts; in this analysis, we consider averages of water diffusion anisotropy measurements along two of the dimensions, which results in a functional observation with scalar argument that is sampled densely along the tract.

Here we study the relationship between the fractional anisotropy along the two well identified white matter tracts, corpus callosum and left corticospinal tracts, and the multiple sclerosis patient cognitive function, as measured by the score at a test, called Paced Auditory Serial Addition Test. Specifically, each multiple sclerosis subject is given numbers at three second intervals and asked to add the current number to the previous one. The score is obtained as the total number of correct answers out of 60.

The study, in its generality, comprises 160 multiple sclerosis patients and 42 healthy controls observed at multiple visits spanning up to four years. For each subject, at each visit, are recorded: diffusion anisotropy measurements along several white matter tracts at many hospital visits, as well as additional information such as age, gender and so on. In this analysis, we use the measurements obtained at the baseline visit. Because Paced Auditory Serial Addition Test was only administered to multiple sclerosis subjects, we limit our analysis to the multiple sclerosis group. Few subjects do not have Paced Auditory Serial Addition Test scores recorded and they are removed from the analysis, leaving 150 multiple sclerosis patients in the study. Part of these data is available in the R-package `refund` (Crainiceanu et al. (2012)). For illustration, Figure 1 shows the fractional anisotropy along the corpus callosum (left panel) and corticospinal tracts (middle) tracts, and the Paced Auditory Serial Addition Test scores (right panel) for all the subjects in the study. Depicted in solid black/solid gray /dashed black are the fractional anisotropy measurements of three different subjects, with each line type representing a subject. Our goal is to test for association between the Paced Auditory Serial Addition Test score in multiple sclerosis patients and the fractional anisotropy along the corpus callosum and the corticospinal tracts, while accounting for age and gender.

Consider first the corpus callosum tract, which has an important role in the cognition function. Fractional anisotropy is measured at 93 locations along this tract: the

measurements include missingness and measurement error. Using our notation, let W_{ij} denote the noisy fractional anisotropy observed at location t_{ij} for the ith subject, Z_i is the three-dimensional vector encompassing the intercept, the subject's age and gender, and let Y_i be the Paced Auditory Serial Addition Test score of the ith multiple sclerosis patient. We assume a partial functional linear model for the dependence between the Paced Auditory Serial Addition Test score and true the fractional anisotropy along the corpus callosum tract of the form (6), where Y_i and Z_i are defined above, and $X_i(\cdot)$ is the underlying smooth fractional anisotropy defined on $\mathcal{T} = [0, 93]$. Here $\beta(\cdot)$ is a parameter function and main object of interest, describing a linear association between the corpus callosum fractional anisotropy and the Paced Auditory Serial Addition Test score, and α is a vector parameter accounting for a linear covariate effect. For simplicity, the age is standardized to have mean zero and variance one and the fractional anisotropy profiles are mean de-trended to have, at each location, mean zero across all the subjects. We are interested in testing the null hypothesis that the parameter function $\beta(\cdot)$ is equal to zero.

As discussed in Section 2 the preliminary step of the hypothesis testing is the estimation of the subject specific functional principal component scores corresponding to the fractional anisotropy profiles along the corpus callosum tract. We use functional principal component analysis through conditional expectation Yao et al. (2005a), and select the number of eigenfunctions using the cumulative explained variance. The results yield that five eigenfunctions are required to explain 90% of the variability in the data, while 15 are required to explain 99% of the variability. For stability reasons, we take a more conservative approach and select the number of eigenfunctions using 90% cumulative explained variance. Figure 2, top three panles and bottom two left most panels display the estimated leading eigenfunctions along with the corresponding estimated eigenvalues; the variance of the measurement error in the functional covariate is estimated to $\tilde{\sigma}_e^2 = 0.002 \times 10^{-2}$.

Then, we test whether the coefficient function $\beta(\cdot)$ is zero along the corpus callosum, by accounting for age and gender effects using the methods discussed in Section 2.2. Figure 2 the bottom right panel depicts the estimated regression function for the fractional anisotropy along the corpus callosum, $\widehat{\beta}(t)$. It indicates that the multiple sclerosis subjects who have a higher than average fractional anisotropy along the middle area of corpus callosum tend to score higher on the Paced Auditory Serial Addition Test. The p-value for testing that $\beta(\cdot) = 0$ reported by the F statistic equals 2.33×10^{-4}, indicating very strong evidence of association. This result is consistent across the other testing procedures: the modified likelihood ratio test p-value is 1.57×10^{-4}, the Wald p-value is 1.03×10^{-4}, while the score p-value is 3.42×10^{-4}. As one annonymous reviewer suggested, we also provide the other estimated model components, for completeness. The estimated intercept is $\widehat{\alpha}_1 = 44.205$, the estimated effects associated with the gender and age are $\widehat{\alpha}_2 = -0.979$ and $\widehat{\alpha}_3 = -0.305$ respectively, and the estimated model variance is $\widehat{\sigma}^2 = 144$. These results confirm our expectation that the cognitive performance of the multiple scelrosis subjects, as assessed via the Paced Auditory Serial Addition Test, is negatively associated with age and furthermore show that it tends to be lower for women than men.

