# Restricted likelihood ratio tests for functional effects in the functional linear model

Bruce J. Swihart*
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
and
Jeff Goldsmith
Department of Biostatistics, Columbia Mailman School of Public Health
and
Ciprian M. Crainiceanu
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

November 1, 2013

**Abstract**

The goal of our article is to provide a transparent, robust, and computationally feasible statistical approach for testing in the context of scalar-on-function linear regression models. Assuming linearity between response and predictors, we are interested in testing for the necessity of functional effects. Our methods are motivated by and applied to a large longitudinal study involving diffusion tensor imaging of intracranial white matter tracts in a susceptible cohort. In the context of this study, we conduct hypothesis tests that are motivated by anatomical knowledge and support recent findings regarding the relationship between cognitive impairment and white matter demyelination. R-code and data are in the examples of `refund::rlrt.pfr()`.

*Keywords:* Functional data analysis; Nonparametric smoothing; Nonregular problem; Penalized splines; Variance components.

technometrics tex template (do not remove)

*

# 1 Introduction

Vast increases in the ability to collect and store functional data have contributed to a proliferation of approaches for regression models involving functions as predictors. There are now many competing methods for parameter estimation in the functional linear model (FLM). However, inferential techniques for these models are less advanced than the estimation procedures. In this manuscript we seek a statistically principled approach to address whether a functional predictor should be included in a functional linear regression model for scalar response (equivalently: scalar-on-function regression model), as detailed in Example 3 of Chapter 1 in Ferraty and Vieu (2006) and chapter 15 of Ramsay and Silverman (2005). Our contribution is two likelihood-based tests that correspond to parametric and semi-parametric models parameterized by coefficient functions. One test has the null of the coefficient function being exactly zero, which tests the "lack of effect in the functional linear model"; the other has the null of a constant not necessarily zero. This test for the "lack of effect in the functional linear model" has been considered nonparametrically (Chapter 9, Horváth and Kokoszka (2012)) and through permutation $F$-tests (Chapter 9, Ramsay et al. (2009)).

Our first approach is related to the standard parametric functional principal components regression (FPCR) method and uses standard likelihood ratio tests for the functional coefficient. Next we modify a penalized approach that casts the FLM in a mixed effects semi-parametric framework and derive likelihood ratio test statistics for variance components that restrict the coefficient function to be a constant under the null hypothesis.

We observe data of the form $\{Y_i, \boldsymbol{X}_i, W_i(t)\}$ for subjects $i \in \{1, \ldots, I\}$, where $Y_i$ is a continuous scalar outcome of interest, $\boldsymbol{X}_i$ are non-functional covariates and $W_i(t)$ for $t \in [0, 1]$ is a functional predictor. The FLM model for data of this form is

$$\mathrm{E}[Y_i] = \alpha + \boldsymbol{X}_i\beta + \int_0^1 W_i(s)\gamma(s)ds \tag{1}$$

where $\gamma(t)$ is a coefficient function that weights the functional predictor $W_i(t)$ to appropriately emphasize portions of the curve in the functional contribution (as represented by the integral) assimilated into the model for scalar outcome $Y_i$. Multiple functional predictors can be easily considered in an appropriately extended model. Increasingly, studies relating

functional exposures to scalar outcomes are longitudinal, where over visits $j = 1, \ldots, J$, functional exposures $W_{ij}(t)$ are measured along with outcome $Y_{ij}$. In accommodation of these repeated measures and the consequent clustering, a further extension includes a subject-specific random effect to account for correlation between repeated scalar observations $Y_{ij}$ across multiple visits.

In testing for functional effects, the following questions have important statistical and scientific considerations. The framework for answering the questions can be facilitated by a hypothesis test comparing a null model $H_0$ to a richer, alternative model $H_A$.

- **Test of functional form: Is functional structure needed?**

  A direct and important question in the context of the FLM is whether the functional structure of observations $W_i(t)$ is needed to explain association with the outcome, or if a simpler summary of these curves suffices. We therefore wish to test

$$\begin{aligned} H_0 &: \mathrm{E}[Y_i] = \alpha + \boldsymbol{X}_i\beta + \overline{W}_i\beta_W \\ H_A &: \mathrm{E}[Y_i] = \alpha + \boldsymbol{X}_i\beta + \int_0^1 W_i(s)\gamma(s)ds \end{aligned} \tag{2}$$

  where $\overline{W}_i = \int_0^1 W_i(s)ds$. Rejecting $H_0$ in favor of $H_A$ would indicate that honoring the functional structure of $W_i(t)$ is worthwhile, while failing to reject would indicate that the mean conveys all the information that a function has related to the outcome. An equivalent null hypothesis would be $\gamma(s) = c$ for some constant $c$.

- **Test of inclusion: Does a functional predictor improve the model?**

  In the now-common context of multiple functional predictors, a reasonable question to ask is which (if any) of the predictors are related to the outcome, as in

$$\begin{aligned} H_0 &: \mathrm{E}[Y_i] = \alpha + \boldsymbol{X}_i\beta + \int_0^1 W_{i1}(s)\gamma_1(s)ds \\ H_A &: \mathrm{E}[Y_i] = \alpha + \boldsymbol{X}_i\beta + \int_0^1 W_{i1}(s)\gamma_1(s)ds + \int_0^1 W_{i2}(s)\gamma_2(s)ds. \end{aligned} \tag{3}$$

  Rejecting $H_0$ in favor of $H_A$ would indicate that including and modeling the functional structure of $W_{i2}(t)$ is worthwhile in a model selection sense. Investigations into potential functional confounding, that is the perturbations in the shape of $\gamma_1(t)$ in

the presence of $W_{i2}(t)$ (under $H_A$) relative to the absence of $W_{i2}(t)$ (under $H_0$), are also possible in this construction. An equivalent null hypothesis would be $\gamma_2(s) = 0$.

The distinguishing feature in fitting FLMs is modeling the integral in the predictor, which involves appropriately aggregating and representing the subject-specific $W_i(t)$ so that a meaningful, shared weighting function $\gamma(t)$ can be estimated. There are many ways to do this, two of which are related to the methods we propose: the widely-used functional principal components regression (FPCR) described in Ramsay and Silverman (2005) and the penalized functional regression (PFR) approach of Goldsmith et al. (2011). Traditional FPCR projects functional observations onto a low-dimensional functional principal components basis and uses scores as predictors in a standard regression model. The PFR approach uses a flexible spline basis to express the functional coefficients and induces smoothness through penalization in a mixed model framework. Here we introduce modifications to both techniques so that the coefficient function can be reduced to a non-zero constant (corresponding to the null hypothesis in (2)). Using these modifications, we develop testing procedures that address the statistical questions described above for both approaches.

