Classical Testing in Functional Linear Models

Dehan Kong, Ana-Maria Staicu and Arnab Maity
Department of Statistics, North Carolina State University

Abstract

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 when the number of principal components diverges, and for both densely and sparsely observed functional covariates. 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 in simulation experiments and using two real data applications.

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

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 in different realistic scenarios, such as when the functional covariate is observed on a sparse irregularly spaced grid, and possibly with measurement error. We discuss four testing procedures, investigate their theoretical properties and study their finite sample performance 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 irregularly observed functional covariate situation; 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 pre-fixed basis systems, such as eigenbasis, 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 et al. (1999), Yao et al. (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. There are two main reasons to consider the problem of testing in the context of functional linear models to be of importance. 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 et al. (2003) discussed a testing procedure based on the norm of the cross covariance operator of the functional predictor and the scalar response. Later, Cardot et al. (2004) proposed two computational approaches by using a permutation and F tests. 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. 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 that the functional covariate is observed fully and without error. Also, 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. Recently, Swihart, Goldsmith and Crainiceanu (2013, unpublished manuscript) addressed a similar testing problem using likelihood ratio tests and restricted likelihood ratio tests and investigated their properties numerically, via simulation studies, but did not present their theoretical properties.

In this paper, we consider the situation where the functional predictor is observed either at densely set of points, or at sparsely, irregularly spaced grid, and possibly 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 simple linear 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 irregularly spaced and sparse observation of the the functional covariates, where smoothing of individual curves are practically impossible. Second, we can easily account for possible measurement errors in the functional observations. 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 et al. (2006), Hall and Hosseini-Nasab (2009), Zhang and Chen (2007), and Yao et al. (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 theoretical properties of our proposed testing procedures. In particular, we derive the null distributions of the test statistics under both dense and sparse irregularly spaced designs, and provide asymptotic theoretical alternative distributions under the dense design. 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 from a pilot or preliminary study, and is interested in determining the sample size of a future larger study within the same cohort. We extend our testing procedures to the partial functional linear model (Shin, 2009), where an additional vector valued covariate is observed and included in the model as a parametric term.

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 and power, for 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.

2 Methodology

2.1 Model specification

Suppose for $i = 1, \ldots, n$, we observe a real values 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 $W_{ij} = W_i(t_{ij})$ is a proxy observation of the true underlying process $X_i(\cdot)$, such that $W_i(t) = \eta(t) + X_i(t) + e_i(t)$, where $\eta(\cdot)$ is the mean function, and $e_i(\cdot)$ is a Gaussian process with mean 0 and covariance $\text{cov}\{e_i(t), e_i(s)\} = \sigma_e^2 I(t = s)$, where $I(t = s)$ is the indicator function which equals 1 if $t = s$, and 0 otherwise. Furthermore, it is assumed that the true process $X_i(\cdot) \in L^2(\mathcal{T})$ has mean 0 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. 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 or sparsely and with measurement error. We develop four testing procedures to test H_0, study their theoretical properties, and compare numerically their performances for dense as well as sparse designs of the functional covariate.

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 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 the pseudo-model

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

and the hypothesis testing problem can be reduced to

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

We consider four classical testing procedures, namely Wald, Score, likelihood ratio and F-test and examine their application in the context of the pseudo-model (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 $\{\xi_{ij}, 1 \leq i \leq n, 1 \leq j \leq s_n\}$, the pseudo log likelihood function from (3) can be written as

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

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 $\mathcal{I}_{22} = B^\top B/\sigma^2$. 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 defined as

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

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 we only need 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. 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 $n \log(\widetilde{\sigma}^2 / \widehat{\sigma}^2)$. Using the same argument as in Wald test, in this case also, we use 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 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), Zhang and Chen (2007). 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%. 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).

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 the next section, we show indeed that one can approximate the null distributions by the above traditional ones under linear model settings despite the fact that we truncate the number of functional principal component scores and plug in the estimates instead of the true scores.

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 it 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, likelihood ratio and F tests only; the asymptotic properties of the Wald test follow trivially from the results of Müller and Stadtmüller (2005).

