



---

Johns Hopkins University, Dept. of Biostatistics Working Papers

---

11-2-2012

# LONGITUDINAL FUNCTIONAL MODELS WITH STRUCTURED PENALTIES

Madan G. Kundu

*Indiana University School of Medicine, Department of Biostatistics*

Jaroslav Harezlak

*Indiana University Fairbanks School of Public Health, Department of Biostatistics, harezlak@iupui.edu*

Timothy W. Randolph

*Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics Program*

---

## Suggested Citation

Kundu, Madan G.; Harezlak, Jaroslav; and Randolph, Timothy W., "LONGITUDINAL FUNCTIONAL MODELS WITH STRUCTURED PENALTIES" (November 2012). *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 248. <http://biostats.bepress.com/jhubiostat/paper248>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

## Longitudinal Functional Models with Structured Penalties

Madan G. Kundu<sup>1</sup>, Jaroslaw Harezlak<sup>1,2,\*</sup>, and Timothy W. Randolph<sup>3</sup>

<sup>1</sup>Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202

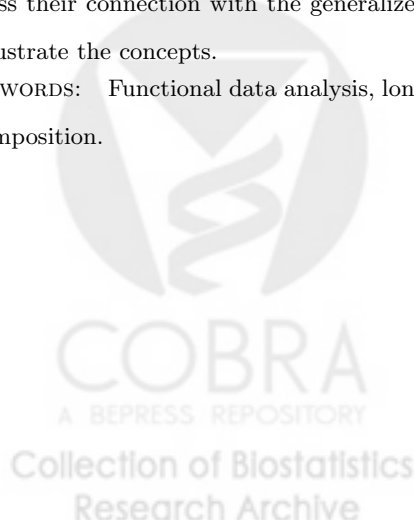
<sup>2</sup>Department of Biostatistics, Indiana University Fairbanks School of Public Health, Indianapolis, IN 46202

<sup>3</sup>Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

\**email*: harezlak@iupui.edu

**SUMMARY:** Collection of functional data is becoming increasingly common including longitudinal observations in many studies. For example, we use magnetic resonance (MR) spectra collected over a period of time from late stage HIV patients. MR spectroscopy (MRS) produces a spectrum which is a mixture of metabolite spectra, instrument noise and baseline profile. Analysis of such data typically proceeds in two separate steps: feature extraction and regression modeling. In contrast, a recently-proposed approach, called partially empirical eigenvectors for regression (PEER) (Randolph, Harezlak and Feng, 2012), for functional linear models incorporates a priori knowledge via a scientifically-informed penalty operator in the regression function estimation process. We extend the scope of PEER to the longitudinal setting with continuous outcomes and longitudinal functional covariates. The method presented in this paper: 1) takes into account external information; and 2) allows for a time-varying regression function. In the proposed approach, we express the time-varying regression function as linear combination of several time-invariant component functions; the time dependence enters into the regression function through their coefficients. The estimation procedure is easy to implement due to its mixed model equivalence. We derive the precision and accuracy of the estimates and discuss their connection with the generalized singular value decomposition. Real MRS data and simulations are used to illustrate the concepts.

**KEY WORDS:** Functional data analysis, longitudinal data, mixed model, structured penalty, generalized singular value decomposition.

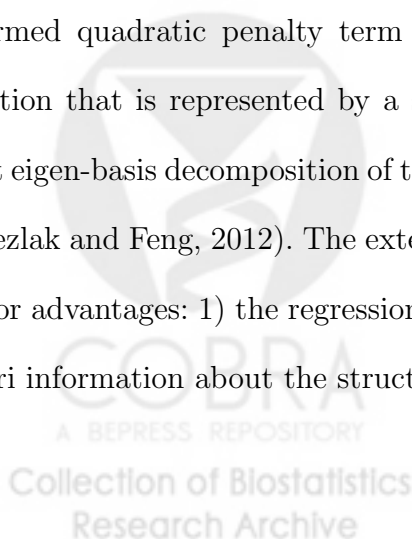


This paper has been submitted for consideration for publication in *Biometrics*

## 1. Introduction

Technological advancements and increased availability of storage of large datasets have allowed for the collection of functional data as part of time-course or longitudinal studies. In the cross-sectional setting, there have been many proposed methods for estimating a regression function in a so-called functional linear model (fLM). This function is a functional (continuous) analogue of a vector of (discrete) regression coefficients; it connects the scalar response,  $y$  to a functional covariate,  $w \equiv w(s)$ . Although these models have recently been well studied, extensions to longitudinally-collected functions have not received much attention. A recently-proposed approach called longitudinal penalized functional regression (LPFR) approach extends cross-sectional fLM to a longitudinal setting by incorporating subject-specific random intercepts (Goldsmith et al., 2012). One of the key assumption in LPFR is that the regression function remains constant over time. Due to this restrictive assumption, LPFR is not suited for situations in which the association between the functional predictors and scalar response may evolve over time. In this manuscript, we propose a technique that extends the analysis of functional linear models by relating scalar outcomes to functional predictors, both observed longitudinally, and allows the estimation of a time-varying regression function.

The method fits into a generalized ridge regression framework by imposing a scientifically-informed quadratic penalty term into the estimation process. The resulting estimate is a function that is represented by a set of “partially empirical” eigenvectors that arise from a joint eigen-basis decomposition of the predictor functions and the penalty term; see (Randolph, Harezlak and Feng, 2012). The extension of this framework to the longitudinal setting has two major advantages: 1) the regression function is allowed to vary over time; and 2) external or a priori information about the structure of the regression function can be incorporated directly



into the estimation process. We formulate the estimation procedure within a mixed-model framework making the method computationally efficient and easy to implement.

Ramsay and Dalzell (1991) introduced the term functional data analysis (FDA) in the statistical literature. The cross-sectional fLM with scalar response can be stated as follows (see e.g., Yao and Müller, 2010)

$$E(y|W) = \mu_y + \int_{\Omega} W(s)\gamma(s)ds$$

where  $\mu_y$  is the mean of  $y$ ,  $\Omega$  denotes the domain of the predictor functions  $W(s)$ ,  $s \in \Omega$ , and  $\gamma(s)$  is a square integrable function that models the linear relationship between the functional predictors and scalar response. We will assume that  $W(\cdot)$  denotes a mean-centered function ( $E[W(s)] = 0$  for almost all  $s \in \Omega$ ). As there is no unique  $\gamma(\cdot)$  that solves this equation, additional regularization or constraint is required. Typically, some form of smoothness is imposed on  $\gamma(\cdot)$ , one approach being to expand both regression function  $\gamma(\cdot)$  and predictor function  $W(\cdot)$  in terms of a set of spline basis functions such as B-splines and then obtain the regularized estimate of  $\gamma(\cdot)$  (Ramsay and Silverman, 1997). Another approach is to express the regression function  $\gamma(\cdot)$  in terms of the orthonormal eigenfunctions of covariance of  $W(\cdot)$  using Karhunen-Loève (K-L) basis expansion (see e.g., Müller, 2005). A third approach is to combine the above two approaches, known as penalized functional regression (PFR) approach (Goldsmith et al., 2011). In PFR approach, a spline basis is used to represent the regression function and a basis of eigenfunctions from the set of predictors is used to represent each  $W(\cdot)$ . Another approach is to use wavelet basis, instead of eigenfunctions, to represent the predictor functions (Morris and Carroll, 2006).

Here we adopt an approach by Randolph, Harezlak and Feng (2012) which does not begin by explicitly projecting onto a pre-specified basis of functions. Instead, prior information about functional structure is incorporated into the estimation process by way of a penalty operator.

“Partially empirical eigenvectors for regression” (PEER) approach exploits the fact that, in the familiar framework of penalized least-squares regression, the estimate arises from a set of basis functions that is determined jointly by the covariance (empirical spatial structure) and the penalty (imposed structure). This naturally extends ridge regression (noninformative structure) and smoothing penalties such as a second-derivative penalty (smooth regression function assumption). In this paper, we extend the scope of the PEER approach to the longitudinal setting in a manner that allows the estimated regression function  $\gamma \equiv \gamma(t, \cdot)$  to vary with time.

The problem we address involves repeated observations from each of  $N$  subjects. At each observation time,  $t$ , we collect data on a scalar response variable,  $y$ , and a (idealized) predictor function,  $W(\cdot)$ . We are interested in longitudinal regression models of the following form:

$$y_t = x_t^\top \beta + \int_{\Omega} W_t(s) \gamma(t, s) ds + \epsilon_t \quad (1)$$

Here  $\gamma(t, \cdot)$  denotes the regression function at time  $t$ ,  $x_t$  is a vector of scalar-valued (non-functional) predictors. In a similar spirit to that of a linear mixed model with time-related slope for longitudinal data, we assume that  $\gamma(t, \cdot)$  can be decomposed into several time-invariant component functions; e.g.,  $\gamma(t, \cdot) = \gamma_0(\cdot) + t \gamma_1(\cdot)$ .