Next, we are interested to assess whether the fractional anisotropy along the left corticospinal tract adds significantly to a model fit for the Paced Auditory Serial Addition Test score with the fractional anisotropy along the corpus callosum. Fractional anisotropy is measured at 55 locations along the corticospinal tracts; the missingness along this tract is notably larger than along the corpus callosum. We assume modeling framework as in (7), $Y_i = Z_i^T \alpha + \int_{\mathcal{T}_{CCA}} \beta_{CCA}(t) X_{i1}(t) dt + \int_{\mathcal{T}_{lCST}} \beta_{lCST}(t) X_{i2}(t) dt + \epsilon_i$; here Y_i and Z_i are defined as above, $X_{i1}(\cdot)$ and $X_{i2}(\cdot)$ are the underlying smooth fractional anisotropy along the corpus callosum and left corticospinal tract, respectively, $\mathcal{T}_{CCA} = [0, 93]$ and

$\mathcal{T}_{lCST} = [0, 55]$. Furthermore, $\beta_{CCA}(\cdot)$ and $\beta_{lCST}(\cdot)$ are the parameter functions quantifying the effect of the fractional anisotropy along the two tracts onto the test score. We are interested to test the null hypothesis that $\beta_{lCST}(\cdot)$ is equal to zero.

As before, we first apply functional principal component analysis to both sets of functional covariates; for the fractional anisotropy along the left corticospinal tract we select the number of eigenfunctions using 90% explained variance (which results to 8 eigenfunctions) and estimate the functional principal component scores. The percentage of explained variance was again selected for stability reasons; in particular 99% variability is explained by 15 eigenfunctions. Using the methods discussed in the paper to assess the testing hypothesis of no relationship between the test score and the fractional anisotropy along the left corticospinal tract while accounting for the other covariates, we obtain a p-value of 0.0771 using F test (0.0624 with modified likelihood ratio test, 0.0670 using Wald and 0.0641 with score test statistic). Thus there is no significant relationship between the cognitive function as assessed by Paced Auditory Serial Addition Test and the corticospinal tracts tract, as measured by fractional anisotropy at level of significance 5%, when the model accounts for fractional anisotropy along the corpus callosum.

Overall, our findings corroborate the specialists' prior expectations that the cognitive function is associated with the corpus callosum tract. Further results (not included here) show surprising association of the cognitive function with the corticospinal tract, when the model accounts only for age and gender. However as we show above, the association is not significant if the model accounts for the corpus-callosum fractional anisotropy. Interestingly, both findings are in agreement with Swihart et al. (2014), who used the fractional anisotropy along the two tracts of the multiple sclerosis subjects measured at all the available hospital visits and a restricted likelihood ratio-based testing approach.

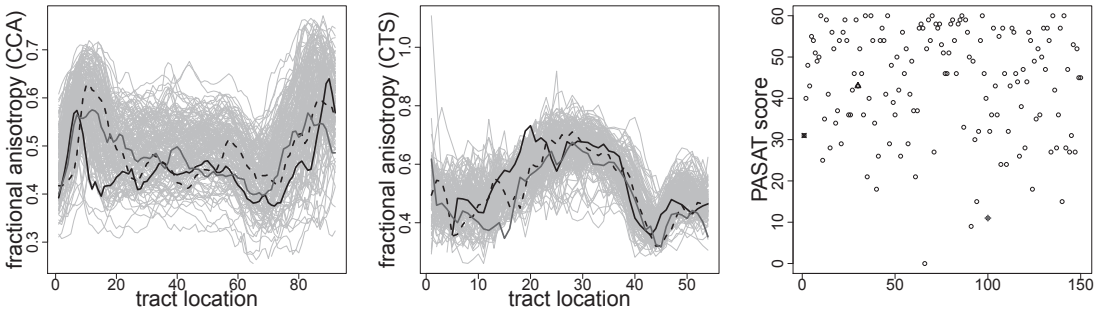


Figure 1. Fractional anisotropy profiles along corpus callosum (left) and corticospinal tracts (middle) and the associated Paced Auditory Serial Addition Test scores (right panel) in the group of multiple sclerosis patients. Depicted in different colors and line/symbols styles are the measurements of three subjects.

6.2. *The Microsoft Xbox auction data*

Next, we consider an application from electronic commerce (eCommerce) field. The eBay auction data set (Wang, Jank, and Shmueli 2008) consists of time series of bids placed over time for 172 auctions for *Microsoft Xbox* gaming systems, which are very popular items on eBay. For each auction, the associated time series is composed of bids made by users located at various geographical locations, and thus it shows very uneven features. In addition, the time between the start and the end of an auction varies across auctions, and furthermore the actions duration varies across actions. Nevertheless, as Jank and Shmueli