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. We begin with introducing some notation. For any two random variables, H_n and G_n, where the subscript is to point their dependence on sample size n, define $H_n \hookrightarrow G_n$ if $P(H_n \leq x) - P(G_n \leq x) \to 0$, as $n \to \infty$. Moreover define $H_n \sim G_n$ if $P(H_n \leq x) = P(G_n \leq x)$. In the following we use T_S for the score statistic, T_L for the likelihood ratio test, and T_F for the F test statistic.

Theorem 1 *Assume model (1) holds. Then, if the null hypothesis, that $\beta(t) = 0$ for all t, is true, we have that: (i) $T_S \hookrightarrow \chi^2_{s_n}$, (ii) $T_L \hookrightarrow \chi^2_{s_n}$, and (iii) $T_F \sim F_{s_n, n-s_n-1}$.*

The assumptions required by Theorem 1 are mild and require $X_i \in L^2(\mathcal{T})$ and $s_n < n$. 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 if the estimated functional principal component scores are used instead of the true ones. Thus, 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. In particular, for fixed truncation value s_n, the null distribution of the F test statistic is exactly $F_{s_n, n-s_n-1}$ and the null distribution of the score test and the likelihood ratio test statistic is $\chi^2_{s_n}$.

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). However, when the design is sparse, the estimation of the functional principal component scores has a lower performance; for example the estimators of the scores have a convergence rate of order $o_P(1)$ (Yao et al., 2005a). Thus the asymptotic distribution of the tests under alternative is different, and the results are far from obvious. In the sparse case we investigate the alternative distribution of the tests only empirically, via numerical simulation.

We begin with describing the assumptions required by our theoretical developments. Throughout this section, let $\mu_i = E\{Y_i \mid X_i(\cdot)\} = \int X_i(t)\beta_a(t)dt$, $\mu = (\mu_1, \ldots, \mu_n)^\top$, and, with a slight abuse of notation, let C denote a generic constant term.

(A) The number of principal components selected, s_n, satisfies $\lambda_{s_n}^{-4} s_n^3 \delta_{s_n}^{-1} n^{-1/2} = o(1)$, where λ_{s_n} is the smallest eigenvalue and δ_{s_n} is the smallest spacing between any two adjacent eigenvalues λ_j and λ_{j+1} for $1 \leq j \leq s_n$.

Condition (A) concerns the divergence of the number of functional principal component with n. Specifically, it is assumed that this divergence also depends on the smallest eigenvalue and the spacing between adjacent eigenvalues. In particular, when the true number of functional principal components is assumed finite Li, Wang, and Carroll (2010), then this condition is met. 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 $s_n^{10\alpha+8} = o(n)$. 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 each integer $r \geq 1$, $\lambda_j^{-r} E(\int_{\mathcal{T}} [X_i(t) - E\{X_i(t)\}]\phi_j(t)dt)^{2r}$ is bounded uni-

formly in j.
(B3) Let $\widetilde{X}_i(\cdot)$ be the centered version of $X_i(\cdot)$ to have null mean function, i.e. $\widetilde{X}_i(t) = X_i(t) - E\{X_i(t)\}$. Assume $R(t_1, t_2, t_3, t_4) = E\{\widetilde{X}_i(t_1)\widetilde{X}_i(t_2)\widetilde{X}_i(t_3)\widetilde{X}_i(t_4)\} - K(t_1, t_2)K(t_3, t_4)$ exists and is finite, for $t_1, t_2, t_3, t_4 \in \mathcal{T}$.
Conditions (B1)-(B3) are common in functional data analysis; see Hall and Hosseini-Nasab (2006) and Li et al. (2010). For example, (B1) and (B2) are met when we have a Gaussian process with Hölder continuous sample paths (Hall and Hosseini-Nasab, 2006).