Our motivation for developing this technique is a study that was seeking to relate the Paced Auditory Serial Addition Test score (PASAT) (Gronwall, 1977) to the microstructure of intracranial white matter in the the corpus callosum (CCA) and the right corticospinal tract (RCST) in multiple sclerosis patients (Figure 1). Naturally, a scalar-on-function regression method fitting a FLM would be appropriate: the PASAT is scalar and the CCA and RCST tracts are functional, comprised of water diffusivity metrics from diffusion tensor imaging on a dense, regular grid of 93 and 55 points, respectively. Subjects were seen multiple times (2 to 8) and PASAT scores and tract images were collected each visit, longitudinally. Even though the sampling was functional, whether the effect was functional – with some locations being hotspots for relating to PASAT or if the whole-tract contributes to an overall average effect regardless of location – was scientifically unconfirmed. Unprecedentedly, test (2) allows this scientific question to be answered in a statistically principled way. The test is scientifically motivated to guide model selection with evidence from the data as opposed to choice of convenience ("simple linear regression on averaged functions are easier to fit and explain") or heuristics ("the $\beta(t)$ coefficient function from a FLM is

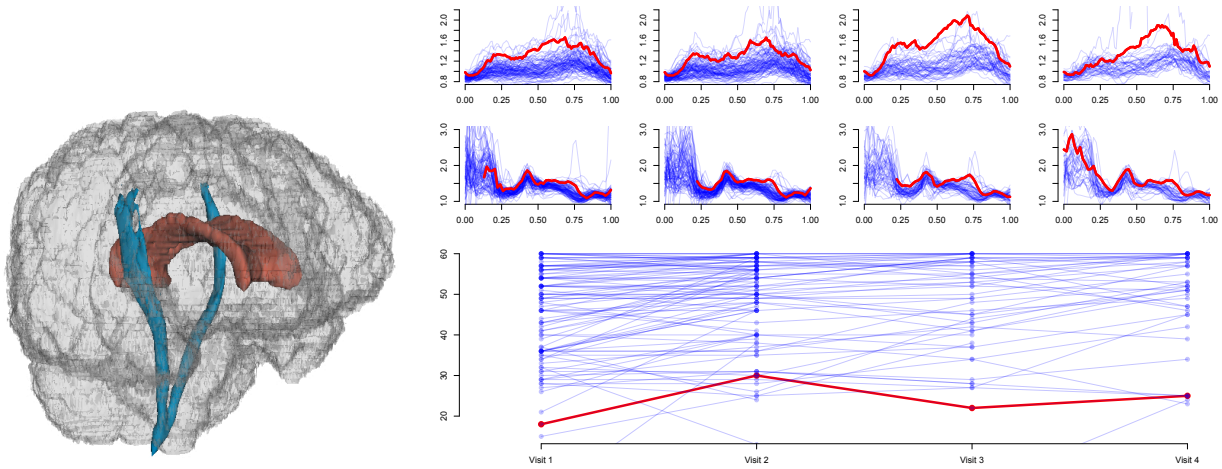


Figure 1: The left panel displays the anatomical structures interest: the corpus callosum is shown in red and the corticospinal tracts in blue. On the right we show the observed data: in the top four plots are observed corpus callosum tract profiles across four visits; in the middle four plots are observed right corticospinal tract profiles across four visits; in the bottom panel is a plot of the longitudinally observed cognitive function outcomes. In all plots, the data observed for a single subject is highlighted in red. Corpus callosum profiles are parameterized anterior to posterior. Corticospinal profiles are parameterized inferior to superior. This figure appears in color in the electronic version of this article.

so dynamic and the confidence bands exclude zero at some locations – it must be significant!"). Test (3) allows a slightly different question to be answered in that considering two functions simultaneously, is one not significantly adding to the fit. In our case, the CCA is a tract associated with cognitive function (connecting the hemispheres of the brain) and RCST is mainly involved in motor functioning. Being that PASAT is a cognitive test, do both a cognitive tract and motor function tract contribute to explaining PASAT variation, or is solely the cognitive track sufficient? Again, this is a scientific question that motivated our statistical method development, in turn, the development will enable and facilitate scientific discovery in a statistically principled way.

The functional data literature contains a rich collection of methods for estimating scalar-on-function regression models. The following is intended as an overview of functional regression methods and is not exhaustive. FPCR, described above, was an early approach; later extensions of this basic method imposed explicit penalties on the roughness of the co-

efficient function (Reiss and Ogden, 2007). A similar collection of techniques uses functional partial least squares in place of principal components (Goutis and Fearn, 1996; Reiss and Ogden, 2007). Several penalized spline approaches distinct from PFR have been proposed (Cardot et al. (2003); Marx and Eilers (1999), see Ferraty et al. (2011) for a discussion on presmoothing methods). Extensions of the FLM to allow nonlinear effects of functional contributions, similar to generalized additive models, are described in (James and Silverman, 2005; McLean et al., 2012), and the adaptation of single-index regression to functional predictors is described in (Eilers et al., 2009). Specifically, nonparametric and semi-parametric extensions have proliferated for this purpose – for an overview, consult Ferraty and Vieu (2006) (Chen et al., 2011; Ferraty et al., 2012; Aneiros-Pérez and Vieu, 2008; Delsol et al., 2011). For coherency and focus, herein we assume linear effects of functional contributions and assume that an interpretable coefficient function is of interest.

Despite the body of work related to estimation for scalar-on-function regression, there is relatively little work related to inference for coefficient function estimates. Most notably and recently is a lasso-type functional variable selection based on prediction performance (Gertheiss et al., 2013). Confidence intervals for functional coefficients in a low-dimension approach to the FLM have been derived (Müller and Stadtmüller, 2005). For penalized approaches, bootstrap confidence intervals have been developed (Reiss and Ogden, 2007; James et al., 2009) and the mixed model framework to construct model-based confidence intervals have been utilized (Goldsmith et al., 2011). Cardot et al. (2003) develop tests based on the covariance of the scalar outcome and functional predictor, but do not extend these tests to consider multiple predictors or longitudinal settings.

Two past approaches have commented on the potential for hypothesis tests in the FLM through the use of tests for zero variance components, but neither fully developed a method or study the properties of a hypothesis test (Reiss and Ogden, 2010; Gertheiss et al., 2012) (the former in the context of scalar-on-image regression). The theory for classical testing for $H_0 : \gamma(s) = 0$ in FLM has been developed, but still lacks computational implementability and the ability to test other null hypotheses Kong et al. (2013). Tests for zero variance components, as proposed herein, are readily implementable and allow for testing a variety of null hypotheses.

Tests for zero variance components have been used in the penalized spline literature. Penalized-spline additive models are a well-documented semiparametric method enabling scatterplot smoothing and can be represented and fitted as a mixed model (Ruppert et al., 2003; Marx and Eilers, 1998; Aerts et al., 2002; Crainiceanu et al., 2005; Wand, 2003; Ngo and Wand, 2004). Likelihood ratio tests (LRTs) and restricted likelihood ratio tests (RLRTs) have been theoretically developed and computationally implemented to test the necessity of the splines against an embedded polynomial regression (Crainiceanu and Ruppert, 2004; Greven et al., 2008; Scheipl et al., 2008), as have score tests (Verbeke and Molenberghs, 2003; Molenberghs and Verbeke, 2007; Zhang and Lin, 2003; Tzeng and Zhang, 2007; Zhang and Lin, 2008) and a Wald-type test (Wood, 2012). Because (R)LRTs and score tests are asympotically equivalent selecting between the two may be a matter of practicality; however score tests may require numerical optimization techniques for infimum calculations while (R)LRTs require a comparison of null and alternative models. Verbeke and Molenberghs (2003) note that practicing statisticians have accessible software to fit and compare a variety of models containing several variance components which may allow (R)LRTs to be more widely employed. Scheipl et al. (2008) demonstrate (R)LRTs outperforming Wald-type tests in situations of several variance components. We therefore solely focus on (R)LRTs herein.

The (R)LRTs for testing functional predictors in models with multiple functional predictors can justify models with multivariable effects (averaged functional effects) and functional effects, an important parsimony given the complexity of functional datasets. In a similar vein of dimension reduction, but not as extreme, Ferraty et al. (2010) suggest a way of reducing the very high dimension of a functional predictor to a low number of dimensions chosen so as to give the best predictive performance. In James et al. (2009), the authors impose a shrinkage penalty which results in coefficient function estimates containing regions equal to zero. A point-impact model for scalar-on-function regression in which one (or a few) unknown locations in the function domain affect the outcome has been proposed for binary outcomes (Lindquist and McKeague, 2009).

We develop approaches to the functional linear model for the problem of hypothesis testing in Section 2. We briefly overview 0-variance component (R)LRT testing method-

ologies in Section 3. Sections 4 through 6 contain a simulation study, data application and concluding remarks, respectively.