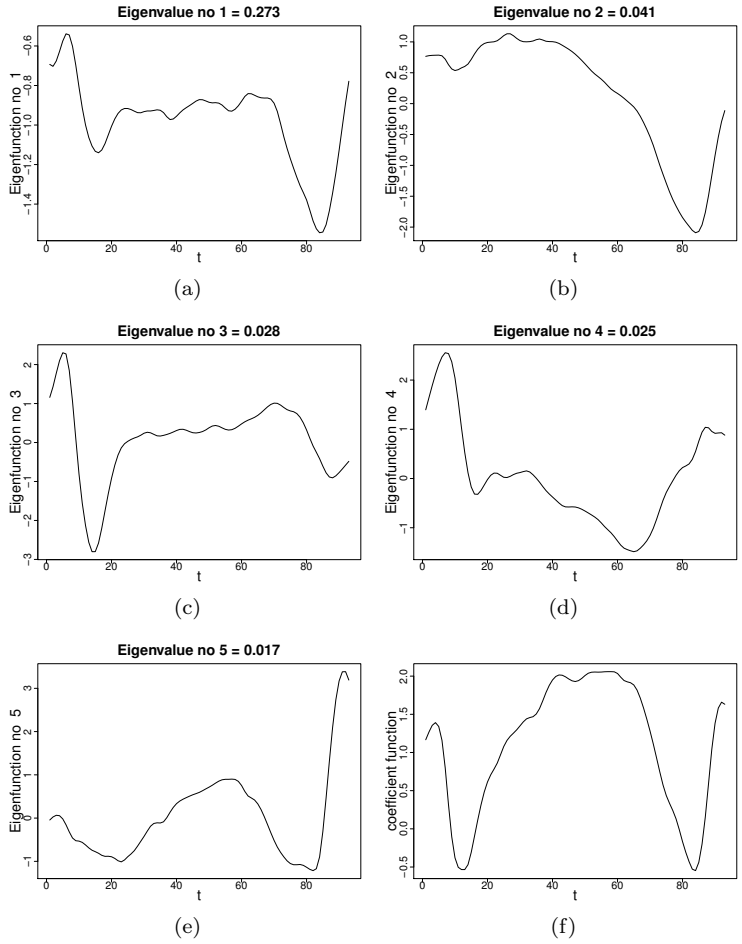


Figure 2. Panels (a),(b),(c), (d), and (e) show the top five estimated eigenfunctions $\tilde{\phi}_k(\cdot)$, along with the associated eigenvalues, $\tilde{\lambda}_k$ (at a scale of 10^{-2}). Panel (f) shows the estimated coefficient function $\tilde{\beta}(\cdot)$.

(2006) argues "bidding in eBay auctions tends to be concentrated at the end, resulting in very sparse bid-arrivals during most of the auction except for its final moments, when the bidding volume can be extremely high". The dynamics of the bids has attracted large interest, especially in the literature of functional data (Liu and Müller 2008). Here we investigate whether the dynamics of the bids in the first part of the auction duration is related to the auction's closing price.

To handle the challenge of different starting times and durations of the auctions, we think of the bids for an action as varying with the percentile of the auction length (see also Jank and Shmueli (2006)). For example if an auction has a length of 7 days, then the bid placed in the 5th day from the starting time corresponds to 71.4 percentile of the auction's duration. Here we focus on the bids placed in the first 71.4% of the auction's duration, and study whether their dynamics influences various measures of the closing price of the auction. To be specific define the formation of the price during the first 71.4% of duration of an action as the process of interest observed with noise. Using the notation in Section 2, let W_{ij} denote the bid placed for action i at the $100 \times t_{ij}$ percentile of the auction's length, where $t_{ij} \in [0, .714]$, and assume that W_{ij} represents the true auction's price $X_i(t_{ij})$ observed at $100 \times t_{ij}$ percentile with noise. We investigate whether the underlying partial auction curve influences: (1) the relative change in the final price of the auction, and (2) the rate of change in the final price.

Before we tackle these two important problems, we carefully examine the data. A close inspection confirms that most auctions have a duration of at least 7 days and thus the auctions with length less than 7 days are removed. Also we remove all the auctions for which there is only one bid in the first 71.4% of the auction duration. The remaining data set contains bids from 125 Xboxes auctions. Moreover, for very action, the number of bids placed in the first 71.4% of the auction's duration, varies between 2 to 14. Our analysis regards the observed partial auction curve as a noisy functional predictor observed at sparse and irregular time points in $\mathcal{T} = [0, .714]$.

For the first objective, the response for each action i, is taken as the relative change in the final price, as defined as $Y_i = (V_i - W_{im_i})/W_{im_i}$, where V_i is the final auction price, W_{im_i} is the bid placed at the largest percentile less than or equal to 71.4 for auction i. We assume that the relation between the underlying partial auction curve and the relative change in the final price is modeled using a functional linear model of the form (1) and are interested to test that there is no association between them. We apply the methods outlined in Section 2, and in particular, we begin with a functional principal component analysis for sparse sampling design through conditional expectation (Yao et al. 2005a). The top four eigenfunctions are required to explain 99% explained variance and the functional principal component scores are estimated using conditional expectation. We have plotted them in Figure 3. Then we perform the test statistics: the p-value reported by the F statistic equals 5.4×10^{-4} indicating very strong evidence of association. This result is similar for the other testing procedures: the modified likelihood ratio test p-value is 4.2×10^{-4}, the Wald p-value is 2.7×10^{-4}, while the score p-value is 8.3×10^{-4}.

One might also be interested in the relationship between the auction price during the first part of the week and the final price. We performed this analysis and found the following results: p-value reported by the F statistic is 0, the likelihood ratio p-value is 0, the Wald p-value is 0, and the score p-value is 9.2×10^{-15}. These results indicate very strong evidence of association between the price at the beginning of the week and the final price.

Next, we turn to the second objective, and re-define the response for each action i, as the rate of change in the final price. Specifically let $Y_i = (V_i - W_{im_i})/(1 - t_{im_i})$, where V_i and W_{im_i} are defined as above, and $100 \times t_{im_i}$ is the percentile of the ith auction's length corresponding to W_{im_i}. The interest is to test that there is no association between the rate of change in the final auction's price and the underlying partial auction curve. We use the estimated functional principal component scores obtained earlier and test the hypothesis of no association via the four testing procedures. We find that the p-values for the F, score, modified likelihood ratio test, Wald tests are 0.0011, 0.0015, 0.0006 and 0.0009 respectively, indicating significant association. In conclusion, our analysis provides novel insights into the bidding dynamics: namely that the bidding trajectory during the first 71.4% of an auction's length is associated with both the relative change of the final auction price as well as its rate of change.