(C1) The observed time points t_{ik} are independent identically distributed random design points with density function $g(\cdot)$, where g is bounded away from 0 on $\mathcal{T}$ and is continuously differentiable.
(C2) $\max_{2\leq k\leq m_i} \{t_{ik} - t_{i(k-1)}\} = O(m^{-1})$, where $m = \min_i m_i$.
(C3) $m \geq Cn^{\kappa}$, with $\kappa > 5/4$.
(C4) $\sum_{j=1}^{\infty} \lambda_j \beta_{ja}^2 < \infty$.
Conditions (C1)-(C4) regards the sampling design and the regression parameter $\beta(\cdot)$. In particular, (C2) and (C3) are standard for a regular dense design; see for example Li et al. (2010).Condition (C4) is mild; for example it suffice to have $\int E\{X_i^2(t)\}dt < \infty$ and $|| \beta_a(\cdot) ||< \infty$ in order for (C4) to hold.

The following result presents the asymptotic distribution of the score test statistic, T_S, the likelihood ratio test, T_L, and the F test statistic, T_F, under the alternative hypothesis. The results are restricted to a dense sampling design.

Theorem 2 *Assume model (1) holds. Furthermore assume the conditions (A),(B1)-(B3),(C1)-(C4) are met. Then under the assumption that $H_a : \beta(\cdot) = \beta_a(\cdot)$ is true, we have: (i) $T_S \hookrightarrow \chi^2_{s_n}(\Lambda_n)$, (ii) $T_L \hookrightarrow \chi^2_{s_n}(\Lambda_n)$, and (iii) $T_F \hookrightarrow F_{s_n,n-s_n-1}(\Lambda_n)$, where $\Lambda_n = n\vartheta(1 + o(1))$ with $\vartheta = \int \beta_a(t_1)\beta_a(t_2)K(t_1, t_2)dt_1dt_2$.*

The proof is included in the Supplementary Material. Our theoretical development uses the approach of "smoothing first, then estimation" described in Zhang and Chen (2007), where each noisy trajectory is first smoothed individually, using local polynomial kernel smoothing with a global bandwidth. It is assumed that the kernel bandwidth h satisfies $h = O(n^{-\kappa/5})$, where n is the sample size and κ is specified in (C3); see also Li et al. (2010).

Corollary 1 *Theorem 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 $\Lambda_a = \int \beta_a(t_1)\beta_a(t_2)K(t_1, t_2)dt_1dt_2$. 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 β_a 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 $K(\cdot,\cdot)$ can be estimated from prior studies. Section 6.2 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.*

4 Extension to partial functional linear regression

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 5.1 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 et al. (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 pseudo-model, 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 pseudo 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.

In practice the tests 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. The asymptotic distribution of these test statistics under the null hypothesis that $\beta(\cdot) \equiv 0$ can be easily derived following similar arguments to Theorem 1, irrespective of the sampling design for the functional covariates. Specifically, the null distribution of T_W, T_S and T_L is $\chi^2_{s_n}$, while the null distribution of T_F is $F_{s_n, n-s_n-(p+1)}$, where the degrees of freedom are changed from (1) to account for the dimension of the nuisance parameter.

5 Real data application

5.1 The Diffusion Tensor Imaging data

Consider our motivating application, the Diffusion Tensor Imaging (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 Goldsmith et al. (2011); Greven et al. (2010); Staicu et al. (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 et al., 1994, 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 full 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 either corpus callosum tract or the left corticospinal tracts tract, 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 parameter of interest, describing a linear association between the 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 5 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. Then we test whether the coefficient function $\beta(\cdot)$ is zero along these directions, by accounting for age and gender effects using the methods discussed in Section 2.2. The p-value 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 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}.

Next, consider the left corticospinal tracts tract, and investigate the association between the true fractional anisotropy along this tract and the cognitive disability as measured by the Paced Auditory Serial Addition Test score. Fractional anisotropy is measured at 55 locations along the corticospinal tracts tract; the missingness along this tract is notably larger than along the corpus callosum tract. We assume a similar partial functional linear model to relate the underlying smooth fractional anisotropy along the corticospinal tracts tract and the Paced Auditory Serial Addition Test performance and test for no relationship between them. As before, we first apply functional principal component analysis, 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 we obtain a p-value of 0.0285 using F test (0.0233 with likelihood ratio test, 0.0223 using Wald and 0.0293 with score test statistic). The results show that there is 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%.