The recently proposed LPFR approach assumes the regression function in (1) is independent of time and proceeds in three steps: uses a truncated set of K-L vectors to represent the predictor functions; expresses the regression function using a spline basis; and fits the longitudinal model using an equivalent mixed-model framework that incorporates subject-specific random effects. In contrast, we model the coefficient function  $\gamma(t, \cdot)$  as a time-dependent combination of several time-invariant component functions,  $\{\gamma_d(\cdot)\}_{d=0}^D$ , each of which is estimated via a penalty operator that is informed by the structure of the data or a scientific question.

Our work is motivated by a study in which magnetic resonance (MR) spectra have been

collected longitudinally from late stage HIV patients (Harezlak et al., 2011). The observed spectra from MR spectroscopy (MRS) are a mixture of pure metabolite spectra, baseline profile and instrument noise.

Pure metabolite spectra provide information about the spatial structure of the observed spectra and consequently can be used to inform the regression function. Figure 1 shows the baseline MR spectra,  $W(\cdot)$ , from a brain region called basal ganglia for 114 subjects and also plots nine pure metabolite spectra. We are interested in studying the association of  $y$  with the metabolite concentrations. It is natural that the spatial structure (shape) of these metabolite spectra should inform the process of estimating the relationship between  $y$  and  $W(\cdot)$ ; i.e., the subspace spanned by the functional structure of the pure metabolite spectra may be more informative than structures such as B-spline or cosine functions that are “external” to the problem. Therefore, the statistical methodology applied in this paper estimates  $\gamma$  focusing on a subspace that is spanned by a basis of generalized singular vectors that arise jointly from the predictors,  $W(\cdot)$ , and the metabolite spectra.

[Figure 1 about here.]

The cross-sectional fLM with scalar response has been a focus of various investigations (Ramsey and Silverman, 1997; Faraway, 1997; Fan and Zhang, 2000; Cardot, Ferraty and Sarda, 1999, 2003; Cai and Hall, 2006; Cardot et al., 2007; Reiss and Ogden, 2009). Some of these methods estimate regression functions in two steps. For example, principal components regression (PCR) estimates of the regression function are obtained first and then these PCR estimates are projected onto a B-spline basis (Cardot, Ferraty and Sarda, 2003) or vice-versa; i.e., PCR fitting is performed only after projection onto B-splines (Reiss and Ogden, 2009). Extensions of fLM have been made towards generalized linear model with functional predictors (James, 2002; Müller and Stadtmüller, 2005) and quadratic functional regression (Yao and

Müller, 2010). Another class of models, known as Functional Analysis of Variance (FANOVA), decompose repetitively-observed functional predictors into several (fixed and random) groups and subject-specific component functions (Brumback and Rice, 1998; Guo, 2002; Di et al., 2009; Greven et al., 2011). However, it is important to distinguish our work from FANOVA methods which do not relate functional predictor(s) to the scalar response in longitudinal set up, the way we did it in this paper. To our knowledge, LPFR (Goldsmith et al., 2012) is the only published method addressing regression estimation in the longitudinal functional predictor framework.

Section 2 establishes notation and specifics on the longitudinal functional model considered in this paper. In the applications, we employ a decomposition-based penalty and this is review in Section 3.3. First, in Section 3.1, the concept of a generalized ridge estimate (or Tikhonov-Phillips estimate, Tikhonov (1963), Phillips (1962)) is discussed while Section 3.2 shows that under weak assumptions, our longitudinal generalized ridge estimate, along with its bias and precision, can be obtained in terms of generalized singular (GS) vectors. These estimates can be obtained as best linear unbiased predictors (BLUP) through mixed model equivalence and this is discussed in Section 4.1. Expressions for precision are derived in Section 4.2.

Numerical illustrations are provided in Section 5. In particular, the simulation in Section 5.1 compares LPFR with the method proposed in this paper. The simulation in subsection 5.2 evaluates the influence of sample size and relative contribution of prior spatial information on the proposed method using a decomposition-based penalty. An application based on experimental data is also illustrated in section 6. The methods discussed in this paper have been implemented in the `refund` package (Crainiceanu et al., 2012) in R through the `peer()` and `lpeer()`. Throughout the presentation we consider a single functional predictor. However, the

proposed method has a natural extension to settings with more than one functional predictor and this is discussed in the last section.

## 2. Statistical Model

We consider  $\Omega = [0, 1]$ , a closed interval in  $\mathbb{R}$  and let  $W(\cdot)$  denotes a random function in  $L^2(\Omega)$ . Let  $W_{it}(\cdot)$  denotes a functional predictor from the  $i^{\text{th}}$  subject ( $i = 1, \dots, N$ ) at the  $t^{\text{th}}$  timepoint ( $t = t_1, \dots, t_{n_i}$ ). Technically, an observed predictor arises as a discretized sampling from an idealized function, and we will assume that each observed predictor is sampled equally at  $p$  locations,  $s_1, \dots, s_p \in [0, 1]$ , with sampling that is appropriately regular and dense enough to capture informative spatial structure, as seen, for instance, in the MRS data in Section 6. Let  $w_{it} := [w_{it}(s_1), \dots, w_{it}(s_p)]^\top$  be the  $p \times 1$  vector of values sampled from the realized function  $W_{it}(\cdot)$ . Then, the observed data are of the form  $\{y_{it}; x_{it}; w_{it}\}$ , where  $y_{it}$  is a scalar outcome,  $x_{it}$  is a  $K \times 1$  column vector of measurements on  $K$  scalar predictors, and  $w_{it}$  is the sampled predictor from the  $i^{\text{th}}$  subject at time  $t$ . Denoting the true regression function at time  $t$  by  $\gamma(t, \cdot)$ , the longitudinal functional regression outcome model of interest is

$$y_{it} = x_{it}^\top \beta + \int_0^1 W_{it}(s) \gamma(t, s) ds + z_{it}^\top b_i + \epsilon_{it} \quad (2)$$

where,  $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$  and  $b_i$  is the vector of  $r$  random effects pertaining to subject  $i$  and distributed as  $N(0, \Sigma_{b_i})$ . As usual we assume that  $z_{it}$  is a subset of  $x_{it}$ ,  $\epsilon_{it}$  and  $b_i$  are independent,  $\epsilon_{it}$  and  $\epsilon_{i't'}$  are independent whenever  $i \neq i'$  or  $t \neq t'$  or both, and  $b_i$  and  $b_{i'}$  are independent if  $i \neq i'$ . Here  $x_{it}^\top \beta$  is the standard fixed effect from  $K$  univariate predictors,  $z_{it}^\top b_i$  is the standard random effect and  $\int_0^1 W_{it}(s) \gamma(t, s) ds$  is the subject/time specific functional effect. We assume that  $\gamma(t, \cdot) \in L^2(\Omega)$ , for all  $t$ .

When the association between functional predictor and response changes over time, the regression function  $\gamma(t, s)$  varies over both spatial and time domain. For example,  $\gamma(t, s)$  may vary



linearly with time,  $\gamma(t, s) = \gamma_0(s) + t\gamma_1(s)$ , or quadratically,  $\gamma(t, s) = \gamma_0(s) + t\gamma_1(s) + t^2\gamma_2(s)$ . This is in a spirit similar to a linear mixed effects model with linear or quadratic time slope (see e.g., Fitzmaurice, Laird and Ware, 2004). In general, we assume that  $\gamma(t, s)$  can be decomposed into several time-invariant component functions  $\gamma_0(s), \dots, \gamma_D(s)$  as

$$\gamma(t, s) = \gamma_0(s) + f_1(t)\gamma_1(s) + \dots + f_D(t)\gamma_D(s)$$

where,  $f_1, \dots, f_D$  are  $D$  prescribed linearly independent functions of  $t$  and  $f_d(0) = 0$  for all  $d$ ; the time component  $t$  enters into  $\gamma(t, s)$  through these terms.

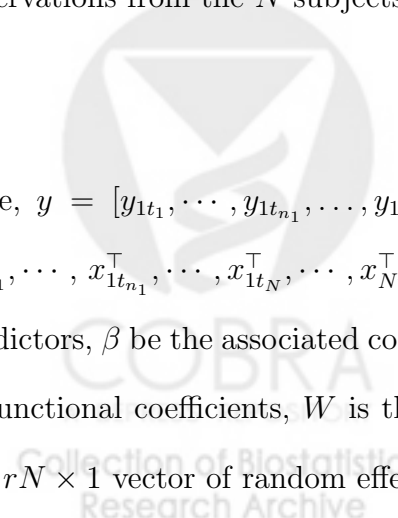
At  $t = 0$ ,  $\gamma(t, s)$  reduces to  $\gamma_0(s)$  and has the obvious interpretation of a baseline regression function pertaining to the sampling points  $s$ . When  $D = 0$ ,  $\gamma(t, s) \equiv \gamma_0(s)$  is independent of  $t$ , a situation considered by Goldsmith et al. (2012). In general, each  $f$  may be any function of  $t$  with  $f(0) = 0$ , e.g.,  $f(t) = t$  or  $t \exp(t)$ . We can rewrite the equation (2) as

$$y_{it} = x_{it}^\top \beta + \int_0^1 W_{it}(s) \{ \gamma_0(s) + f_1(t)\gamma_1(s) + \dots + f_D(t)\gamma_D(s) \} ds + z_{it}^\top b_i + \epsilon_{it}$$