7. Simulation study

The performance of the Wald, score, modified likelihood ratio test and F tests in terms of Type I error and power is investigated in a simulation experiment. First we consider a functional linear model and study the tests performance under various sample sizes and sampling designs for the functional covariate (Section 7.1). Moreover, we illustrate how to use the asymptotic alternative distribution of the tests to calculate the ideal sample size to detect a specified alternative (Section 7.2). Finally, we consider a partial functional

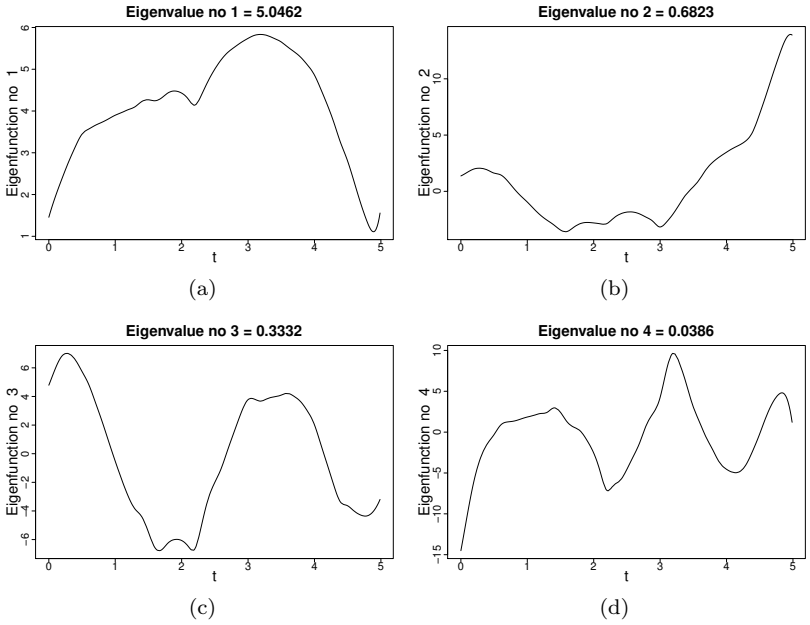


Figure 3. Panels (a),(b),(c), and (d) show the top four estimated eigenfunctions $\tilde{\phi}_k(\cdot)$, along with the associated eigenvalues, $\tilde{\lambda}_k$ (at a scale of 10^3).

linear model, in an attempt to mimic the Diffusion Tensor Imaging data generation process, and evaluate the tests performance, when the model is misspecified (Section 7.3).

7.1. *Functional linear model*

The underlying generating process for the ith functional covariate is $X_i(t) = \sum_{j\geq 1} \xi_{ij}\phi_j(t)$, where ξ_{ij} are generated independently as $N(0, \lambda_j)$, for $\lambda_1 = 16$, $\lambda_2 = 12$, $\lambda_3 = 8$, $\lambda_4 = 4$, $\lambda_5 = 2$, $\lambda_6 = 1$ and $\lambda_k = 0$ for $k \geq 7$. Also ϕ_k are Fourier basis functions on $[0, 10]$ defined as $\phi_1(t) = \cos(\pi t/10)/\sqrt{5}$, $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, $\phi_3(t) = \cos(3\pi t/10)/\sqrt{5}$, $\phi_4(t) = \sin(3\pi t/10)/\sqrt{5}$, $\phi_5(t) = \cos(5\pi t/10)/\sqrt{5}$, $\phi_6(t) = \sin(5\pi t/10)/\sqrt{5}$, $0 \leq t \leq 10$. The observed functional covariate is taken as $W_i(t) = X_i(t) + e_i(t)$, where the measurement error process $e_i(\cdot)$ is assumed Gaussian with mean zero and covariance $\text{cov}\{e_i(t), e_i(s)\} = I(t = s)$.

We consider there types of sampling designs for the functional covariate.

- Design 1: (Dense design). The observed points on each curve are an equally spaced grid of 300 points in $[0, 10]$.
- Design 2: (Moderately sparse design with a few points). The number of points per curve, m_i, is moderate and varies across subjects. Specifically, m_i is chosen randomly from a discrete uniform distribution on $\{5, 6, 7, 8, 9, 10\}$. Each curve is assumed to be observed at m_i points that are randomly selected from the set of 501 equally spaced points in $[0, 10]$.
- Design 3: (Very sparse design). The number of points per curve is small and varies across subjects. Similar generating process of the sampling points as Design 2, with exception that the number of measurements m_i is chosen from a discrete uniform distribution on $\{2, 3, 4\}$.

The response Y_i is generated from model (1), where $X_i(\cdot)$ are generated as above, $\epsilon_i \sim$

$N(0,1)$ and the coefficient function $\beta(\cdot)$ is equal to

$$\beta_c(t) \quad = \quad c\{1+\exp(1-0.1t)\}^{-1}, \tag{8}$$

where $c \geq 0$ is a parameter that controls the departure from the null function. The performance of the tests was assessed in testing the hypothesis $H_0 : \beta(\cdot) \equiv 0$, when the sample size increases from 50 to 500. For Type I error rate performance, we consider data generated from the above model when $\beta(\cdot) = 0$ corresponding to $c = 0$. For power performance, we consider $\beta(\cdot) = \beta_c(\cdot)$ corresponding to $c > 0$ for c taking values in grid of 15 equally spaced points in $[0.02, 0.3]$.

The four tests were calculated as described in Section 2, after having estimated the functional principal component scores as a preliminary step. For the latter, the estimation of the functional principal component scores was obtained using the Matlab package, PACE, available at http://anson.ucdavis.edu/∼ntyang/PACE. The number of functional principal components is selected such that the cumulative explained variance is 99%; other threshold levels have been also investigated, and the results remained in general unchanged. We used 5000 simulated data sets are used to estimate the Type I error rate and 1000 simulated data sets to estimate the power.