Overall, our findings corroborate the specialists prior expectations that the cognitive function is associated with the corpus callosum tract, as well as point out surprising association of the cognitive function with the corticospinal tract. Interestingly, both findings are in agreement with Swihart et al. (2013), 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.

5.2 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 (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

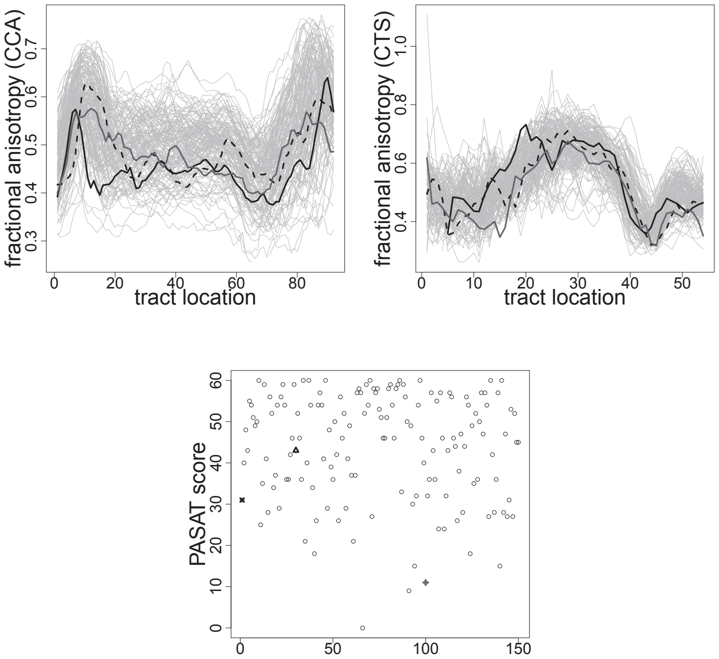


Figure 1: Fractional anisotropy profiles along corpus callosum (left) and corticospinal tracts (middle) and the associated Paced Auditory Serial Addition Test scores (right panel) in the group of multiple sclerosis patients. Depicted in different colors and line/symbols styles are the measurements of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

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 (see (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. Then we perform the test statistics: the p-value reported by the F statistic equals 5.5×10^{-4} indicating very strong evidence of association. This result is similar for the other testing procedures: the 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}.

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 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, 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.

6 Simulation study

The performance of the Wald, score, 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 6.1). 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 6.2). Finally, we consider a partial functional 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 6.3).

6.1 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 ϵ_i 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

$$\beta_c(t) \quad = \quad c\{1 + \exp(1 - 0.1t)\}^{-1}, \tag{7}$$

where $c \geq$ 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 12 equally spaced points in $[0.02, 0.1]$.

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/$\sim$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.

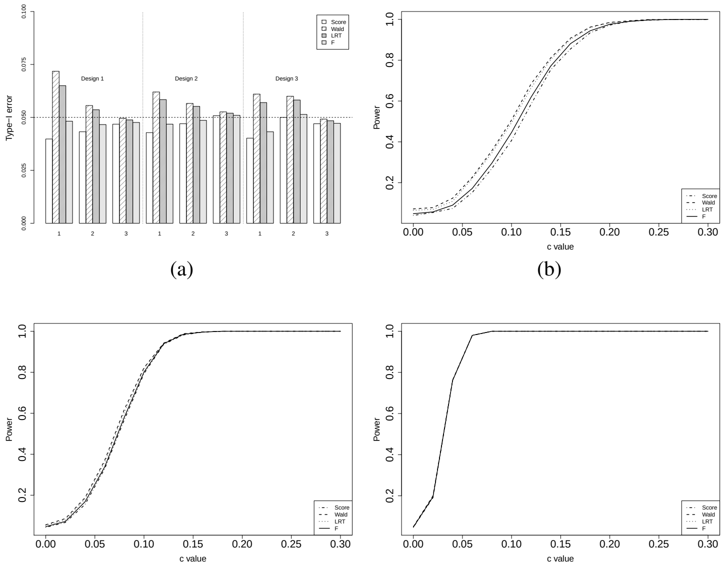
The results are presented in Figure 6.1, and correspond to fixing the level of significance at 5%. Figure 6.1 (a) shows the performance of the tests with respect to Type I error rate for various sampling designs and as the sample size increases from 50 to 500. In particular, F test gives reasonable type-I 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 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 6.1 (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.