In a PEER approach, the dependence of  $y_{it}$  on  $W_{it}$  is seen as a linear dependence on observations at  $p$  sampling points,  $w_{it}$ ; spatial (functional) structure is imposed directly into the estimation of  $\gamma_d = [\gamma_d(s_1), \dots, \gamma_d(s_p)]^\top$ , for  $d = 0, \dots, D$ . Combining all  $n_\bullet = \sum_{i=1}^N n_i$  observations from the  $N$  subjects obtained across all time points, we express the model as

$$y = X\beta + W\gamma + Zb + \epsilon. \tag{3}$$

Here,  $y = [y_{1t_1}, \dots, y_{1t_{n_1}}, \dots, y_{1t_N}, \dots, y_{Nt_{n_N}}]^\top$  is an  $n_\bullet \times 1$  vector of all responses,  $X = [x_{1t_1}^\top, \dots, x_{1t_{n_1}}^\top, \dots, x_{1t_N}^\top, \dots, x_{Nt_{n_N}}^\top]^\top$  is an  $n_\bullet \times K$  design matrix pertaining to  $K$  univariate predictors,  $\beta$  be the associated coefficient vector,  $\gamma = [\gamma_0^\top, \gamma_1^\top, \dots, \gamma_D^\top]^\top$  is a  $(D+1)p \times 1$  vector of functional coefficients,  $W$  is the corresponding  $n_\bullet \times (D+1)p$  design matrix. Further,  $b$  is the  $rN \times 1$  vector of random effects and  $Z$  is the corresponding  $n_\bullet \times rN$  design matrix. The



matrix  $W$  has the structure

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_N \end{bmatrix} \quad W_i = \begin{bmatrix} w_{it_1}^\top & f_1(t_1)w_{it_1}^\top & \cdots & f_D(t_1)w_{it_1}^\top \\ \vdots & \vdots & \ddots & \vdots \\ w_{it_{n_i}}^\top & f_1(t_{n_i})w_{it_{n_i}}^\top & \cdots & f_D(t_{n_i})w_{it_{n_i}}^\top \end{bmatrix}$$

### 3. Estimation of Parameters with a Penalty

The formal model in (2) is ill-posed and has no unique solution for  $\gamma$ . Common approaches to estimate a regression function in a fLM involve reducing dimension by projecting onto a subspace defined by a few K-L (empirical) eigenvectors or onto the span of a set of spline basis functions. Alternatively, our use of a generalized ridge penalty constrains the estimation of  $\gamma$  in the spatial (or  $s$ ) dimension without preliminary smoothing or explicit dimension reduction. The process encourages structure of a particular type via the choice of penalty operator. In the longitudinal (or  $t$ ) dimension,  $\gamma$  is more explicitly and severely constrained by the choice of  $f_1, \dots, f_D$ .

#### 3.1 Generalized Ridge Estimate

The model of interest described in the previous section can be written as follows:

$$y = X\beta + W\gamma + \epsilon^* \quad (4)$$

where  $\epsilon^* = Zb + \epsilon \sim N(0, V)$  and  $V = Z\Sigma_b Z^\top + \sigma_\epsilon^2 I$ . Further, assume  $L_d$  be the penalty operator for  $\gamma_d$  and  $\lambda_d^2$  be the associated tuning parameter,  $\forall d = 0, \dots, D$ . Then the penalized estimates of  $\beta$  and  $\gamma$  can be obtained by minimizing

$$\|y - X\beta - W\gamma\|_{V^{-1}}^2 + \lambda_0^2 \|\gamma_0\|_{L_0^\top L_0}^2 + \cdots + \lambda_D^2 \|\gamma_D\|_{L_D^\top L_D}^2 \quad (5)$$

here we have used the notation  $\|a\|_B^2 = a^\top B a$ , where  $B$  is a symmetric, positive definite matrix. A generalized ridge estimate of  $\beta$  and  $\gamma$  based on minimizing the above expression is

obtained as (see e.g., Ruppert, Wand and Carroll, 2003, p. 66)

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = (C^\top V^{-1} C + D)^{-1} C^\top V^{-1} y \quad (6)$$

where,  $C = [X \ W]$ ,  $D = \text{blockdiag}\{0, L^\top L\}$  and  $L = \text{blockdiag}\{\lambda_0 L_0, \dots, \lambda_D L_D\}$ .

### 3.2 Connection with the GSVD

A PEER estimate and its mean squared error (MSE) can be expressed in terms of generalized singular (GS) vectors obtained through generalized singular value decomposition (GSVD) for the cross-sectional fLM with single functional predictor and no random effect Randolph, Harezlak and Feng (2012). For background on the GSVD and its computation, see Van-Loan (1976); Paige and Saunders (1981); Bjorck (1996); Golub and Van-Loan (1996). Here, we derive a similar expression for the generalized ridge estimate  $\hat{\gamma}$  of the regression function in (6).

After some algebra, the generalized ridge estimate in (6) for  $\gamma$  can be expressed as

$$\hat{\gamma} = -A_1 X^\top V^{-1} y + A_2 W^\top V^{-1} y$$

where

$$A_1^\top = (X^\top V^{-1} X)^{-1} X^\top V^{-1} W [W^\top V^{-1} W + L^\top L - W^\top V^{-1} X (X^\top V^{-1} X)^{-1} X^\top V^{-1} W]^{-1}$$

$$A_2 = W^\top V^{-1} W + L^\top L - W^\top V^{-1} X (X^\top V^{-1} X)^{-1} X^\top V^{-1} W$$

When  $X = 0$  (a situation without any scalar predictors) or  $X^\top V^{-1} W = 0$  the generalized ridge estimation of  $\gamma$  can be put into a PEER estimation framework in terms of GS vectors, as discussed below.

With  $X = 0$  or  $X^\top V^{-1} W = 0$ , the  $\hat{\gamma}$  reduces to  $[W^\top V^{-1} W + L^\top L]^{-1} W^\top V^{-1} y$ . Moreover, in this case generalized ridge estimate of  $\beta$  becomes  $[X^\top V^{-1} X]^{-1} X^\top V^{-1} y$ . Now, if we transform  $\tilde{W} := V^{-1/2} W$  and  $\tilde{y} := V^{-1/2} y$ , we can rewrite  $L$  as

$$L = \lambda_0 \text{blockdiag} \left\{ L_0, \frac{\lambda_1}{\lambda_0} L_1, \dots, \frac{\lambda_D}{\lambda_0} L_D \right\} = \lambda_0 L^s$$

Here,  $L^s$  can be interpreted as a scaled  $L$  where scaling is done for all the tuning parameters associated with the ‘longitudinal’ part of the regression function with respect to the ‘baseline’ tuning parameter.

Set  $\tilde{p} = (D + 1)p$ , let  $m$  denote the number of rows in  $L$  and set  $c = \dim[\text{Null}(L)]$ . Further, assume that  $n_\bullet \leq m \leq \tilde{p} \leq m + n_\bullet$  and the rank of the  $(n_\bullet + m) \times \tilde{p}$  matrix  $[\tilde{W}^\top \quad (L^s)^\top]^\top$  is  $\tilde{p}$ . The following describes the GSVD of the pair  $(\tilde{W}, L^s)$ : there exist orthogonal matrices  $\mathcal{U}$  and  $\mathcal{V}$ , a nonsingular  $\mathcal{G}$  and diagonal matrices  $S$  and  $M$  such that Randolph, Harezlak and Feng (2012)

$$\begin{aligned}\tilde{W} &= \mathcal{U}S\mathcal{G}^{-1} & S &= [0 \quad S] & S &= \text{blockdiag}\{S_1, I_{\tilde{p}-m}\} \\ L^s &= \mathcal{V}\mathcal{M}\mathcal{G}^{-1} & \mathcal{M} &= [M \quad 0] & M &= \text{blockdiag}\{I_{\tilde{p}-n_\bullet}, M_1\}\end{aligned}$$

Submatrices  $S_1$  and  $M_1$  have  $\ell = n_\bullet + m - \tilde{p}$  diagonal entries ordered as

$$\begin{aligned}0 < \sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_\ell < 1 & \quad \text{where,} \quad \sigma_k^2 + \mu_k^2 = 1, \quad k = 1, \dots, \ell \\ 0 > \mu_1 \geq \mu_2 \geq \dots \geq \mu_\ell > 1\end{aligned}$$

Here, the columns  $\{g_k\}$  of  $\mathcal{G}$  are the GS vectors determined by the GSVD of the pair  $(\tilde{W}, L^s)$ .