The results are presented in Figure 4, and correspond to fixing the level of significance at 5%. Figure 4 (a) shows the performance of the tests with respect to Type I error rate for various sampling designs and as the sample size increases from 50 to 500. In particular, F test gives reasonable Type I errors for all the designs and various sample sizes. The score test seems to be somewhat conservative for small samples for all the sampling designs, while Wald and the modified likelihood ratio test indicate an inflated Type I error for small and moderate sample sizes ($n = 50$ or $n = 100$). For large sample size ($n = 500$), all of the tests give Type I error rates close to the nominal level.

Figure 4 (b)-(d) display the power performance of the tests for the dense sampling design and various sample sizes. The tests have comparable power for all sample sizes investigated. The results are similar for the other two designs and are included in the Supplementary Material: as expected, the power of the tests decreases with the sparseness of the design.

One neat property of selecting the number of principal components using the cumulative percentage of explained variance of the functional covariates, and thus not involving the data Y, is that the technique does not require multiple testing correction. For example, Hilgert et al. (2013) proposed AIC-based selection of the number of principal components using some grid search $\{S := 1, 2, 4, 8 \dots s_{max}\}$ combined with a multiple testing procedure. However, the downside of our approach is that selecting an unnecessarily large number of components may result in a loss of power of the testing procedure. To gain more insight, we compared numerically our proposed methods and the minimax adaptive testing method proposed in Hilgert et al. (2013); this comparison is included in the Supplementary Material, Section 5.2, due to space limitation of the paper. Furthermore we compared the F test with our discussed asymptotic null distribution with two other alternatives: 1) the F test with the bootstrap-based approximation of its null distribution discussed in González-Manteiga et al. (2014) and 2) the likelihood ratio test of Swihart et al. (2014) for single functional covariate. The results are described in the Supplementary Material. Some of the competitive methods, namely the minimax and the boostrap methods, are not applicable to the case when the functional covariate is measured at an irregular sparse design, nor corrupted with measurement error; thus we restricted the comparison to the dense design scenario and when the covariate is measured without noise. We found that the Type I error is significantly inflated for the

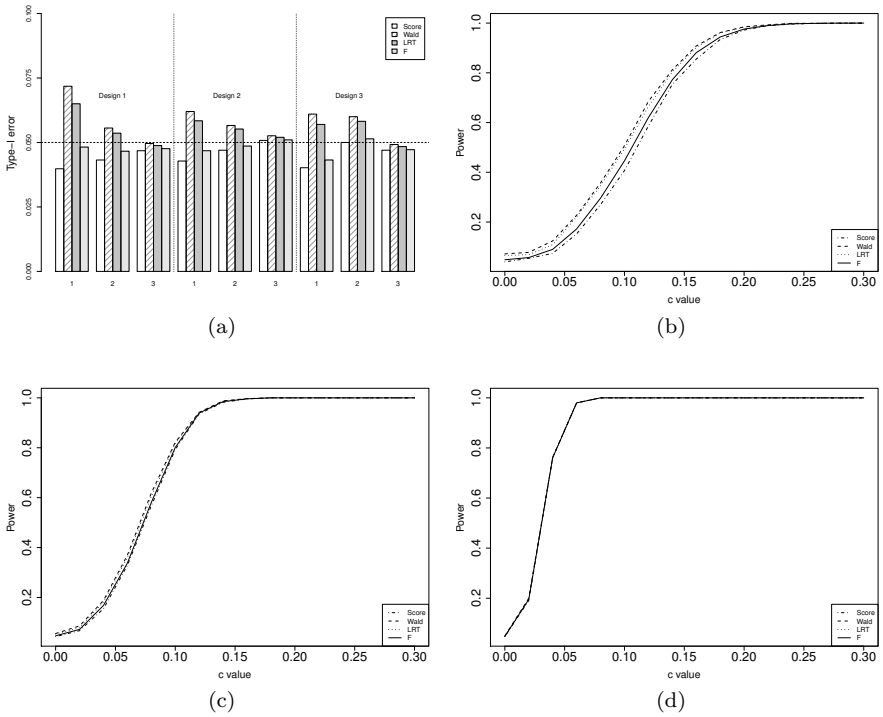


Figure 4. Panel (a) shows the estimated Type I error (depicted as the height of the bars) for all the four tests in nine settings obtained from combining three sampling designs and sample sizes when the nominal level is 5% (horizontal dashed red line). The bars are first grouped according to the sample size (50, 100, and 500, labeled by the digits 1, 2, and 3 respectively on the horizontal axis), and then separated by designs (Design 1, Design 2, and Design 3). Panel (b),(c) and (d) correspond to the changes of the power for Design 1, sample size 50, 100, and 500 respectively.

likelihood ratio test; this is in agreement with our own numerical experience, that the likelihood ratio test yields inflated Type I error. The results show similar performance in terms of Type I error rate and power for the minimax and the bootstrap methods. The advantage of our methods, over with the minimax and the bootstrap methods, is that they accommodate irregular or sparse designs and measurement error in the functional covariate. Finally, we investigated the robustness of the testing procedures to normality: the true functional covariate is generated such that the functional principal component scores have the scaled t_3 (heavy tailed) distribution. Our finding is that the procedures are not sensitive to non-normal distribution of the scores; the results are described in the Supplementary Material.

We have also conducted simulations for the case when $X_i(t)$ is generated from a large number of basis functions, and we have performed simulations to see the performance of our test when we use different thresholds of percentages of variance explained (85%, 90%, 99%) to choose the number of principal components. We have included the results in Figures S4–S13 in the supplementary materials. We have found that the Type I error performance is quite similar when we use different thresholds. For the power performance, when sample size is small ($n = 50$), there would be a little bit power loss when we use a larger threshold 99% compared with smaller thresholds 85%, 90%, but when the sample size becomes large ($n = 500$), the power performance is quite robust to the choice of the thresholds. It would be desirable to develop a method to select the number of basis functions, and it is an interesting topic for future research. As one referee pointed out, Su and Hsu (2016) developed a method to select the number of basis functions when

studying the same testing problem presented in our paper.