6.2 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 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 $n\Lambda_c$, $F^*_{\alpha,s,n-s-1}$ denotes



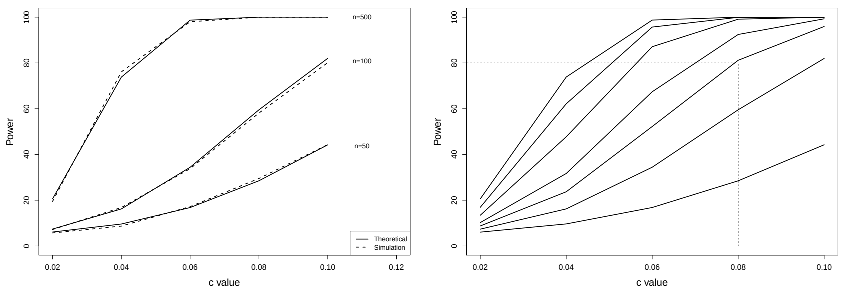
(a)

(b)

(c)

(d)

Figure 2: 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.



(a) (b)

Figure 3: 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).

the critical value corresponding to right tail probability of α under $F_{s,n-s-1}(0)$, and $\Lambda_c = \int \beta_c(t_1)\beta_c(t_2)K(t_1,t_2)dt_1dt_2$.

Figure 6.2 (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 6.1 panels (b)-(d) and restricted to the domain $(0, 0.10]$. 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 6.2 (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. For the Wald, score, and likelihood ratio tests, close agreement between the asymptotic and empirical power approximations occurs when the sample size is large. Because of these considerations, our recommendation is to use F test for sample size calculations.

6.3 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 5.1. We analyze these data using the partial functional linear model approach discussed in Section 4; 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 likelihood ratio tests. All the tests have comparable powers.

Acknowledgement

A.-M. Staicu's research was supported by U.S. National Science Foundation grant number DMS 1007466. A. Maity's research was supported by National Institute of

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

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 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., 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 & Probability Letters*, 45, 11–22.

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*, 30.

Crainiceanu, C., (Coordinating authors), P. R., Goldsmith, J., Greven, S., Huang, L., and (Contributors), F. S. (2012), *refund: Regression with Functional Data*, 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*, to appear.

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.

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.

— (2009), "Theory for high-order bounds in functional principal components analysis," *Mathematical Proceedings of the Cambridge Philosophical Society*, 146, 225–256.

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.

Jank, W. and Shmueli, G. (2006), "Functional data analysis in electronic commerce research," *Statistical Science*, 21, 155–166.

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, 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*, Hoboken, NJ: Wiley, Statist. Practice, pp. 269–289.

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

Müller, H.-G. and Wang, J.-L. (2012), *PACE: Functional Data Analysis and Empirical Dynamics*, mATLAB package version 2.15.

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, with discussion and a reply by the authors.

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

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.

Shin, H. (2009), "Partial functional linear regression," *Journal of Statistical Planning and Inference*, 139, 3405–3418.

Staicu, A.-M., Crainiceanu, C. M., Ruppert, D., and Reich, D. (2011), “Modeling functional data with spatially heterogeneous shape characteristics,” *Technical report*.

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.

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.

— (2005b), “Functional linear regression analysis for longitudinal data,” *The Annals of Statistics*, 33, 2873–2903.

Zhang, J.-T. and Chen, J. (2007), “Statistical inferences for functional data,” *The Annals of Statistics*, 35, 1052–1079.