Denote the columns of  $\mathcal{U}$  and  $\mathcal{V}$  by  $u_k$  and  $v_k$ , respectively. Now, it can be shown that

$$[W^\top V^{-1}W + L^\top L]^{-1}W^\top V^{-1} = [W^\top V^{-1}W + \lambda_0^2(L^s)^\top L^s]^{-1}W^\top V^{-1} = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{G}^\top \tilde{W}^\top V^{-1/2}$$

and consequently,  $\hat{\gamma}$  can be expressed as

$$\hat{\gamma} = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{S}^\top \mathcal{U}^\top \tilde{y} = \sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\sigma_k^2}{\sigma_k^2 + \lambda_0^2 \mu_k^2} \frac{1}{\sigma_k} u_k^\top \tilde{y} g_k + \sum_{k=\tilde{p}-c+1}^{\tilde{p}} u_k^\top \tilde{y} g_k$$

Further, the bias and variance can be expressed as

$$\begin{aligned}\text{Bias}[\hat{\gamma}] &= (I - W^\#W)\gamma = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} (\lambda_0^2 \mathcal{M}^\top \mathcal{M}) \mathcal{G}^{-1} \\ &= \sum_{k=1}^{\tilde{p}-n_\bullet} g_k \tilde{g}_k^\top \gamma + \sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\lambda_0^2 \mu_k^2}{\sigma_k^2 + \lambda_0^2 \mu_k^2} g_k \tilde{g}_k^\top \gamma \\ \text{Var}[\hat{\gamma}] &= W^\#V(W^\#)^\top = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{S}^\top \mathcal{S} (\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{G}^\top \\ &= \sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\sigma_k^2}{(\sigma_k^2 + \lambda_0^2 \mu_k^2)^2} g_k g_k^\top + \sum_{k=\tilde{p}-c+1}^{\tilde{p}} g_k g_k^\top\end{aligned}$$

where,  $W^\# = [W^\top V^{-1}W + L^\top L]^{-1}W^\top V^{-1}$  and  $\tilde{g}_k$  denotes the  $k$ th column of  $\mathcal{G}^{-T} = (\mathcal{G}^{-1})^\top = (\mathcal{G}^\top)^{-1}$ . Further, we can express bias as  $[W^\top V^{-1}W + L^\top L]^{-1}L^\top L\gamma$  which means  $\hat{\gamma}$  will be unbiased only when  $\gamma \in \text{Null}(L)$ .

For estimates obtained using this technique, the bias and variance can be expressed in terms of generalized singular vectors, provided the assumption of  $X^\top V^{-1}W = 0$  applies. In this case, one can show that  $\hat{\beta}$  is simply the generalized least squares estimate from the linear model  $y = X\beta + \epsilon^*$ , and  $\hat{\gamma}$  is the generalized ridge estimate from  $y = W\gamma + \epsilon^*$  with penalty  $L$ . That is,  $\beta$  is estimated as if  $W\gamma$  were not present, and  $\gamma$  is estimated as if  $X\beta$  were not present. The PEER estimate discussed in this section can be thought of as an extension of the estimation discussed in Randolph, Harezlak and Feng (2012) in two ways: we allow for a general covariance matrix  $V$  (for  $y$ ) and we also extend the penalty operator to apply across multiply-defined domains,  $L_0, \dots, L_D$ .

### 3.3 Decomposition based penalty

Our goal is to estimate  $\gamma$  imposing some presumed functional structure. In other words, the aim is to supplement (not to smooth or otherwise alter) the predictor function with knowledge about spatial structure in a mathematically tractable way. A common approach to incorporate spatial structure into the functional regression model is to use the strongest structure from the predictors by considering only first few K-L vectors (Hall, Poskitt and Presnell, 2001; Cardot, Ferraty and Sarda, 2003). However, we will incorporate spatial structure through an informed choice of penalty operator as proposed by Randolph, Harezlak and Feng (2012).

Let  $\gamma_d \equiv \gamma_{L_d, \lambda_d}$  be the estimate obtained from the penalty operator  $L_d$  and tuning parameter  $\lambda_d^2$ , for each  $d = 0, \dots, D$ . For example,  $L_d$  may denote  $I_p$  (a ridge penalty) or a second-order derivative penalty (giving an estimate having continuous second derivative). Alterna-

tively, with prior knowledge about potentially-relevant structure in a regression function, a decomposition-based targeted penalty can be defined in terms of a subspace defined by such structure (Randolph, Harezlak and Feng, 2012). To be precise, if it is appropriate to impose scientifically-informed constraints on the “signal” being estimated by  $\gamma$ , this prior may be implemented by encouraging the estimate to be in or near a subspace,  $\mathcal{Q} \subset L^2(\Omega)$ .

Returning to our notation that reflects functional predictors observed at  $p$  sampling points, we represent  $\mathcal{Q}$  by the range of a  $p \times J$  matrix  $Q$  whose columns are  $q_1, \dots, q_J$ . Consider the orthogonal projection  $P_Q = QQ^+$  onto the  $\text{Range}(Q)$ , where  $Q^+$  is Moore-Penrose inverse of  $Q$ . Then a decomposition-based penalty is defined as

$$L_Q = \alpha_0 P_Q + \alpha_1 (I - P_Q) \quad (7)$$

for scalars  $\alpha_0$  and  $\alpha_1$ . To see how  $L_Q$  works, let  $\tilde{\gamma}_d$  be any estimate of  $\gamma_d$ . When  $\tilde{\gamma}_d \in Sp(Q)$ , we have  $L_Q \tilde{\gamma}_d = \alpha_0 \tilde{\gamma}_d$ , but when  $\tilde{\gamma}_d \notin Sp(Q)$ , we have  $L_Q \tilde{\gamma}_d = \alpha_1 \tilde{\gamma}_d$ . The condition  $\alpha_1 > \alpha_0$  imposes more penalty for  $\tilde{\gamma}_d \notin P_Q$  compared to when  $\tilde{\gamma}_d \in P_Q$ . The weights  $\alpha_1$  and  $\alpha_0$  determine the relative strength of emphasizing  $\mathcal{Q}$  in the estimation process. Note, in particular, that taking  $\alpha_1 = \alpha_0$  results in a ridge estimate and that  $L_Q$  is invertible, provided  $\alpha_1$  and  $\alpha_0$  are nonzero. Analytical properties of estimates from this family of penalties have been discussed in Randolph, Harezlak and Feng (2012).

#### 4. Mixed model representation

Estimates of  $\beta$  and  $\gamma$  obtained by minimizing the expression in equation (5) correspond to a generalized ridge estimate. In this section we aim to construct an appropriate mixed model that minimizes the expression in equation (5). In general, the penalty,  $L$ , is not required to be invertible but for simplicity this will be assumed here. The mixed model approach provides an automatic selection of tuning parameters. REML-based estimation of the tuning parameters

has been shown to perform as well as the other criteria and under certain conditions it is less variable than GCV-based estimation (Reiss and Ogden, 2009).

#### 4.1 Estimation of parameters

Using Henderson's (1950) justification [21], one can show that the model  $y = X\beta + W\gamma + \epsilon^*$  where,  $\epsilon^* \sim N(0, V)$  and  $\gamma_d \sim N(0, \frac{1}{\lambda_d^2}(L_d^\top L_d)^{-1})$ , for each  $d = 0, \dots, D$ , minimizes the expression in equation (5) to obtain the BLUP. Thus the generalized ridge estimate of  $\beta$  and  $\gamma$  correspond to the BLUP from the following model:

$$y = X\beta + W^*\gamma^* + \epsilon$$

where,  $W^* = [W \ Z]$ ,  $\gamma^* = [\gamma^\top \ b^\top]^\top \sim N[0, \Sigma_{\gamma^*}]$  and  $\epsilon \sim N(0, \sigma_\epsilon^2 I)$  with

$$\Sigma_{\gamma^*} = \text{blockdiag}\{(L^\top L)^{-1}, \ \Sigma_b\} \quad \Sigma_b = \text{blockdiag}\{\Sigma_{b_1}, \dots, \Sigma_{b_N}\}$$

This representation allows us to estimate fixed and functional predictors simply by fitting a linear mixed model (e.g., using the `lme()` of the `nlme` package in R or PROC MIXED in SAS).

#### 4.2 Precision of Estimates

Our ridge estimate is the BLUP from equivalent mixed model, hence the variance of the estimate depends on whether the parameters are random or fixed. Randomness of  $\gamma$  is a device used to obtain the ridge estimate while  $\epsilon$  and  $b$  in our case are truly random. With this argumentation, it can be advocated that variance of estimates to be conditional on  $\gamma$ , but not on  $b$  (Ruppert, Wand and Carroll, 2003). BLUP of  $\beta$ ,  $\gamma$  and  $b$  can be expressed as (see e.g., Robinson, 1991; Ruppert, Wand and Carroll, 2003):

$$\tilde{\beta} = (X^\top V_1^{-1} X)^{-1} X^\top V_1^{-1} y \quad \tilde{\gamma} = (L^\top L)^{-1} W^\top V_1^{-1} (y - X\tilde{\beta}) \quad \tilde{b} = \Sigma_b Z^\top V_1^{-1} (y - X\tilde{\beta})$$

where,  $V_1 = V + W(L^\top L)^{-1}W^\top$ .  $\tilde{\beta}$  is an unbiased estimator of  $\beta$ , but  $\tilde{\gamma}$  is not unbiased. It is trivial to see that  $\text{Cov}(y|\gamma) = V$ . Thus, the variances of  $\tilde{\beta}$  and  $\tilde{\gamma}$ , conditional on  $\gamma$ , are:

$$\text{Cov}(\tilde{\beta}|\gamma) = (X^\top V_1^{-1}X)^{-1} X^\top V_1^{-1}V V_1^{-1}X (X^\top V_1^{-1}X)^{-1}$$

$$\text{Cov}(\tilde{\gamma}|\gamma) = A_\gamma V A_\gamma^\top \quad A_\gamma = (L^\top L)^{-1}W^\top V_1^{-1}\{V_1 - X(X^\top V_1^{-1}X)^\top\}V_1^{-1}$$