7.2. *Sample size calculation*

In this section, we discuss how to employ the asymptotic distribution of the tests under the alternative hypothesis to calculate appropriate sample sizes for detection of the effect, when both the power and the precision are a priori specified. This research direction is novel and has not been addressed hitherto in the literature of functional data analysis. We begin by assessing the accuracy of the asymptotic distribution of the tests under the alternative hypothesis in finite sample sizes. The intuition is that if the alternative asymptotic distribution of a test has good performance in finite samples, then this distribution can be used for sample size calculation, just as in typical linear regression.

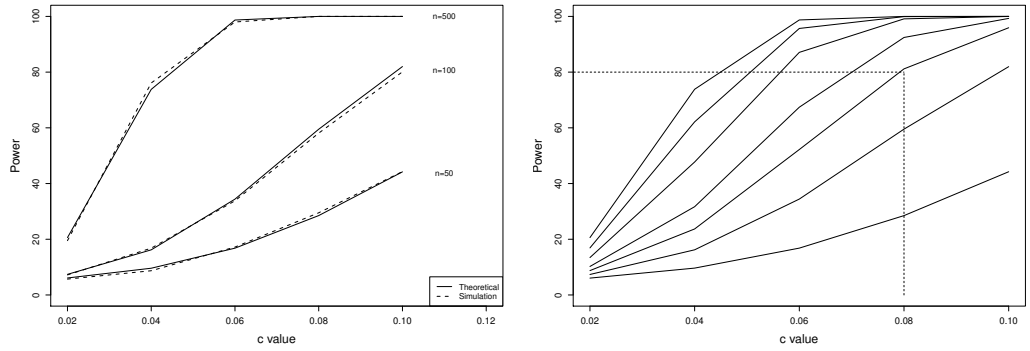
Consider model (1) where the response Y_i is generated as described in the previous section, and the covariate X_i is observed at dense design (Design 1). Also the true regression parameter function is $\beta(\cdot) = \beta_c(\cdot)$, for $c > 0$, where the scaling parameter c controls the departure of the parameter function $\beta_c(\cdot)$ from the null function. The results focus on the F test, T_F, employed for testing the null hypothesis $H_0 : \beta(\cdot) = 0$. The theoretical power of the test can be calculated using Theorem 3.2, and following the approach outlined in Section 3. In particular, for sample size n, the power curve, as a function of c, can be approximated by $P\{F_{s,n-s-1}(n\Lambda_c') > F^*_{\alpha,s,n-s-1}\}$, where $F_{s,n-s-1}(n\Lambda_c')$ denotes F distribution with degrees of freedom s and $n-s-1$, respectively, and non-centrality parameter $n\Lambda_c'$, where $F^*_{\alpha,s,n-s-1}$ denotes the critical value corresponding to right tail probability of α under $F_{s,n-s-1}(0)$, $\Lambda_c' = \int \beta_c(t_1)\beta_c(t_2)K_s(t_1,t_2)dt_1dt_2$, and s is the leading number of eigenfunctions of the covariance function $K(\cdot,\cdot)$.

Figure 5 (a) displays the power of the F test, as a function c, when the level of significance is fixed at 5%. Empirical and theoretical power curves are compared for varying sample sizes, $n = 50$, $n = 100$ and $n = 500$. The empirical power curves (dashed lines) are basically the power curves of the F test that are shown in Figure 4 panels (b)-(d) and restricted to the domain $(0, 0.1]$. Theoretical power curves (solid lines) are calculated using `R` software to compute various probabilities and quantiles corresponding to F distribution of various degrees and different values for the non-centrality parameter.

For fixed sample sizes, the theoretical and empirical power curves are very close, indicating that the asymptotic distribution of the F test under alternative is reliable for calculation of sample sizes. For example, consider model (1), assume that there is a linear association between the response and the functional covariate, and that the true regression parameter is $\beta(\cdot) = \beta_{0.08}(\cdot)$. Then, corresponding to a power level of at least 80%, the smallest sample size at which one can detect significant association at tolerance level of 0.05 is $n = 150$. In Figure 5 (b) this is represented by tracing up the vertical line at $c = 0.08$ that corresponds to parameter function $\beta_{0.08}(\cdot)$ to intersect the power curves of different sample size, at different power levels. The smallest sample size at which the power level is at least 80% is the desired sample size.

The sample size calculation is illustrated on the F test, mainly because the alternative asymptotic distribution of this test is very accurate, even for smaller samples. In particular Wald and the modified likelihood ratio tests yield inflated Type I error rate for moderately large sample size. For the score test, close agreement between the asymptotic and empirical power approximation occurs when the sample size is large. Because of these considerations, our recommendation is to use F test for sample size calculations.

The sample size calculation is an important novelty of this paper. Indeed it depends on the true covariance surface $K(\cdot,\cdot)$ and the desired magnitude of effect one hopes to



(a) (b)

Figure 5. Panel (a) shows the empirical (dashed line) and theoretical (solid) power curves for Design 1, and different sample sizes. Panel (b) displays theoretical power curves corresponding to several sample sizes: 50, 100, 150, 200, 300, 400, 500 (from bottom to top).

detect with the test. The logic follows the typical sample size calculation techniques in classical regression models. Nevertheless, the complexity of the quantities involved in this calculation arises from the complex objects - random functions - that we deal with. To be more specific, one can use pilot studies to estimate this covariance surface and calculate the sample size for future, much larger studies depending on the magnitude of the effect.