# 2 Techniques for Scalar-on-Function Regression

In this section we present two unique approaches to estimation in the FLM given in equation (1). The first is based on the widely-used FPCR approach, and the second on the more recent PFR method. Both approaches are presented to facilitate testing under the null hypothesis of a constant coefficient function, although other parametric forms for the coefficient under the null are easily considered.

## 2.1 FPCR

Functional principal components regression (FPCR) uses a low-dimensional principal component basis to express both the predictors and the coefficient function. Here we modify this approach to separate a constant and a functional effect.

First, an FPC decomposition is estimated from the observed curves. Briefly, define the covariance operator $\Sigma^{\boldsymbol{W}}(s,s') = \text{Cov}[W_i(s), W_i(s')]$ and let $\sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(s')$ be the spectral decomposition of $\Sigma^{\boldsymbol{W}}(s,s')$. Here $\boldsymbol{\psi}(s) = \{\psi_k(s) : k \in \mathbb{Z}^+\}$ are orthonormal eigenfunctions and $\lambda_1 \geq \lambda_2 \geq \ldots$ are the corresponding non-increasing eigenvalues. In practice, functions are observed on a dense (or sparse at the subject level) grid and possibly with measurement error. To account for this, we estimate $\Sigma^{\boldsymbol{W}}(s,s')$ using a method-of-moments approach and smooth the off-diagonal elements of this estimated covariance matrix to remove the effect of measurement error (Staniswalis and Lee, 1998; Yao et al., 2003). A truncated Karhunen-Loève approximation for $W_i(s)$ is $W_i(s) = \mu(s) + \sum_{l=1}^{K_w} c_{ik}\psi_k(s)$, where $K_w$ is the truncation lag, the $c_{ik} = \int_0^1 \{W_i(s) - \mu(s)\}\psi_k(s)ds$ are uncorrelated random variables with mean 0 and variance $\lambda_k$, and $\mu(s) = \text{E}[W(s)]$. The scores $\boldsymbol{c}_i$ are estimated either through numeric integration or as random effects in a mixed model (Xiao et al., 2013; Crainiceanu et al., 2009; Di et al., 2009; Yao et al., 2005). The choice of $K_w$ can be guided by the proportion of variability explained by each component or the leveling of the loglikelihood for increasing $K_w$ (James et al., 2000).

We express the coefficient function using the basis $\boldsymbol{\phi}(s) = \{\phi_1(s), \ldots, \phi_{K_g}(s)\}$; that is, we let

$$\gamma(s) = \boldsymbol{\phi}(s)\boldsymbol{\gamma} = \sum_{k=1}^{K_g} \gamma_k \phi_k(s)$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_{K_g}\}^T$. In a departure from standard functional principal components regression, we include curve means $\bar{W}_i$ as scalar covariates and expand centered curves using $[W_i(s) - \bar{W}_i] = \sum_{l=1}^{K_w} c_{ik}^* \psi_k(s)$ where $c_{ik}^* = \int_0^1 \left\{W_i(s) - \bar{W}_i\right\} \psi_k(s) ds$. This formulation, suggested by a reviewer, allows the separation of the constant and functional contributions using

$$\int_0^1 W_i(s)\gamma(s)ds = \bar{W}_i\gamma_0 + \int_0^1 \boldsymbol{c}_i^{*'}\boldsymbol{\psi}^T(s)\boldsymbol{\phi}(s)\boldsymbol{\gamma}ds = \bar{W}_i\gamma_0 + \boldsymbol{c}_i^{*'}\boldsymbol{\gamma}.$$

Doing so facilitates testing for constancy by modeling deviations from a constant coefficient function using the parameter vector $\boldsymbol{\gamma}$.

Next we pose the FLM as a standard linear model. Let $\boldsymbol{C}$ be the row-stacking of $\boldsymbol{c}_i^{*'}$, $\boldsymbol{X} = [1\ X\ \boldsymbol{C}]$ be the matrix consisting of non-functional covariates (including the curve-specific means $\bar{W}_i$) and the matrix $\boldsymbol{C}$, and $\boldsymbol{\beta}^T = [\alpha, \beta, \gamma_0, \boldsymbol{\gamma}]$ be the vector of coefficients. The FLM can be written

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y}] &= X\beta + \int_0^1 W(s)\gamma(s)ds \\ &= \boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

and the parameters can be estimated using standard least squares. We are particularly interested in the coefficients $\{\gamma_1, \ldots, \gamma_{K_g}\}$ which model deviations from a constant function. Note the parameter $K_g$ acts as a tuning parameter to control smoothness in $\gamma(s)$, and is typically chosen to be relatively small. The choice can be quite influential and is probably best guided by a cross-validated approach (James et al., 2000; Ruppert, 2002).

This formulation contrasts with the standard FPCR in in the separation of constant and functional effects for observed curves $W_i(s)$. Doing so allows for testing the constancy of the functional coefficient as in (2): one must only perform a test of the hypothesis $\gamma_1 = \gamma_2 = \ldots = \gamma_{K_g} = 0$ using a standard likelihood ratio test. Additionally, one can test for inclusion of the functional predictor (with constant coefficient or varying) by testing $\gamma_0 = \gamma_1 = \gamma_2 = \ldots = \gamma_{K_g} = 0$, which is again a standard LRT. Briefly, to test including $W_2$

in the presence of $W_1$, as in test (3), each would be decomposed separately with possibly distinct $K_g'$ and $\gamma_2(s) = \gamma_{20} = \gamma_{21} = \gamma_{22} = \ldots = \gamma_{2K_g'} = 0$ would be tested in a standard LRT. Optimally selecting $K_w$ and $K_g$ for FPCR is an open problem in functional data analysis. The next method, PFR, is much less sensitive to these choices as one can set each moderately large.

## 2.2 PFR

Alternatively to FPCR, PFR (Goldsmith et al., 2011) uses a large number of FPC basis functions to expand the predictors and a flexible spline basis for the coefficient function. Smoothness in the coefficient function is imposed using a mixed model construction. The PFR method allows for a range of basis function and penalty specifications, and here we construct a basis that reduces to a constant function under certain conditions.

As above, the functional predictor $W_i(s)$ is expressed (and estimated) using principal components decomposition with basis functions $\boldsymbol{\psi}(s)$ and scores $\boldsymbol{c}_i$, so that $W_i(s) = \mu(s) + \sum_{k=1}^{K_w} c_{ik}\psi_k(s)$. Next, the functional coefficient $\gamma(s)$ is expressed in terms of a flexible spline basis $\boldsymbol{\phi}(s) = \{\phi_0(s), \phi_1(s), \ldots, \phi_{K_g}(s)\}$. Here we take $\boldsymbol{\phi}(s)$ to be a B-spline basis in which $\phi_0(s) = 1$ and $\{\phi_1(s), \ldots, \phi_{K_g}(s)\}$ model deviations from a constant. Thus

$$\gamma(s) = \boldsymbol{\phi}(s)\boldsymbol{g} = \gamma_0 + \sum_{k=1}^{K_g} g_k\phi_k(s)$$

where $\boldsymbol{g} = \{\gamma_0, g_1, \ldots, g_{K_g}\}^T$. Smoothness is induced via a mixed-effects model treating $\{g_k\}_{k=1}^{K_g}$ as random effects shared across individuals (in keeping with standard notation, we use $g_k$ here in place of $\gamma_k$ as in §2.1 to emphasize the distinction between random and fixed effects). We use a modified first order random walk prior on the vector $\{g_k\}_{k=1}^{K_g}$ (Carter and Kohn, 1994; Hastie and Tibshirani, 2000; Fahrmeir and Lang, 2001a,b; Lang and Brezger, 2004; Goldsmith et al., 2011). That is, we assume $g_l \sim \mathrm{N}\left[g_{l-1}, \sigma^2_{\boldsymbol{g}}\right]$ for $2 \leq l \leq K_g$ and let $g_1 \sim \mathrm{N}\left[0, \sigma^2_{\boldsymbol{g}}\right]$. Using these expressions for the predictor and coefficient functions, the functional contribution for subject $i$ is

$$\int_0^1 W_i(s)\gamma(s)ds = a + \int_0^1 \boldsymbol{c}_i'\boldsymbol{\psi}^T(s)\boldsymbol{\phi}(s)\boldsymbol{g}ds = a + \boldsymbol{c}_i'\boldsymbol{M}\boldsymbol{g}$$

where the $(m,n)^{th}$ element of $\boldsymbol{M}$ is $\int_0^1 \psi_m(s)\phi_n(s)ds$ and $a = \int_0^1 \mu(s)\gamma(s)ds$ which is incorporated into the overall intercept $\alpha$.