To obtain the unconditional variance, we need to replace  $V$  by  $V_1$  in the above expressions but this will overestimate the variance of the estimates. The expressions for predicted value of  $y$  and its variance are:

$$\tilde{y} = X\tilde{\beta} + W\tilde{\gamma} + Z\tilde{b} \quad \text{Cov}(\tilde{y}|\gamma) = A_y V A_y^\top$$

where,

$$A_y = [\{V_1 - WL^\top LW - Z\Sigma_b Z^\top\}^{-1}X (X^\top V^{-1}X)^{-1} X^\top V^{-1} + WL^\top LW^\top + Z\Sigma_b Z^\top]V_1^{-1}$$

Let,  $T = [1 \ f_1(t) \ \cdots \ f_d(t)] \otimes I_K$ . Then the discretized version of regression function at time  $t$  is  $\gamma_{(t)} = [\gamma(t, s_1), \cdots, \gamma(t, s_K)] = T\gamma$ . Therefore, the estimate of  $\gamma_{(t)}$  is  $\tilde{\gamma}_{(t)} = T\tilde{\gamma}$  and the estimate of its variance is  $T\text{Cov}(\tilde{\gamma}|\gamma)T^\top$ .

## 5. Simulation

We present two simulation studies to evaluate the properties of the proposed method which we refer to as ‘‘LongPEER’’. The first simulation study (Section 5.1) compares the performance of the LongPEER method with the LPFR approach. In the second simulation study, only the LongPEER method is considered. The purpose of the second simulation study is to evaluate the influence of sample size and the contribution of prior information about spatial structure (as determined by two different tuning parameters  $\alpha_0$  and  $\alpha_1$  in equation 7) on the LongPEER estimate. In both the simulation studies, the simulated predictor functions resemble the MRS data. All results summarized in this section are based on 100 simulated samples.



For each subject and visit, predictor functions were simulated independently. Predictor functions are “bumpy” curves with bumps at some prespecified locations. White noise was added to the predictor functions to account for the instrumental measurement noise. Bumpy regression functions were generated with bumps at some (but, not all) of the bump locations of the predictor function. For the simulation in Section 5.1, the regression function is assumed to be independent of time, whereas it varies with time in simulation in Section 5.2. For both the predictor and regression functions, 100 equispaced sampling points in  $[0,1]$  are used.

For the decomposition based penalty, the matrix  $L_d$  matrix is defined as follows: 1) select the discretized functions  $q_j, j = 1, \dots, J$  spanning the “non-penalized-subspace” and 2) compute  $L_d = QQ^+$ , where  $Q = \text{col}[q_1, \dots, q_J]$ . Vectors  $q_j$  are discretized functions defined to be zero except at one bump corresponding to a region in the simulated predictor functions. For ridge and second-order penalties, the matrices  $L_d$  are defined as  $I$  and  $\mathcal{D}^2$ , respectively, where  $\mathcal{D}^2 = [d_{i,j}]$  is a square matrix with entries  $d_{i,i} = d_{i,i+2} = 1$ ,  $d_{i,i+1} = -2$  and  $d_{i,j} = 0$  otherwise.

Estimation error was summarized in terms of the mean squared error (MSE) of the estimated regression function defined as  $\|\gamma - \tilde{\gamma}\|^2$ , where  $\tilde{\gamma}$  denotes the estimate of  $\gamma$ . Further, MSE was decomposed into the trace of the variance and squared norm of bias. We also calculated sum of squares of prediction error (SSPE) as  $\|y - \tilde{y}\|^2$ , where  $\tilde{y}$  denotes the estimate of  $y$ . The estimates based on the proposed methods, including the LongPEER estimate, were obtained as BLUPs from the mixed model formulation described in Section 4.1.

### 5.1 Comparison with LPFR

As previously stated, LPFR estimates a regression function that does not vary with time. Therefore, in the first set of simulations, we generated the outcomes using a time-invariant regression function (i.e.,  $\gamma(t, s) = \gamma_0(s)$ , for all  $t$ ). The following model was used to generate

the outcome data:

$$y_{it} = \beta_0 + \int_0^1 W_{it}(s)\gamma_0(s)ds + b_i + \epsilon_{it}, \quad i = 1, \dots, 100, \quad (8)$$

$$\text{where } \gamma_0(s) = \sum_{h \in H_{\gamma_0}} a_{0h} \exp \left[ -2500 * \left( \frac{h-s}{100} \right)^2 \right]$$

The bumpy predictor functions were generated from the following equation

$$w_{it}(s) = \sum_{h \in H_1} (\xi_{1h} + c_{1h}) \exp \left[ -2500 * \left( \frac{s-h}{100} \right)^2 \right] + \sum_{h \in H_2} (\xi_{2h} + c_{2h}) \exp \left[ -1000 * \left( \frac{s-h}{100} \right)^2 \right] \\ + (\xi_{31} + 0.9) \exp \left[ -1000 * \xi_{32} \left( \frac{s-50}{100} \right)^2 \right],$$

[Table 1 about here.]

where  $c_{1h}$ ,  $c_{2h}$  and  $a_{0h}$  are defined in Table 1.  $\{\xi_{1h}, h \in H_1\}$ ,  $\{\xi_{2h}, h \in H_2\}$ , and  $\xi_{31}$  were drawn independently from Uniform(0, 0.1). Also,  $\beta_0 = 0.06$ ,  $\epsilon_{it} \sim N[0, (0.02)^2]$  and  $b_i \sim N[0, (0.05)^2]$ .

[Figure 2 about here.]

We applied both LPFR (using `lpfr()` available in the `refund` package in R (Crainiceanu et al., 2012)) and the LongPEER method with different forms of structured penalties to the simulated data. To obtain LPFR estimate, the dimension of both principal components for predictor function and truncated power series spline basis for the regression function were set to default values of 30.

Table 2 displays the MSE and prediction error obtained for different estimates. The range of SSPE for the considered methods is very narrow (1.116 and 1.168) indicating that all methods perform well when the prediction metric is applied.

The LongPEER estimate has much smaller MSE when compared to the other three estimates. Ridge estimate has less bias compared to the estimate obtained using second-order difference penalty. However, due to the larger variance, in overall, the ridge estimate has larger MSE

compared to the estimate obtained using second-order difference penalty. Both the bias and variance are largest for the LPFR estimate and consequently it has the largest MSE of the four estimates.

Figure 2 displays the four estimates of the regression function. The LongPEER estimate is almost identical to the true regression function. Bump locations and magnitudes are very close to the true function and nearly as smooth as the true function. The estimated regression function obtained using ridge penalty is also quite close to the true function including the location and the magnitude of the bumps. However, the ridge estimate is more wiggly than the LongPEER estimate. The LPFR estimate and the estimate obtained using second-order difference penalty are oversmoothed in that the magnitude of the bumps is underestimated or missed altogether. The increased performance of the LongPEER estimate is due to its ability to exploit presumed structural information which is ignored by the other estimates.

[Table 2 about here.]

## 5.2 Simulation with time varying regression function

[Figure 3 about here.]

In this simulation, the regression function varies as a parametric function of time. We are not aware of other longitudinal functional regression methods for estimating a time-varying regression function so we display on the performance of LongPEER. The primary goal is to assess the effects of sample size and the relative contribution of external information as determined by  $\alpha_0$  and  $\alpha_1$  in equation 7, on the regression function estimate.

Without loss of generality, we set  $\alpha_0 = 1$  and vary  $\alpha_1$  on an exponential scale. Larger values of  $\alpha_1$  indicate greater emphasis of prior information on the estimation process. The model considered here is similar to that described in Section 5.1 with the exception that  $\gamma(t, s) =$

$\gamma_0(s) + t \gamma_1(s)$ . The function  $\gamma_0(s)$  is defined in equation (9) and  $\gamma_1(s)$  is of the form

$$\gamma_1(s) = \sum_{h \in H_{\gamma_1}} a_{1h} \exp \left[ -2500 * \left( \frac{h-s}{100} \right)^2 \right]$$

where the value of  $h$  and  $a_{1h}$  are listed in Table 1 and  $\beta_0 = 0.06$ . Realizations of functional predictors were generated as described in section 5.1.

For each simulation, an appropriate  $\sigma_\epsilon^2$  was chosen to ensure that the squared multiple correlation coefficient  $R^2 = s_y^2 / [s_y^2 + \sigma_\epsilon^2]$  is 0.9. Here,  $s_y^2 = \frac{1}{4} \sum_{t=1}^4 \frac{1}{N-1} \sum_{i=1}^N (y_{it} - \bar{y}_{.t})^2$  denotes the average sample variance in the set  $\{y_{it} - \epsilon_{it} : i = 1, \dots, N; t = 1, \dots, 4\}$  with  $\bar{y}_{.t} = \frac{1}{N} \sum_{i=1}^N y_{it}$ .