7.3. *Partial functional linear model*

Next, we investigate the performance of the tests in a partial functional liner model setting that mimics the Diffusion Tensor Imaging data generation process, and we study the robustness of the results when the distribution of the errors is not Gaussian. In particular consider the case-study, where of interest is the association between the Paced Auditory Serial Addition Test score and the fractional anisotropy profiles along the corpus callosum tract in multiple sclerosis, while accounting for the gender and age of the patients; see Section 6.1. We analyze these data using the partial functional linear model approach discussed in Section 5; in the interest of space, the model components estimates are given in the Supplementary Material. We use these estimates to perform a simulation experiment for partial functional linear model.

The estimated eigenfunctions and eigenvalues, are used to obtain the generating process for the underlying functional covariates $\{X_i(t) : t \in [0, 93]\}$. The noisy observations W_{ij} corresponding to points $t_{ij} \in [0, 93]$ are obtained by contaminating $X_i(t_{ij})$ with Gaussian measurement error that has mean 0 and variance equal to the estimated variance of the noise in the study; it is assumed a regular dense design for t_{ij}'s. The additional covariates are taken as the gender and the centered and scaled age of the patients in the study. The response Y_i is generated from the partial functional linear model (6) for $\alpha = \tilde{\alpha}$, $\beta(t) = c\tilde{\beta}(t)$, where $c \geq 0$, $\tilde{\alpha}$ and $\tilde{\beta}(\cdot)$ are the estimated effects from the data analysis. The sample size is set to $n = 150$, the total number of patients in the application. Two settings for the distribution of the random noise ϵ_i are considered: (i) $\epsilon_i \sim N(0, 144)$, (ii) $\epsilon_i \sim \sqrt{48}t_3$, where the variance of the noise is equal to the estimated analogue in the application. The objective of this experiment is to study the performance of the four tests for testing the null hypothesis that $H_0 : \beta(\cdot) \equiv 0$.

The four tests are applied, as discussed in Section 2, where for consistency with the real data analysis, the number of functional principal components is selected using a

threshold level of 90% for the cumulative explained variance. Type I error is estimated based on 5000 simulations when data are generated under the assumption that $\beta(\cdot) \equiv 0$, and the power is estimated based on 1000 simulations when data are generated under the assumptions that $\beta(\cdot) = c\tilde{\beta}(\cdot)$ for $c > 0$, for various values of c.

Table 1 gives the results separately for the two models for the error distribution, when the significance level is 5%. Overall it appears that all the tests are robust to the model mispecification: both the Type I error rate and various powers of the tests seem to be similar under the two error distributions considered. Furthermore, the Type I error rates are close to the nominal level for the score and F tests, while they seem somewhat inflated for the Wald and the modified likelihood ratio tests. All the tests have comparable powers.

Table 1. Percentage of rejected tests at 5% significance level. The results are based on 5000 simulated data sets for Type I error and 1000 simulated data sets for power.

Model	Type of test	$c = 0$	0.2	0.4	0.6	0.8	1
Normal	Score	5.4	11.6	32.2	67.2	91.0	98.6
	Wald	5.8	12.1	33.1	67.9	91.3	98.8
	Likelihood Ratio	5.8	12.3	33.3	68.1	91.3	98.8
	F	5.1	11.2	31.4	66.7	90.8	98.6
t	Score	5.3	12.3	39.5	74.9	93.0	98.0
	Wald	5.7	12.8	40.2	75.6	93.5	98.0
	Likelihood Ratio	5.7	12.9	40.2	75.7	93.5	98.0
	F	5.1	11.9	39.1	74.5	92.8	97.9

Acknowledgement

A.-M. Staicu's research was supported by U.S. National Science Foundation grant numbers DMS 1007466 and DMS 0454942 and the U.S. National Institute of Health grants R01 NS085211 and R01 MH086633. A. Maity's research was supported by U.S. National Institute of Health grant R00ES017744. We thank Ciprian Crainiceanu, Daniel Reich, the National Multiple Sclerosis Society, and Peter Calabresi for the diffusion tensor imaging dataset.

Supplementary material

Supplementary material available online includes details of the estimation of the functional principal component scores, complete proofs of the two main theorems, the expressions of the testing procedures for partial functional linear model, and additional simulations.

References

Basser, P., Mattiello, J., and LeBihan, D. (1994), 'MR diffusion tensor spectroscopy and imaging', *Biophysical Journal*, 66, 259–267.

Basser, P., Pajevic, S., Pierpaoli, C., and Duda, J. (2000), 'In vivo fiber tractography using DT-MRI data', *Magnetic Resonance in Medicine*, 44, 625–632.

Cardot, H., Ferraty, F., and Sarda, P. (1999), 'Functional linear model', *Statistics & Probability Letters*, 45, 11–22. http://dx.doi.org/10.1016/S0167-7152(99)00036-X.

Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003), 'Testing hypotheses in the functional linear model', *Scandinavian Journal of Statistics*, 30, 241–255. http://dx.doi.org/10.1111/1467-9469.00329.

Cardot, H., Goia, A., and Sarda, P. (2004), 'Testing for No Effect in Functional Linear Regression Models, Some Computational Approaches', *Communications in Statistics - Simulation and Computation*, 33, 179–199.

Crainiceanu, C., (Coordinating authors), P.R., Goldsmith, J., Greven, S., Huang, L., and (Contributors), F.S. (2012), *refund: Regression with Functional Data*, http://CRAN.R-project.org/package=refund. R package version 0.1-5.

Goldsmith, A.J., Feder, J., Crainiceanu, C.M., Caffo, B., and Reich, D. (2011), 'Penalized Functional Regression', *Journal of Computational and Graphical Statistics*, 20, 830–851.