We pose the FLM as a standard linear mixed effects model. Let $\boldsymbol{C}$ be the row-stacking of $\boldsymbol{c}_i'$, $\boldsymbol{X} = [1\ X\ (\boldsymbol{CM})^{[,1]}]$ be the matrix consisting of non-functional covariates and the first column of $\boldsymbol{CM}$, $\boldsymbol{Z} = [(\boldsymbol{CM})^{[,2:K_g]}]$ be the matrix consisting of the remaining columns of $\boldsymbol{CM}$, $\boldsymbol{\beta}^T = [\alpha, \beta, \gamma_0]$ be the fixed effects vector and $\boldsymbol{u}^T = \{g_k\}_{k=1}^{K_g}$ the random effects vector. The FLM can be written

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{u}] &= X\beta + \int_0^1 W(s)\gamma(s)ds \\ &= \boldsymbol{X\beta} + \boldsymbol{Zu} \\ \boldsymbol{u} &\sim \mathrm{N}\left[\ \boldsymbol{0}\ ,\ \sigma_{\boldsymbol{g}}^2\boldsymbol{D}\ \right] \end{aligned}$$

where $\boldsymbol{D}$ is the penalty matrix induced by the random walk prior distribution on the B-spline basis coefficients. Using this framework, extensions to regression with multiple functional predictors and to longitudinal functional regression are direct, in that one can appropriately augment the fixed and random effect design matrices according to the structure desired (Goldsmith et al., 2012). As discussed in Goldsmith et al. (2011) and Ruppert (2002), the choice of $K_w$ is less important in the PFR context than in the FPCR framework due to smoothness in $\gamma(s)$ being explicitly induced. Choosing $K_w$ sufficiently large to capture variability in the predictors and coefficient function is the only concern. Choosing $K_w$ sufficiently large would encompass all the relevant projections of $X^TX$ which include those of $X^TY$, whose relevant projections via partial least squares have been noted (Preda et al., 2007; Preda and Saporta, 2005).

Having the functional coefficient modeled in a LMM with one fixed effect and many random effects can be viewed as a problem in semiparametric regression (Ruppert et al., 2003), for which exact likelihood ratio tests (LRT) and restricted likelihood ratio tests (RLRT) have been developed (Crainiceanu et al., 2005). The LRT and RLRT centralize on the issue of testing for 0-variance components. Interpretatively, a 0-variance component ($\sigma_g^2 = 0$) sets all random effects $\{g_k\}_{k=1}^{K_g}$ identically to 0 and constrains the coefficient function to be a constant, as in test (2). Since only variance components are involved in testing, this may be tested as a LRT or RLRT, with RLRT being preferred. Similarly $\sigma_g^2 = 0$ and $\gamma_0 = 0$ restricts the coefficient function to be zero, motivating a test for inclusion

of the predictor as in test (3) via a LRT due to a fixed effect being specified in the test.

Inferentially, the p-values associated with these tests can be obtained from the nonstandard LRT and RLRT distributions of Crainiceanu and Ruppert (2004) and Greven et al. (2008). The tests are deemed nonstandard because the null value of a tested parameter is on the boundary of the parameter space and because the outcome cannot be split into independent subvectors. Previous treatments by Self and Liang (1987) and Stram and Lee (1994) each require some level of independence of the outcome. In the setting of linear mixed models with random effects representing smoothing terms, the assumption of independence under the alternative is violated, resulting in a conservative albeit computationally straightforward test. Details of the 0-variance testing procedure for use in the context of PFR are given in Section 3.

# 3 0-variance component testing methodologies

Here we review the testing procedure for 0-variance components in the LMM framework, emphasizing the applicability of this approach for testing in the PFR context. We use established software and borrow theory for penalized-spline additive models, and use tests based on the restricted likelihood ratio test (RLRT) and likelihood ratio test (LRT).

## 3.1 Restricted likelihood ratio test

The restricted likelihood ratio test statistic

$$RLRT = 2 \sup_{\theta \in H_A} \text{REL}(\theta) - 2 \sup_{\theta \in H_0} \text{REL}(\theta)$$

is suitable for testing any hypothesis that involves solely variance components, where $\text{REL}(\theta)$ denotes the restricted log-likelihood of the parameter vector $\theta$. In particular, the RLRT is relevant for testing the constancy of a coefficient function through the variance component $\sigma_g^2$.

For a LMM with one random effect variance component and fixed effects design matrix $\boldsymbol{X}$ of dimension $n \times p$, a spectral representation of the exact finite sample null distribution

exists for $n$ total observations (Crainiceanu and Ruppert, 2004):

$$RLRT_n \stackrel{d}{=} \sup_{\lambda \geq 0} \left[ (n-p-1) \log \left\{ 1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right\} - \sum_{s=1}^{K} \log(1 + \lambda \mu_{s,n}) \right],$$

where $\stackrel{d}{=}$ denotes equality in distribution, with the numerator and denominator terms

$$N_n(\lambda) = \sum_{s=1}^{K} \frac{\lambda \mu_{s,n}}{1 + \lambda \mu_{s,n}}, \quad D_n(\lambda) = \sum_{s=1}^{K} \frac{w_s^2}{1 + \lambda \mu_{s,n}} + \sum_{s=K+1}^{n-p-1} w_s^2.$$

The quantities $w_s,\ (s = 1, \ldots, n-p-1)$ and $\mu_s,\ (s = 1, \ldots, K)$ are independent standard normal random variables and $\mu_{s,n}$ are the $K$ eigenvalues of the $K \times K$ matrix $\boldsymbol{Z}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Z}$. For models with more than one random effect, a pseudo-likelihood approach of Gong and Samaniego (1981) as detailed in Greven et al. (2008) can and will be taken. Interchangeably for the single and multiple random effects models, the test based on critical values from the simulated distribution will be referred to as $RLRT_{CR}$ herein and will be the only test considered in the paper for PFR. Scheipl et al. (2008) found $RLRT_{CR}$ comparable to bootstrap-based competitors with regard to power and size-levels while providing substantial computational time reduction from hours to seconds.

## 3.2 Likelihood ratio test

Consider the likelihood ratio test statistic

$$LRT = 2 \sup_{\theta \in H_A} \mathrm{L}(\theta) - 2 \sup_{\theta \in H_0} \mathrm{L}(\theta)$$

for testing any hypothesis that involves variance components *and* fixed effect coefficients, where $\mathrm{L}(\theta)$ denotes the log-likelihood of the parameter vector $\theta$. Some hypothesis tests will affect the parameterization of the mean in addition to a 0-variance component. In such instances, the LRT is needed.

For a LMM with one random effect variance component, a spectral representation of the exact finite sample null distribution exists (Crainiceanu and Ruppert, 2004):

$$LRT_n \stackrel{d}{=} \sup_{\lambda \geq 0} \left[ n \log \left\{ 1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right\} - \sum_{s=1}^{K} \log(1 + \lambda \xi_{s,n}) \right] + n \left( 1 + \frac{\sum_{s=1}^{q} \mu_s^2}{\sum_{s=1}^{n-p-1} w_s^2} \right).$$

The quantities $w_s$, $\mu_s$, $\mu_{s,n}$, $N_n(\lambda)$, and $D_n(\lambda)$, are as previously defined and $\xi_{s,n}$ are the $K$ eigenvalues of the $K \times K$ matrix $\boldsymbol{Z}'\boldsymbol{Z}$. Scalar $q$ indexing $\sum_{s=1}^{q} \mu_s^2$ is the number of fixed effects being tested; if 0 the $RLRT$ is preferred.