Results in terms of SSPE and MSE are displayed graphically in Figure 3. As the sample size increases, both SSPE and MSE of  $\tilde{\gamma}_0$  and  $\tilde{\gamma}_1$  decrease, providing empirical evidence that the estimates obtained using decomposition based penalty are consistent. For a given sample size, we observe a decrease in MSE as the value of  $\alpha_1$  increases up to 100. However, no further improvement is seen beyond  $\alpha_1 = 100$ , suggesting that choice of  $\alpha_1$  has little or no influence the estimate provided it is sufficiently large.

The estimated  $\gamma_0$  and  $\gamma_1$  are displayed in Figure 4. The estimated  $\gamma_0(\cdot)$  and  $\gamma_1(\cdot)$  are nearly identical to the simulated true functions, even for sample size  $N = 100$ . Further inspection reveals a small estimation bias at the points where both  $\gamma_0$  and  $\gamma_1$  have peaks, but this bias gradually disappeared as the sample size increased to  $N = 400$ .

[Figure 4 about here.]

## 6. MRS study application

[Figure 5 about here.]

We applied LongPEER to study the association of metabolite spectra obtained from basal ganglia on the global deficit score (GDS) in a longitudinal study of late stage HIV patients

(Harezlak et al., 2011). GDS is often used as a continuous measure of neurocognitive impairment (e.g., Carey et al., 2004) and a large GDS score is an indicator of high degree of impairment. The collected MRS spectra are composed of the combination of pure metabolite spectra, instrument noise and baseline profile. The spectra were sampled at  $n = 399$  distinct sampling points each. We collected a total of  $n_{\bullet} = 306$  observations from  $N = 114$  subjects. The longitudinal observations for each subject were within 3 years from baseline. The number of observations per subject ranged from 1 to 5 with median 3.

Information on spectra obtained from nine pure metabolites was available and hence we were able to use this to define a decomposition penalty  $L_Q$  as in equation (7) (with  $\alpha_1 = 1000$ ). Spectra of basal ganglia at baseline and pure metabolite spectra are displayed in Figure 1. The pure metabolite spectra include spectra of Creatine (Cr), Glutamate (Glu), Glucose (Glc), Glycerophosphocholine (GPC), *myo*-Inositol (Ins), N-Acetylaspartate (NAA), N-Acetylaspartylglutamate (NAAG), *scyllo*-Inositol (Scyllo) and Taurine (Tau). It can be verified that at each locations of major bumps in basal ganglia spectra (left panel), there are bumps in at least one of pure metabolite profiles (right panel). We fitted the following model to data

$$y_{it} = \beta_0 + \beta_1 t + \int_{\Omega} W_{it}(s)\gamma(t, s)ds + b_i + \epsilon_{it}, \quad (9)$$

where,  $\gamma(t, s) = \gamma_0(s) + t \gamma_1(s)$  and  $y_{it}$  and  $W_{it}(\cdot)$  are the GDS and basal ganglia spectrum for subject  $i$  at time  $t$ , respectively. We assume that  $\epsilon_{it} \sim N(0, \sigma_{\epsilon}^2)$  and  $b_i$  is the subject-specific random intercept distributed as  $N(0, \sigma_b^2)$ . The estimates are obtained as the BLUP from the mixed model formulation described in Section 4.1.

Estimates of  $\sigma_{\epsilon}^2$  and  $\sigma_b^2$  were 0.08037 and 0.33115, respectively. Figure 5 displays the estimates of  $\gamma_0(\cdot)$  and  $\gamma_1(\cdot)$  with pointwise 95% confidence bands. To make the interpretation easier, we also include the selected pure metabolite spectra. These figures reveal that  $\hat{\gamma}_0(\cdot)$  (the ‘baseline’

part of the regression function) is different from zero at the locations where at least one of the pure metabolites Cr, Glu, NAA, NAAG and Scyllo has a bump. Similarly, each non-zero part of  $\hat{\gamma}_1(\cdot)$  (the ‘longitudinal’ part of the regression function) coincides with bump locations of one or more pure metabolite profiles of Cr, Glu, NAA, GPC and Ins.

Pointwise confidence interval for  $\gamma_0(\cdot)$  and  $\gamma_1(\cdot)$  contain the “zero” line over most of the intervals of interest. The estimated  $\gamma_0$  is significant in the region  $s \in (0.4, 0.5)$  and estimated  $\gamma_1$  is significant in a wider region of  $s \in (0.4, 0.6)$ . To be precise, the peaks of the both  $\hat{\gamma}_0(\cdot)$  and  $\hat{\gamma}_1(\cdot)$  are significant at the locations where the at least one of the pure metabolite profiles of NAA and Glu have bumps.

The finding of the significant negative ‘longitudinal’ effect of NAA is worth noticing. This suggests that GDS increases as the NAA concentration decreases in basal ganglia. This finding is consistent with the studies where reduced concentration of NAA has been found to be associated with decrease in neuronal mass (Christiansen et al., 1993; Lim and Spielman, 1997; Soares and Law, 2009).

## 7. Discussion

We have proposed a novel estimation method for longitudinal functional regression and derived some properties of the coefficient function estimate. Within this framework, the LongPEER method is the first to allow a coefficient function to *vary with time*. It extends the scope of generalized ridge regression to the realm of longitudinal data. The approach may be viewed as an extension of longitudinal mixed effects models, replacing scalar predictors by functional predictors. Advantages of this method include: 1) a framework that allows the regression function to vary over time; 2) the ability to incorporate structural information into the estimation process; and 3) easy implementation through the linear mixed model equivalence.

The emphasis here is on a general statistical framework for incorporating scientific knowledge into the estimation process when both the scalar response and predictor functions are observed longitudinally. In absence of prior information, one may impose more vaguely-defined constraints—such as smoothness or re-weighted empirical subspaces—to estimate the coefficient function. An advantageous use of specific prior information is illustrated in the first simulation where smoothness constraints or a spline basis representations perform poorly. The second simulation shows a reduction in estimation error as sample size increases; this empirically suggests the consistency for such an estimate, provided the information about the regression function structure is correct.

Solutions to the generalized ridge regression problem can be expressed in many forms. The linear mixed model equivalence provides an easy computational implementation as well as an automatic choice of the tuning parameters using REML criterion. The GSVD provides algebraic insight and a convenient way to derive the bias and variance expressions of the estimates. Another natural way to obtain the regression function estimates is by using Bayesian equivalence (see e.g., Robinson, 1991) with the informative priors quantifying the available scientific knowledge.

One of the natural extensions of our work can incorporate multiple functional predictors. For example, we can observe two functional predictors  $W_t^{(1)}(\cdot)$  and  $W_t^{(2)}(\cdot)$  with  $\gamma^{(1)}(t, \cdot)$  and  $\gamma^{(2)}(t, \cdot)$  their associated coefficient functions, respectively. Furthermore, we can express  $\gamma^{(1)}(t, s) = \gamma_0^{(1)}(s) + f_1^{(1)}(t)\gamma_1^{(1)}(s) + \dots + f_d^{(1)}(t)\gamma_d^{(1)}(s)$  and  $\gamma^{(2)}(t, s) = \gamma_0^{(2)}(s) + f_1^{(2)}(t)\gamma_1^{(2)}(s) + \dots + f_d^{(2)}(t)\gamma_d^{(2)}(s)$ . If  $W^{(1)}$  and  $W^{(2)}$  represent design matrices for the two functional predictors, then we can estimate  $\gamma^{(1)}(t, \cdot)$  and  $\gamma^{(2)}(t, \cdot)$  by finding the BLUP estimate of  $\gamma^{(1)}$  and  $\gamma^{(2)}$  from the mixed model,  $y = X\beta + W^{(1)}\gamma^{(1)} + W^{(2)}\gamma^{(2)} + Zb + \epsilon$ . The simplified formula for bias and variance derived in Section 3.2 still holds with an additional assumption  $(W^{(1)})^\top V^{-1}W^{(2)} = 0$ .

As presented here, the method addresses models having a continuous scalar outcome, but allowing for either binary and count responses is of interest. Indeed, an important problem that arises in MRS data is that of understanding the neurocognitive impairment status of HIV patients, defined as a binary variable, based on functional predictors collected over time. Extending our approach to these general settings is possible and currently being pursued.

**Acknowledgements:** The authors thank Dr B. Navia who provided the MRS data used as an example in the manuscript. Partial research support was provided by the National Institutes of Health grants U01-MH083545, R01-CA126205 and U01-CA086368.