González-Manteiga, W., González-Rodríguez, G., Martínez-Calvo, A., and García-Portuguéss, E. (2014), 'Bootstrap independence test for functional linear models', *unpublished manuscript*.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010), 'Longitudinal functional principal component analysis', *Electronic Journal of Statistics*, 4, 1022–1054.

Hall, P., and Horowitz, J.L. (2007), 'Methodology and convergence rates for functional linear regression', *The Annals of Statistics*, 35, 70–91. http://dx.doi.org/10.1214/009053606000000957.

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

Hall, P., Müller, H.G., and Wang, J.L. (2006), 'Properties of principal component methods for functional and longitudinal data analysis', *The Annals of Statistics*, 34, 1493–1517. http://dx.doi.org/10.1214/009053606000000272.

Hilgert, N., Mas, A., and Verzelen, N. (2013), 'Minimax adaptive tests for the functional linear model', *The Annals of Statistics*, 41, 838–869.

Jank, W., and Shmueli, G. (2006), 'Functional data analysis in electronic commerce research', *Statistical Science*, 21, 155–166. http://dx.doi.org/10.1214/088342306000000132.

Kong, D., Xue, K., Yao, F., and Zhang, H.H. (2016), 'Partially functional linear regression in high dimensions', *Biometrika*, 103, 147–159. http://dx.doi.org/10.1093/biomet/asv062.

Li, Y., Wang, N., and Carroll, R.J. (2010), 'Generalized functional linear models with semiparametric single-index interactions', *Journal of the American Statistical Association*, 105, 621–633. http://dx.doi.org/10.1198/jasa.2010.tm09313. Supplementary materials available online.

Liu, B., and Müller, H.G. (2008), 'Functional data analysis for sparse auction data', in *Statistical methods in e-commerce research*, Statist. Practice, Hoboken, NJ: Wiley, pp. 269–289, http://dx.doi.org/10.1002/9780470315262.ch12.

McLean, M.W., Hooker, G., and Ruppert, D. (2014), 'Restricted Likelihood Ratio Tests for Linearity in Scalar-on-Function Regression.', *Statistics and Computing*, to appear.

Mclean, M.W., Hooker, G., Staicu, A.M., Scheipl, F., and Ruppert, D. (2014), 'Functional Generalized Additive Models', *Journal of Computational and Graphical Statistics*, 23, 249–269. http://dx.doi.org/10.1080/10618600.2012.729985.

Müller, H.G., and Stadtmüller, U. (2005), 'Generalized functional linear models', *The Annals of Statistics*, 33, 774–805. http://dx.doi.org/10.1214/009053604000001156.

Müller, H.G., and Wang, J.L. (2012), *PACE: Functional Data Analysis and Empirical Dynamics*, http://anson.ucdavis.edu/ mueller/data/pace.html. MATLAB package version 2.15.

Müller, H.G., Wu, Y., and Yao, F. (2013), 'Continuously additive models for nonlinear functional regression', *Biometrika*, 100, 607–622. http://dx.doi.org/10.1093/biomet/ast004.

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. http://links.jstor.org/sici?sici=0035-9246(1991)53:3¡539:STFFDA¿2.0.CO;2-Worigin=MSN. With discussion and a reply by the authors.

Ramsay, J.O., and Silverman, B.W. (2005), *Functional Data Analysis*, 2nd ed., Springer Series in Statistics, Springer.

Rao, C.R. (1948), 'Large sample tests of statistical hypotheses concerning several parameters with

applications to problems of estimation', *Mathematical Proceedings of the Cambridge Philosophical Society*, 44, 50–57. http://dx.doi.org/10.1017/S0305004100023987.

Shin, H. (2009), 'Partial functional linear regression', *Journal of Statistical Planning and Inference*, 139, 3405–3418. http://dx.doi.org/10.1016/j.jspi.2009.03.001.

Staicu, A.M., Crainiceanu, C.M., Ruppert, D., and Reich, D. (2012), 'Modeling functional data with spatially heterogeneous shape characteristics', *Biometrics*, 17, 331–343.

Su, D.C., Yu-Ru, and Hsu, L. (2016), 'Hypothesis testing in functional linear models', *works.bepress.com/di/22/download/*.

Swihart, B., Goldsmith, J., and Crainiceanu, C. (2014), 'Restricted Likelihood Ratio Tests for Functional Effects in the Functional Linear Model', *Technometrics*, p. to appear.

Verzelen, N., and Villers, F. (2010), 'Goodness-of-fit tests for high-dimensional Gaussian linear models', *The Annals of Statistics*, 38, 704–752. http://dx.doi.org/10.1214/08-AOS629.

Wang, S., Jank, W., and Shmueli, G. (2008), 'Explaining and forecasting online auction prices and their dynamics using functional data analysis', *Journal of Business & Economic Statistics*, 26, 144–160. http://dx.doi.org/10.1198/073500106000000477.

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. http://dx.doi.org/10.1198/016214504000001745.

Yao, F., Müller, H.G., and Wang, J.L. (2005b), 'Functional linear regression analysis for longitudinal data', *The Annals of Statistics*, 33, 2873–2903. http://dx.doi.org/10.1214/009053605000000660.

Zhang, J.T., and Chen, J. (2007), 'Statistical inferences for functional data', *The Annals of Statistics*, 35, 1052–1079. http://dx.doi.org/10.1214/009053606000001505.

Zhu, H., Yao, F., and Zhang, H.H. (2014), 'Structured functional additive regression in reproducing kernel Hilbert spaces', *Journal of the Royal Statistical Society. Series B.*, 76, 581–603. http://dx.doi.org/10.1111/rssb.12036.