For models with more than one random effect, a pseudo-likelihood approach is taken as described for the $RLRT$. Interchangeably for the single and multiple random effects models, the test based on critical values from the simulated distribution will be referred to as $LRT_{CR}$ herein.

# 4 Simulation

In this section we explore the properties of the inferential procedures developed in Sections 2 and 3 paper. We focus on the test for functional structure as presented in equation (2).

## 4.1 Testing in the Standard FLM

Our simulations are motivated by the neuroimaging application considered in Section 5. Functional predictors are generated using the observed principal component basis functions $\boldsymbol{\psi}^O(s)$, score variances $\boldsymbol{\lambda}^O$, mean function $\mu^O(s)$, and measurement error variance $\sigma^{2,O}$ obtained from a FPC decomposition of the observed curves $W_i^O(s)$ for subjects in the study. To construct simulated predictors $W_i^S(s)$ with measurement error, $\epsilon(s)$, we generate subject-specific PC loadings using $\boldsymbol{c}_i^2 \sim \mathrm{N}\left[0, \mathrm{diag}(\boldsymbol{\lambda})\right]$ and let $W_i^S = \mu^O(s) + \sum_{k=1}^{15} c_{ik}^S \psi_k^O(s) + \epsilon(s)$, $1 \leq i \leq I$ where $\epsilon(s)$ is iid $\mathrm{N}\left[0, \sigma^{2,O}\right]$ for each $s$.

Simulated outcomes $Y_i^S$ are given by $Y_i^S = \alpha + \int_0^1 W_i^S(s)\gamma^{(S)}(s) + \epsilon_i$ where $\alpha = 2$ and $\epsilon \sim \mathrm{N}\left[0, 1\right]$. The coefficient function $\gamma^S(s)$ in simulations is based on the estimate $\gamma^O(s)$ from the real data analysis: $\gamma^S = \overline{\gamma^O(s)} + r(\gamma^O(s) - \overline{\gamma^O(s)})$ where $r$ is a scaling factor that we vary. Constructed in this way, $\gamma^S(s)$ is a combination of the constant and deviation from constant observed in the real data coefficient function. Figure 2 illustrates the simulation design by plotting the observed curves, a single generated dataset, and the $\gamma^S(s)$ that result from several choices of the scaling factor $r$.

Using the procedure above, ten thousand data sets are generated for each of the following parameter combinations:

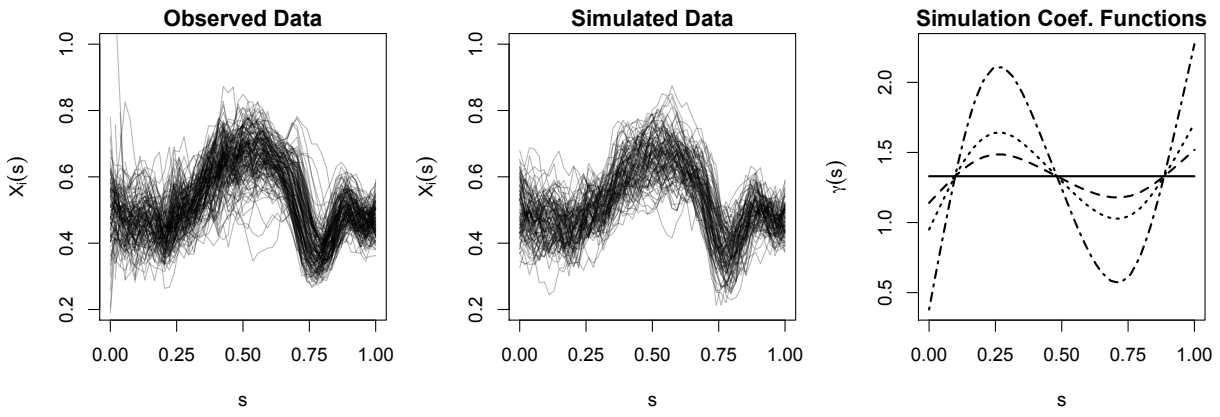


Figure 2: The left panel displays the observed functional predictors upon which simulated data is based. The middle panel shows one simulated dataset with $I = 100$. The right panel shows four coefficient functions used in simulations, with scaling factor ranging from $r = 0$ (null scenario) to $r = .1$.

1. Sample sizes (a) $I = 100$; (b) $I = 250$; (c) $I = 500$;

2. Scaling factors (b) $r = 0$; (b) $r = 0.05$; (c) $r = 0.10$; (d) $r = 0.25$.

This gives a total of 12 possible simulation designs. For each simulated data set under each design, the FLM is fitted via PFR with $K_g = K_w = 30$ and no scalar covariates. For the FPCR approach in Section 2 of the main paper, the tuning parameter $K_g$ is either fixed at $K_g = 2$ or chosen as the minimum number of PCs needed to explain at least 90% of variability in simulated functional predictors. We test the null hypothesis of a constant coefficient function in the PFR setting using the procedures developed in Section 3 of the main paper and in the FPCR setting using the standard LRT. Note that the null hypothesis is true for simulations in which $r = 0$.

Table 1 reports the probability of rejecting the null hypothesis at the .05 and .01 levels under each of the three testing scenarios, labeled as "$\text{FPCR}_2$" for the FPCR approach with $K_g = 2$, "$\text{FPCR}_{pve}$" for the FPCR approach with $K_g$ chosen using the percent variance explained, and "PFR" for the penalized approach. Also included in Table 1 is a comparison of the average mean squared error (AMSE), which is the sum of each simulation's integrated $\text{MSE} = \int_0^1 \left(\widehat{\gamma^S}(s) - \gamma^S(s)\right)^2 ds$ divided by the number of simulations (10,000) . To compute

the MSE, $\widehat{\gamma^S}(s)$ is the estimate under the null model if $p > .05$ and is the estimate under the alternative model otherwise.

| | Reject at .05 | | | Reject at .01 | | | AMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\text{FPCR}_2$ | $\text{FPCR}_{pve}$ | PFR | $\text{FPCR}_2$ | $\text{FPCR}_{pve}$ | PFR | $\text{FPCR}_2$ | $\text{FPCR}_{pve}$ | PFR |
| $I = 100$ | | | | | | | | | |
| $r = 0.00$ | 0.059 | 0.066 | 0.041 | 0.012 | 0.014 | 0.007 | 1.374 | 3.290 | 0.914 |
| $r = 0.05$ | 0.132 | 0.119 | 0.078 | 0.039 | 0.032 | 0.018 | 3.162 | 5.868 | 2.458 |
| $r = 0.10$ | 0.318 | 0.274 | 0.207 | 0.138 | 0.112 | 0.078 | 6.922 | 11.964 | 6.341 |
| $r = 0.25$ | 0.931 | 0.939 | 0.911 | 0.812 | 0.822 | 0.787 | 19.586 | 26.155 | 15.158 |
| $I = 250$ | | | | | | | | | |
| $r = 0.00$ | 0.062 | 0.058 | 0.043 | 0.014 | 0.014 | 0.008 | 0.516 | 1.253 | 0.323 |
| $r = 0.05$ | 0.242 | 0.198 | 0.150 | 0.095 | 0.068 | 0.052 | 1.967 | 3.699 | 1.650 |
| $r = 0.10$ | 0.666 | 0.602 | 0.536 | 0.426 | 0.361 | 0.318 | 4.443 | 8.018 | 4.255 |
| $r = 0.25$ | 0.999 | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | 15.763 | 14.162 | 7.292 |
| $I = 500$ | | | | | | | | | |
| $r = 0.00$ | 0.073 | 0.066 | 0.048 | 0.017 | 0.015 | 0.010 | 0.294 | 0.770 | 0.171 |
| $r = 0.05$ | 0.446 | 0.370 | 0.297 | 0.230 | 0.172 | 0.134 | 1.551 | 3.111 | 1.369 |
| $r = 0.10$ | 0.929 | 0.914 | 0.883 | 0.812 | 0.773 | 0.740 | 3.218 | 5.137 | 2.521 |
| $r = 0.25$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 14.834 | 9.756 | 4.611 |