## References

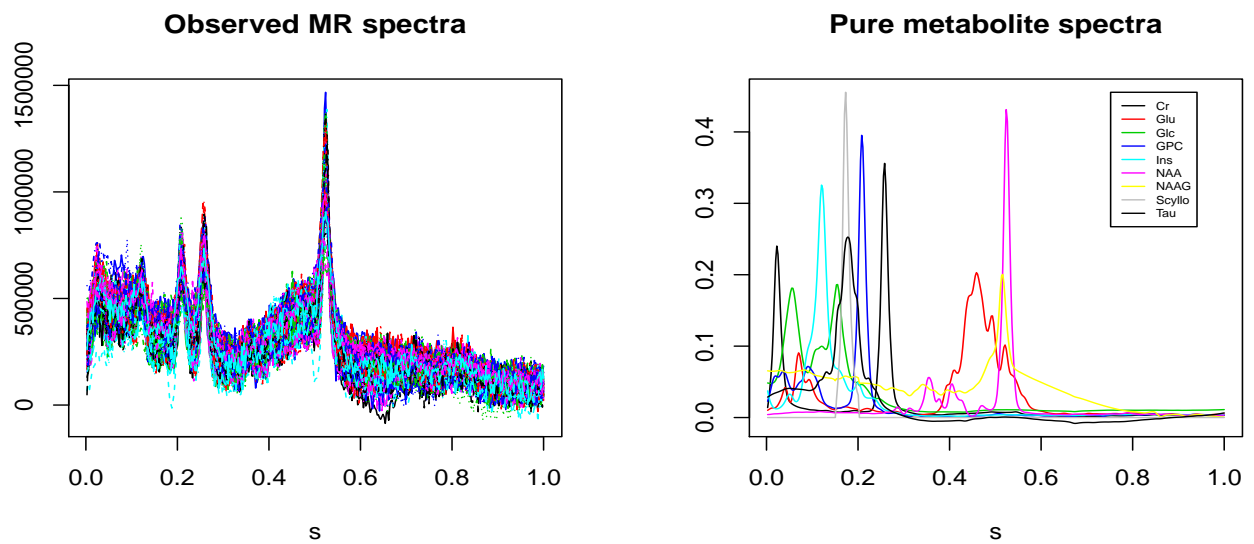
- Bjorck, A. (1996). Numerical methods for least square problems, 1st edition. Philadelphia: SIAM.
- Brumback, B. and Rice J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of American Statistical Association* **93(443)**, 961–976.
- Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34(5)**, 2159–2179.
- Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45(1)**, 11–22.
- Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.
- Cardot, H., Crambes, C., Kneip, A. and Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics Data Analysis* **51(10)**, 4832–4848.
- Carey, C., Woods, S., Gonzalez, R., Conover, E., Marcotte, T., Grant, I. and Heaton, R. (2004).



- Predictive validity of global deficit scores in detecting neuropsychological impairment in HIV infection. *Journal of Clinical and Experimental Neuropsychology* **26(3)**, 307–319.
- Christiansen, P., Toft, P., Larsson, H., Stubgaard, M. and Henriksen, O. (1993). The concentration of N-acetyl aspartate, creatine + phosphocreatine, and choline in different parts of the brain in adulthood and senium. *Magnetic resonance imaging* **11(6)**, 799–806.
- Crainiceanu, C., Reiss, P., Goldsmith, A., Huang, L., Huo, L. and Scheipl F. (2012). *refund: Regression with Functional Data* (R package version 0.1-6). [<http://CRAN.R-project.org/package=refund>].
- Di, C., Crainiceanu, C., Caffo, B. and Punjabi, N. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* **4**, 458–288.
- Fan J., and Zhang, J. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62(2)**, 303–322.
- Faraway, J. (1997). Regression analysis for a functional response. *Technometrics* **39(3)**, 254–261.
- Fitzmaurice, M., Laird, G. and Ware, J. (2004). *Applied Longitudinal Analysis*, 1st edition. Wiley Series in probability and statistics.
- Goldsmith, J., Bobb, J., Crainiceanu, C., Caffo, B. and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20(4)**, 830–851.
- Goldsmith, J., Bobb, J., Crainiceanu, C., 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 (Applied Statistics)* **61(3)**, 453–469.
- Golub G. and Van-Loan C. (1996) *Matrix computations*, Baltimore: John Hopkins University Press.
- Greven, S., Crainiceanu, C., Caffo, B. and Reich, D. (2011). Longitudinal functional principal

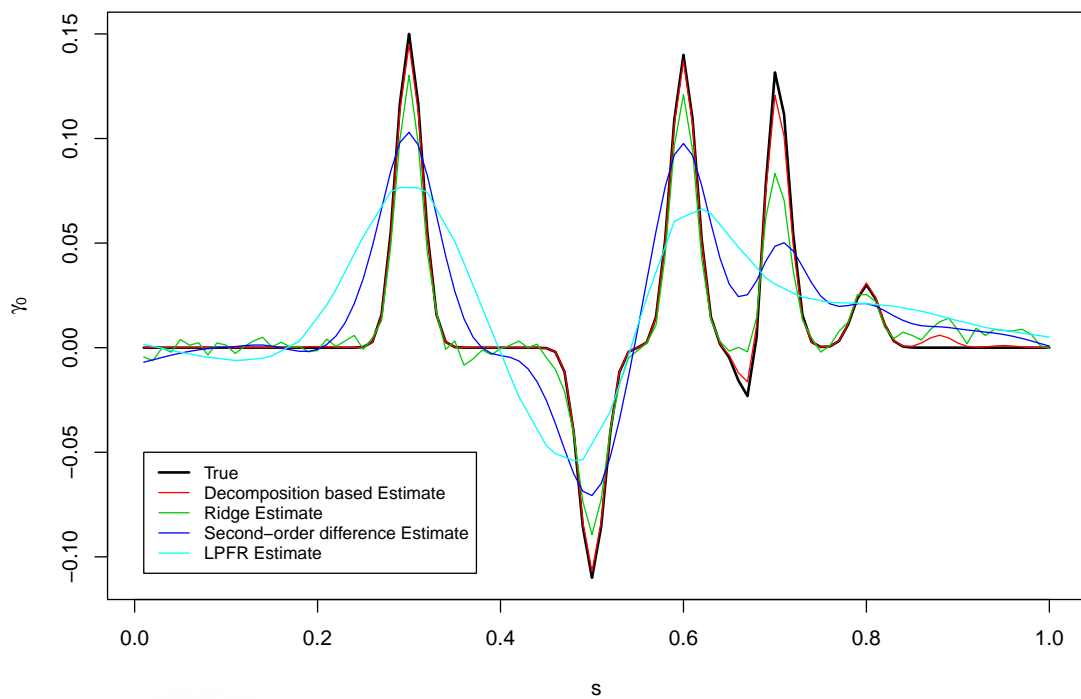
- component analysis. Recent Advances in Functional Data Analysis and Related Topics, 1st Edition. Physica-Verlag HD pp.149–154.
- Guo,W. (2002). Functional mixed effects models. *Biometrics* **58(1)**, 121–128.
- Hall, P., Poskitt, D., and Presnell. B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43(1)**, 1–9.
- Harezlak, J., Buchthal, S., Taylor, M., Schifitto, G., Zhong, J., Daar, E., Alger, J., Singer, E., Campbell, T. and Yiannoutsos, C. (2011). Persistence of HIV-associated cognitive impairment, inflammation, and neuronal injury in era of highly active antiretroviral treatment. *AIDS* **25**, 625–633.
- Henderson,C. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21(1)**.
- James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64(3)**, 411–432.
- Lim, K. and Spielman, D. (1997). Estimating NAA in cortical gray matter with applications for measuring changes due to aging. *Magnetic resonance in medicine* **37(3)**, 372–377.
- Morris, J. and Carroll, R. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68(2)** 179–199.
- Müller, H. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32(2)**, 223–240.
- Müller, H. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33(2)**, 774–805.
- Paige, C. and Saunders, M. (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis* **18(3)**, 398–405.
- Phillips, D. L. (1962), A technique for the numerical solution of certain integral equations of the first kind, *J. Associat. Comput. Mach.*, **9**, 84–97.

- Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* **53(3)**, 539–572.
- Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, 1st edition. Springer-Verlag, Berlin.
- Randolph, T., Harezlak, J. and Feng, Z. (2012). Structured penalties for functional linear models – partially empirical eigenvectors for regression. *Electronic Journal of Statistics* **6**, 323–353.
- Reiss, P. and Ogden, R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71(2)**, 505–523.
- Robinson, G. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science* **6(1)**, 15–32.
- Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*, 1st edition. Cambridge Series in Statistical and Probabilistic Mathematics.
- Van-Loan, C. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis* **13(1)**, 76–83.
- Soares, D. and Law, M. (2009). Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clinical radiology* **64(1)**, 12–21.
- Tikhonov, A. N. (1963), Solution of incorrectly formulated problems and the regularization method, *Dokl. Akad. Nauk SSSR*, **151(3)**, 501–504 (in Russian); *English transl.: Soviet Math. Dokl.*, **4(4)**, 1035–1038.
- Yao, F. and Müller, H. (2010). Functional quadratic regression. *Biometrika* **97(1)**, 49–64.

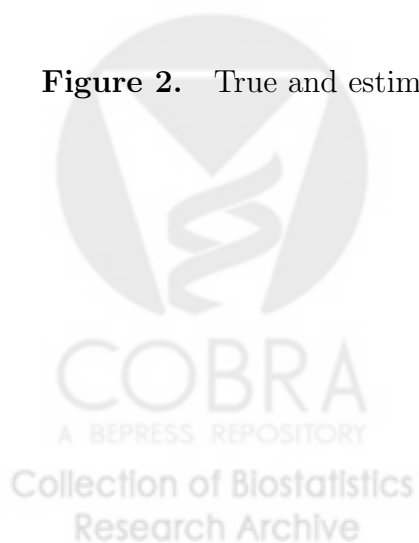


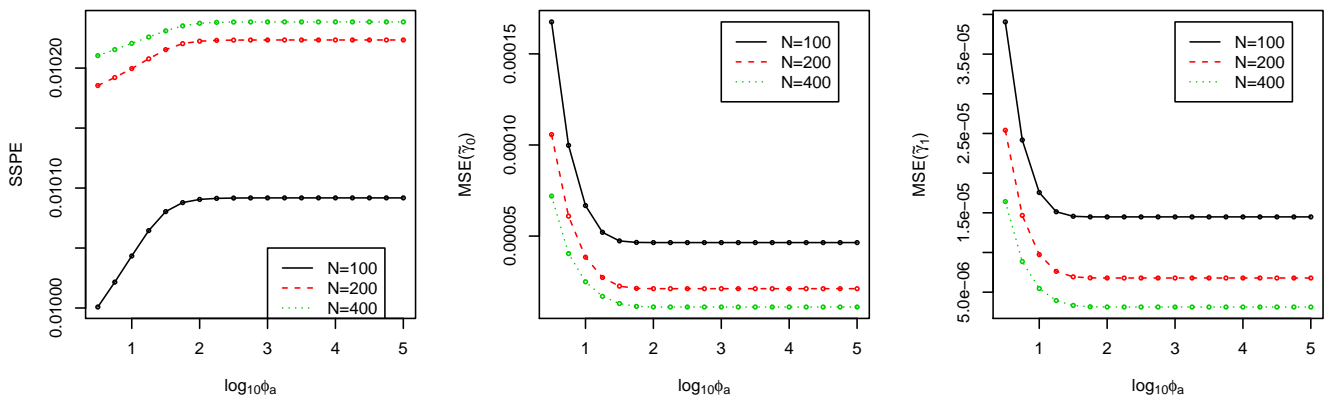
**Figure 1.** Observed MR spectra from basal ganglia for 114 subjects collected at baseline ( $t = 0$ ) are shown in the left panel and the 9 pure metabolite profiles in the right panel.





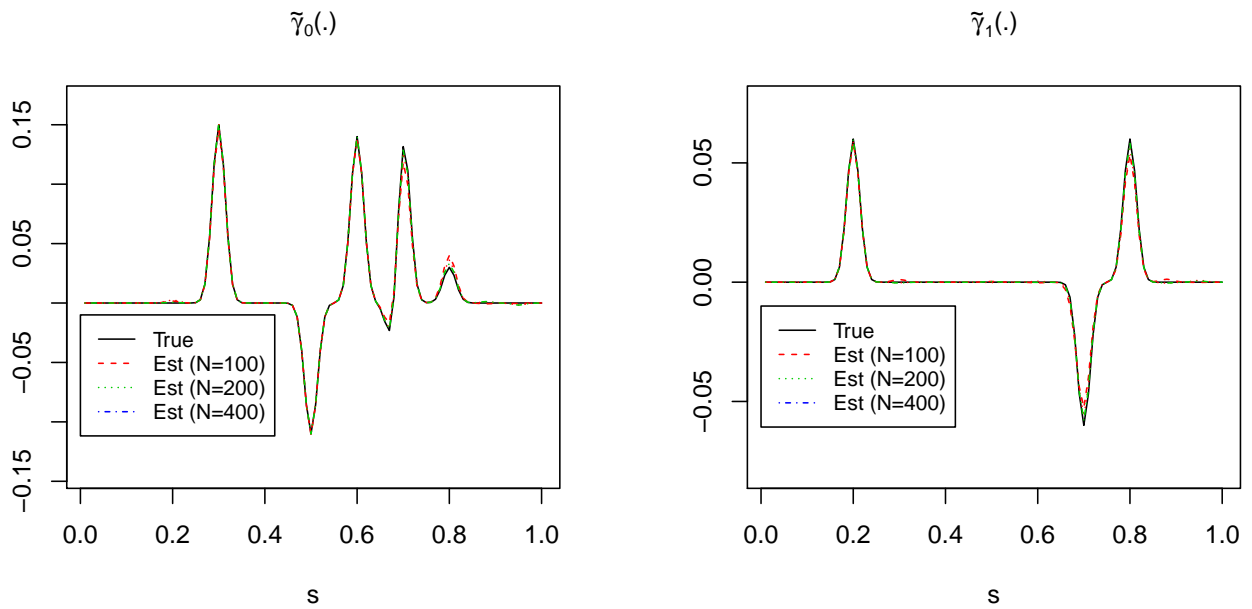
**Figure 2.** True and estimated regression for simulation study in Section 5.1.





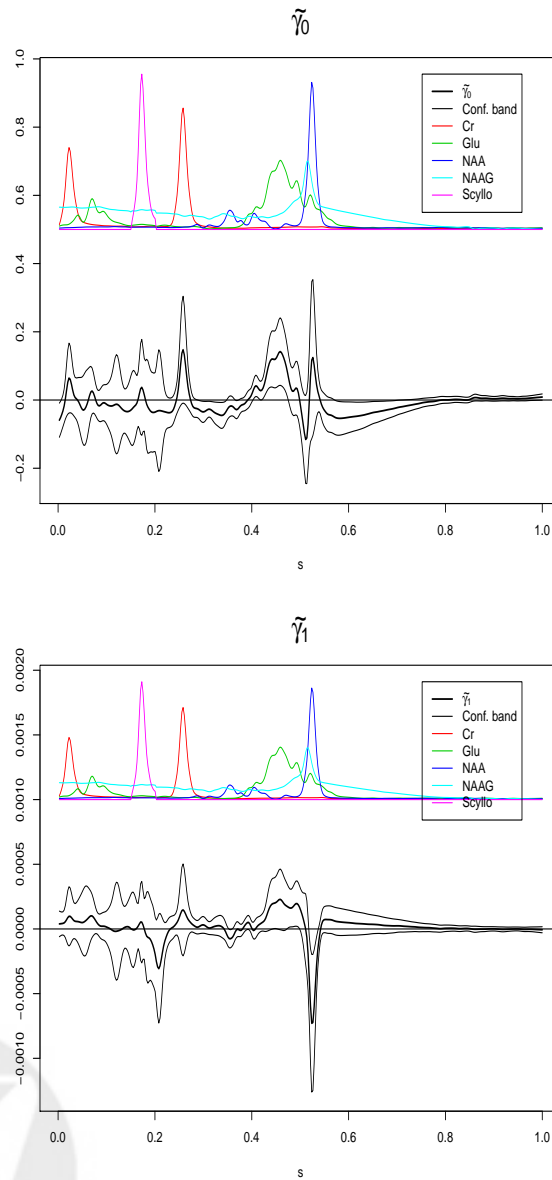
**Figure 3.** SSPE and MSE for simulation study in Section 5.2.





**Figure 4.** Regression function estimates for the simulation study described in Section 5.2. Left panel shows the estimate of  $\tilde{\gamma}_0(\cdot)$  and right panel the estimate of  $\tilde{\gamma}_1(\cdot)$





**Figure 5.** Estimates of the regression function (with 95 % point-wise confidence band) for the analysis described in Section 6. The top panel shows the estimated  $\gamma_0(\cdot)$  and bottom panel the estimate of  $\gamma_1(\cdot)$ . Scaled pure metabolite profiles are also shown on both plots.



**Table 1**

Values of  $c_{1h}$ ,  $c_{2h}$ ,  $a_{0h}$  and  $a_{1h}$  for generation of predictor and regression function in simulation studies in Sections 5.1 and 5.2.

$h \in H_1$		$h \in H_2$		$h \in H_{\gamma_0}$		$h \in H_{\gamma_1}$	
$h$	$c_{1h}$	$h$	$c_{2h}$	$h$	$a_{0h}$	$h$	$a_{1h}$
20	0.05	70	0.6	30	0.15	20	0.06
30	0.10	80	0.5	50	-0.11	70	0.06
60	0.08	88	0.3	60	0.14	80	0.06
68	0.05	95	0.4	68	0.05		
				70	0.15		
				80	0.03		



**Table 2**

*Estimation and prediction errors obtained using Ridge penalty, second-derivative penalty ( $D^2$ ), LPFR and LongPEER based on 100 simulated datasets. The sample size is set at  $N=100$  and number of observations at  $n_i = 4$ .*

	LPFR	Ridge penalty	$D^2$ penalty	Decomposition penalty
MSE( $\gamma_0$ )	0.2227	0.0980	0.0820	0.0095
$\ \text{Bias}\ ^2(\gamma_0)$	0.1065	0.0099	0.0551	0.0006
Trace of Variance( $\gamma_0$ )	0.1162	0.0881	0.0269	0.0089
SSPE of $Y$	1.1685	1.1157	1.1579	1.1535