Table 1: Average rejection probability at the .05 and .01 thresholds for the null hypothesis of a constant coefficient function. Tests are performed using the FPCR approach with $K_g = 2$ ("$\text{FPCR}_2$"), the FPCR approach with $K_g$ chosen as the smallest value needed to explain 90% of observed variability ("$\text{FPCR}_{pve}$"), and the PFR method ("PFR"). $100\times$ AMSE for the coefficient function is also provided.

Several key points are apparent in Table 1. First, the tests we propose have the appropriate size under the null hypothesis, although we note that the PFR is slightly conservative for smaller samples and the FPCR approaches are anti-conservative for all sample sizes. Second, power to detect a true alternative hypothesis increases both as sample size increases and as the size of the effect increases. The FPCR approaches have power that is greater than or equal to the PFR approach in all circumstances; the non-inferior power is likely related to the relative simplicity of the model and to the anti-conservatism under the null. Third, the PFR method uniformly outperforms the FPCR methods in terms of AMSE, often substantially. The relative performance of the $\text{FPCR}_2$ and $\text{FPCR}_{pve}$ changes as $r$ increases. For low values of $r$ the extra PCs used in the $\text{FPCR}_{pve}$ method lead to overfitting, while for larger values of $r$ these become useful basis functions for expanding

the coefficient. Table 2 shows the size-corrected power for the simulation, for which PFR uniformly bests $\text{FPCR}_{pve}$ at size 0.01. Finally, although not shown we note that the computational burden of the three approaches is similar (For 100, 250, and 500 subjects, 1 run on average takes 6, 8, and 10 seconds, respectively). The most computationally expensive step is the estimation of a FPC decomposition, which is common to all methods, and the model fitting is done using efficient implementations.

| | Reject at .05 | | | Reject at .01 | | |
|---|---|---|---|---|---|---|
| | $\text{FPCR}_2$ | $\text{FPCR}_{pve}$ | PFR | $\text{FPCR}_2$ | $\text{FPCR}_{pve}$ | PFR |
| $I = 100$ | | | | | | |
| $r = 0.00$ | 0.050 | 0.050 | 0.050 | 0.010 | 0.010 | 0.010 |
| $r = 0.05$ | 0.118 | 0.093 | 0.093 | 0.033 | 0.023 | 0.025 |
| $r = 0.10$ | 0.294 | 0.230 | 0.236 | 0.125 | 0.086 | 0.100 |
| $r = 0.25$ | 0.922 | 0.919 | 0.922 | 0.796 | 0.785 | 0.820 |
| $I = 250$ | | | | | | |
| $r = 0.00$ | 0.050 | 0.050 | 0.050 | 0.010 | 0.010 | 0.010 |
| $r = 0.05$ | 0.217 | 0.179 | 0.166 | 0.076 | 0.050 | 0.059 |
| $r = 0.10$ | 0.629 | 0.574 | 0.558 | 0.381 | 0.309 | 0.338 |
| $r = 0.25$ | 0.999 | 1.000 | 1.000 | 0.998 | 1.000 | 1.000 |
| $I = 500$ | | | | | | |
| $r = 0.00$ | 0.050 | 0.050 | 0.050 | 0.010 | 0.010 | 0.010 |
| $r = 0.05$ | 0.382 | 0.320 | 0.302 | 0.171 | 0.129 | 0.139 |
| $r = 0.10$ | 0.903 | 0.892 | 0.884 | 0.753 | 0.721 | 0.746 |
| $r = 0.25$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 2: Size-adjusted average rejection probability at the .05 and .01 thresholds for the null hypothesis of a constant coefficient function. Tests are performed using the FPCR approach with $K_g = 2$ ("$\text{FPCR}_2$"), the FPCR approach with $K_g$ chosen as the smallest value needed to 90% of observed variability ("$\text{FPCR}_{pve}$"), and the PFR method ("PFR"). $100\times$ average MSE for the coefficient function is also provided. Cutoffs are chosen to ensure nominal coverage under the null hypothesis.

# 5 Application

We turn our attention to the study of intracranial white matter microstructure that is the motivation for our work. Of interest is whether differences in cognitive function can be explained by changes in white matter observed longitudinally in a cohort of multiple

sclerosis (MS) patients. MS is an immune-mediated inflammatory disease that is associated with the demyelination of white matter fibers. Because the myelin sheath surrounds and protects the axons which rapidly propagate electrical signals in the brain, damage to this insulation can result in severe cognitive and motor disability. To quantify white matter properties, diffusion tensor imaging is used to produce detailed images of white matter tissue by tracing the diffusion of water in the brain (Basser et al., 1994, 2000; LeBihan et al., 2001; Mori and Barker, 1999). From these images, continuous summaries of major white matter structures called tract profiles can be obtained.

Our data contain 100 subjects with between 2 and 8 visits each, for a total of 340 visits. Corpus callosum (CCA) and right corticospinal tract (RCST) profiles and tests of cognitive ability were obtained at each visit, longitudinally. The CCA and RCST profiles were registered to a common, regularly spaced, dense grid of 93 and 55 points respectively. In this analysis we focus on the Paced Auditory Serial Addition Test (PASAT) as a measure of cognitive performance (Gronwall, 1977). Our goal is to understand the relationship between this score and tract profiles of the CCA and RCST. The CCA is a major white matter structure connecting the left and right hemispheres of the brain, and damage to this structure has previously been linked to a decline in cognitive performance among MS patients (Ozturk et al., 2010). The right and left corticospinal tracts connect the motor cortex to the brain stem. Figure 1 illustrates the position of the corpus callosum and corticospinal tracts in the brain. We also show the tract profiles and scalar outcomes observed to illustrate the longitudinal functional data structure we address. Note this data set has been considered previously in (Goldsmith et al., 2012).

We conduct three tests to evaluate the strength of association between the spatially dynamic white matter integrity of the corpus callosum ($W_{ij1}$) and right corticospinal tract ($W_{ij2}$) with the PASAT score. All models considered include a binary variable indicating each subject's first visit to adjust for the learning effect common in cognitive testing. We also include subject-specific random intercepts to account for repeated observations within subjects (additional tests indicate that the random intercepts are a crucial component of any model considered). All PFR models used ($K_w = 10,\ K_g = 50$) and ($K_w = 30,\ K_g = 30$) and all FPCR models $K_w = 30$ and $K_g = 10$.

As discussed in the Introduction, two scientific questions motivated the statistical development: 1) Is the effect of a tract indeed functional, varying over location or is the effect constant over location as a whole-tract average? and 2) Does accounting for a motor functioning tract in the presence of a cognitive tract add significantly to the model fit, that is to say is some part of PASAT's variance not just cognitive but related to motor functioning damage? Scientific question 1) is addressed by Test 1 and Test 2 (hypotheses (4) and (5)) below for each tract; scientific question 2) is addressed by Test 3 (hypothesis (6)).

- Test 1: Does the functional structure of the corpus callosum significantly improve beyond a mean-only model? To answer this we test

$$\begin{aligned} H_0 &: \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \overline{W}_{ij1}\beta_W \\ H_A &: \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \int_0^1 W_{ij1}(s)\gamma_1(s)ds. \end{aligned} \tag{4}$$

  With $RLRT_{CR} = 3.58$ and $p_{CR} = 0.01$ for $(K_w = 10,\ K_g = 50)$, the test rejects $H_0$ in favor of $H_A$, indicating that the functional structure modeled for the corpus callosum within the analysis can be justified. Results for the $(K_w = 30,\ K_g = 30)$ analysis agree: $RLRT_{CR} = 4.12$ and $p_{CR} = 0.02$.

- Test 2: Does the functional structure of the right corticospinal tract significantly improve beyond a mean-only model? To answer this we test

$$\begin{aligned} H_0 &: \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \overline{W}_{ij2}\beta_W \\ H_A &: \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \int_0^1 W_{ij2}(s)\gamma_1(s)ds. \end{aligned} \tag{5}$$

  With $RLRT_{CR} = 3.00$ and $p_{CR} = 0.02$ for $(K_w = 10,\ K_g = 50)$, the test rejects $H_0$ in favor of $H_A$, indicating that the functional structure modeled for the right corticospinal tract within the analysis can be justified. Results for the $(K_w = 30,\ K_g = 30)$ analysis agree: $RLRT_{CR} = 2.74$ and $p_{CR} = 0.02$.

- Test 3: Does modeling RCST ($W_{ij2}$) add significantly to a model fit with CCA ($W_{ij1}$)

alone? We test

$$H_0 : \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \int_0^1 W_{ij1}(s)\gamma_1(s)ds \tag{6}$$
$$H_A : \mathrm{E}[Y_{ij}] = \alpha + b_i + \boldsymbol{X}_i\beta + \int_0^1 W_{ij1}(s)\gamma_1(s)ds + \int_0^1 W_{ij2}(s)\gamma_2(s)ds.$$

With $LRT_{CR} = 0.12$ and $p_{CR} = 0.78$ for ($K_w = 10,\ K_g = 50$), the test fails to reject $H_0$ in favor of $H_A$, indicating that omitting RCST from the analysis can be justified. Results for the ($K_w = 30,\ K_g = 30$) analysis agree: $RLRT_{CR} = 0.00$ and $p_{CR} = 1.00$.

Given the inferential conclusions matched for all PFR and FPCR models and based on our simulated results of PFR producing higher AMSE, we discuss solely the PFR $K_w = 10,\ K_g = 50$) results (for thoroughness, FPCR results: Test 1 $LRT_{FPCR} = 2.45$, $p_{FPCR} = 0.01$; Test 2 $LRT_{FPCR} = 3.48$, $p_{FPCR} < 0.01$; Test 3 $LRT_{FPCR} = 0.73$, $p_{FPCR} = 0.70$).

Figure 3 shows the resulting coefficient functions from fitting the alternative hypotheses in (4)-(6). In both univariate models (models with a single functional predictor) the coefficient functions are dynamic over the domain, suggesting that qualitative assessments might conclude that either predictor's functional structure contributes to the model. Tests (4) and (5) confirm this presupposition. However, the results of the multivariate model show a coefficient function for the right corticospinal tract that is constant and near zero, while the coefficient function for the corpus callosum is largely unchanged from the univariate analysis. These multivariate visual results agree with the result of test (6). PFR sensitivity analyses with pairwise combinations of $K_w \in \{5, 6, 7, 8, 9, 10, 11\}$ and $K_g \in \{K_w + 1, K_w + 5, 20, 40, 60\}$ show no change in inference for any of the tests at the 0.05 confidence level, with p-values ranging from (0.001, 0.04), (0.01, .04), (0.28, 1.0) for tests (4), (5), and (6), respectively. Furthermore, the results stabilize above values $K_w = 9,\ K_g = 20$ for PFR.

The claim of inference being robust to values of $K_w$ and $K_g$ assumes an alpha-level of 0.05 for each of the tests. If the alpha level was lower, inference might be sensitive to $K_w$ and $K_g$. Also, there is an issue of test (6) being conditional on test (4), which could motivate a Bonferroni adjustment making the alpha-level 0.025 for each test, also making the outlined tests sensitive to $K_w$ and $K_g$. Again, the optimal selection of $K_w$ and $K_g$ is

a difficult and open problem. As a matter of practice to protect inferential integrity, the authors advocate selecting $K_w$ so that 99% of the variance is explained and $K_g = 50$, if computationally feasible. Since p-values tend to stabilize for increasing $K_w$ and $K_g$, setting $K_w$ and $K_g$ in such a fashion will ensure they are in the upper range of most contexts and correspondingly p-values being more reflective of the data than of the modeling choice.

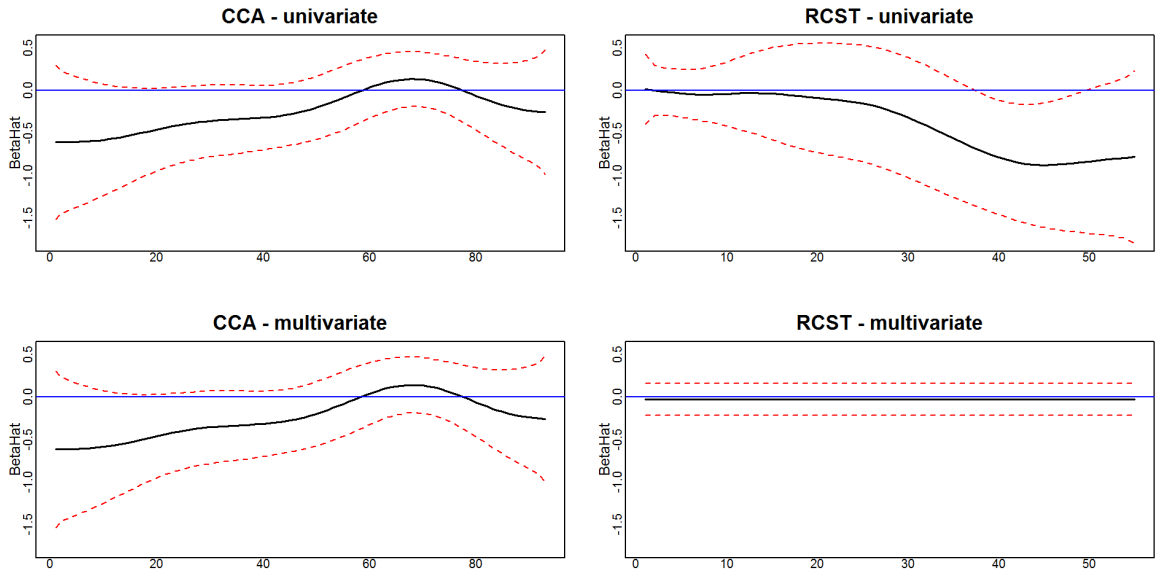


Figure 3: The top panels of this figure show the coefficient function estimates that result from the univariate alternative models in (4) and (5) on the left and right respectively. The bottom panels show the coefficient function estimates that result from the multivariate alternative model in (6). This figure appears in color in the electronic version of this article.

For both univariate tests including functional information is supported and deems the spatial information of a brain tract as important in the relationship between and cognitive performance and demyelination of the tracts in isolation. Our statistical results are also consistent with the scientific information regarding the anatomical structures of interest. In particular we recall that the corticospinal tracts are primarily transmitters of motor signals that should not directly affect cognitive performance. On the other hand, the corpus callosum connects the hemispheres of the brain and may be relevant for the PASAT measure of cognitive function, which involves auditory processing, short term memory, and arithmetic computations. Plausibly, the corticospinal tract could be related to cognitive function as

a measure of overall disease burden and lesion load, as evinced by (5). Additionally, the corticospinal tract is closely correlated with the corpus callosum in the region of interest indicated in the univariate analysis (positions 40-50). The results of these hypothesis tests support a case for functional confounding of the relationship between the PASAT score and the right corticospinal tract by the corpus callosum.

# 6 Concluding remarks

Often, the intuitive assumption is that the functional structure of predictors contains useful information for exploring associations with an outcome of interest. However, the case may often be that the relevant quantities are captured by much simpler forms. In this paper we have developed a framework for rigorous hypothesis testing framework that compares the null of a constant coefficient function to a more flexible, spatially varying coefficient. Under the null hypothesis, only the mean of functional predictors is retained as a covariate in a standard linear model.

The application results of Section 5 emphasize the trouble intuition can cause and the usefulness of explicit hypothesis tests. For both univariate tests (4) and (5) a constant function can comfortably fit between the 95% pointwise confidence bounds of the estimated coefficient function, but in both cases the test of a constant coefficient function is rejected at the .05 level by a hypothesis test for constancy. Moreover, although the corpus callosum is a significant predictor of the PASAT score, the pointwise confidence interval for the coefficient function contains zero over the entire domain. In situations like these heuristic arguments can fail badly and proper inference requires the use of hypothesis testing.

Our approaches to this problem mirror the development of estimation procedures for the FLM. First, we consider a low-dimensional approach based on the use of principal component loadings as predictors in a linear model. This method is a modification of the popular FPCR technique that has a relatively long history in the FDA literature. Next we implement an estimation procedure that uses a flexible spline basis for the coefficient function and induces smoothness via penalization in a mixed model framework. Testing in the first case is relatively straightforward and is based on a standard likelihood ratio test, while in the second case we develop nonstandard (restricted) likelihood ratio tests for zero

variance components. Simulations indicate that the tests achieve the nominal size under the null and have reasonable power to detect true alternative hypotheses.

As is always the case when comparing low-dimensional and penalized approaches to functional regression, there are certain tradeoffs in the context of hypothesis testing. The FPCR approach is more straightforward to implement for both estimation and testing, and in simulation exercises appears to have more power to detect a true alternative (although it is somewhat anti-conservative under the null). The penalized approach allows for more flexible estimation of the coefficient function, but requires more sophisticated estimation and testing techniques. Deciding which approach is most appropriate is often context-specific, although we generally recommend the more flexible penalized approach in the absence of compelling justifications for the low-dimensional method.

Several directions for future work are apparent. The use of the mixed model framework to induce penalization is common in the FDA literature, ranging from smoothing estimates of individual curves to penalization in function-on-function and function-on-scalar regression models. Our work indicates that (restricted) likelihood ratio tests are well-suited to testing in functional settings and could be adapted to the contexts above. In this manuscript we have focused on testing for continuous outcome regression models, but considering functional generalized linear models is important as well.

## ACKNOWLEDGEMENTS

The diffusion tensor imaging data were collected at Johns Hopkins University and the Kennedy-Krieger Institute under the direction of Peter A. Calabresi, MD. Crainiceanu and Swihart were supported by Grant Number R01NS060910 from the National Institute of Neurological Disorders and Stroke. Crainiceanu was also supported by Grant Number R01EB012547 from the National Institute of Biomedical Imaging And Bioengineering.

## SUPPLEMENTARY MATERIAL

**R-code and data:** The defusion tensor imaging data and functions `rlrt.pfr()` and `predict.pfr()` are available in the R-package `refund` (Crainiceanu et al., 2013).

# References


Aerts, M., Claeskens, G., and Wand, M. (2002). Some theory for penalized spline generalized additive models. *Journal of statistical planning and inference* **103,** 455–470.

Aneiros-Pérez, G. and Vieu, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis* **99,** 834–857.

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. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13,** 571–591.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika* **81,** 541–553.

Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* **39,** 1720–1747.

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L., and Scheipl, F. (2013). *refund: Regression with Functional Data.* R package version 0.1-8.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92,** 91–103.

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

Delsol, L., Ferraty, F., and Vieu, P. (2011). Structural test in regression on functional variables. *Journal of Multivariate Analysis* **102,** 422–447.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* **4,** 458–488.

Eilers, P. H. C., Li, B., and Marx, B. D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems* **96,** 196–202.

Fahrmeir, L. and Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50,** 201–220.

Fahrmeir, L. and Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the institute of Statistical Mathematics* **53,** 11–30.

Ferraty, F., Goia, A., Salinelli, E., and Vieu, P. (2012). Functional projection pursuit regression. *Test* pages 1–28.

Ferraty, F., González-Manteiga, W., Martínez-Calvo, A., and Vieu, P. (2011). Presmoothing in functional linear regression. *Statistica Sinica* **22,** 69–94.

Ferraty, F., Hall, P., and Vieu, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97,** 807–824.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer.

Gertheiss, J., Goldsmith, J., Crainiceanu, C., and Greven, S. (2012). Longitudinal scalar-on-functions regression with application to tractography data. *(Submitted)* .

Gertheiss, J., Maity, A., and Staicu, A.-M. (2013). Variable selection in generalized functional linear models. *Stat* **2,** 86–101.

Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20,** 831–851.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C* **61,** 453–469.

Goldsmith, J., Wand, M. P., and Crainiceanu, C. M. (2011). Functional regression via variational bayes. *Electronic Journal of Statistics* **5,** 572–602.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* pages 861–869.

Goutis, C. and Fearn, T. (1996). Partial least squares regression on smooth factors. *Journal of the American Statistical Association* **91,** 627–632.

Greven, S., Crainiceanu, C., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* **17,** 870–891.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills* **44,** 367–373.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science* **15,** 196–223.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87,** 587–602.

James, G. M. and Silverman, B. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association* **100,** 565–576.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics* **37,** 2083–2108.

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

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13,** 183–212.

LeBihan, D., Mangin, J., Poupon, C., and Clark, C. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging* **13,** 534–546.

Lindquist, M. and McKeague, I. (2009). Logistic regression with Brownian-like predictors. *Journal of the American Statistical Association* **104,** 1575–1585.

Marx, B. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* **28,** 193–209.

Marx, B. D. and Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics* **41,** 1–13.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2012). Functional generalized additive models. *Journal of Computational and Graphical Statistics* .

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician* **61,** 22–27.

Mori, S. and Barker, P. (1999). Diffusion magnetic resonance imaging: its principle and applications. *The Anatomical Record* **257,** 102–109.

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

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of statistical software* **9,** 1–54.

Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P., and Reich, D. (2010). MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis* **16,** 166–177.

Preda, C. and Saporta, G. (2005). Pls regression on a stochastic process. *Computational Statistics & Data Analysis* **48,** 149–158.

Preda, C., Saporta, G., and Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics* **22,** 223–235.

Ramsay, J. J. O., Hooker, G., and Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer.

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

Reiss, P. and Ogden, R. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association* **102,** 984–996.

Reiss, P. and Ogden, T. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66,** 61–69.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11,** 735–757.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*, volume 12. Cambridge Univ Pr.

Scheipl, F., Greven, S., and Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* **52,** 3283–3299.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* pages 605–610.

Staniswalis, J. and Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **444,** 1403–1418.

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* pages 1171–1177.

Tzeng, J. and Zhang, D. (2007). Haplotype-based association analysis via variance-components score test. *American journal of human genetics* **81,** 927.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59,** 254–262.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18,** 223–250.

Wood, S. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika* .

Xiao, L., Ruppert, D., Zipunnikov, V., and Crainiceanu, C. (2013). Fast Covariance Estimation for High-dimensional Functional Data. *ArXiv e-prints* .

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

Yao, F., Müller, H. G., Clifford, A., Dueker, S., Follett, J., Lin, Y., Buchholz, B., and Vogel, J. (2003). Shrinkage estimation for functional principal component scores with application to the population. *Biometrics* **59,** 676–685.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

Zhang, D. and Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. *Random effect and latent variable model selection* pages 19–36.