Nested Hierarchical Functional Data Modeling and Inference for the Analysis of Functional Plant Phenotypes

Yuhang Xu
Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583

Yehua Li and Dan Nettleton
Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011

Correspondence: yehuali@iastate.edu

Abstract

In a plant science Root Image Study, the process of seedling roots bending in response to gravity is recorded using digital cameras, and the bending rates are modeled as functional plant phenotype data. The functional phenotypes are collected from seeds representing a large variety of genotypes and have a three-level nested hierarchical structure, with seeds nested in groups nested in genotypes. The seeds are imaged on different days of the lunar cycle, and an important scientific question is whether there are lunar effects on root bending. We allow the mean function of the bending rate to depend on the lunar day and model the phenotypic variation between genotypes, groups of seeds imaged together, and individual seeds by hierarchical functional random effects. We estimate the covariance functions of the functional random effects by a fast penalized tensor product spline approach, perform multi-level functional principal component analysis (FPCA) using the best linear unbiased predictor of the principal component scores, and improve the efficiency of mean estimation by iterative decorrelation. We choose the number of principal components using a conditional Akaike Information Criterion and test the lunar day effect using generalized likelihood ratio test statistics based on the marginal and conditional likelihoods. We also propose a permutation procedure to evaluate the null distribution of the test statistics. Our simulation studies show that our model selection criterion selects the correct number of principal components with remarkably high frequency, and the likelihood-based tests based on FPCA have higher power than a test based on working independence.

Keywords: Akaike information criterion; Functional data analysis; Generalized likelihood ratio test; Penalized splines; Principal components; Permutation test.

Short title: Hierarchical Functional Data Analysis

1 Introduction

Technological advances are allowing plant scientists to measure increasingly complex phenotypes, including dynamic phenotypes that can be captured by recording the growth or development of plants over time. From the statistical point of view, it can be natural and advantageous to consider phenotypes measured over time as functional data. We develop methods of modeling and inference for functional phenotypic data with a nested hierarchical structure that arises naturally in the data acquisition process. Although our methods are quite general, we focus our presentation on a study involving gravitropism.

Gravitropism is the growth movement of a plant in response to gravity. Charles Darwin was one of the first scientists to document that plant roots show positive gravitropism, i.e., they grow in the same direction as gravity. Root gravitropism is an active research area in botany and agriculture because of its importance for plant growth and development (Noh et al., 2003). In a recent Root Image Study (RIS) conducted by Edgar Spalding's lab at the University of Wisconsin-Madison, researchers studied root gravitropism of maize seeds from various genotypes. There were 1762 seeds observed in the RIS, drawn from 235 genotypes with up to 10 seeds for each genotype. Within each genotype, seeds were distributed across up to two dishes. Each dish contained up to 5 seeds and was monitored by one digital camera. Figure 1 (a) shows an image of a group of seeds of the same genotype in the same dish. We refer to a dish of seeds as a 'file' in this paper because images of these seeds were recorded in one camera file. There were a total of 457 files in the RIS data. The lab used 7 cameras and monitored multiple dishes each day. The entire data collection process was completed in about 4 months. The data have a natural three-level nested hierarchical structure, with seeds nested in files and files nested in genotypes. In this paper, genotypes, files and seeds are also referred to as levels one, two and three of the hierarchy, respectively.

Seeds were initially positioned with their root tips approximately perpendicular to the force of gravity, as shown in Figure 1 (a). During the imaging process, seed root tips

turned downward due to root gravitropism. The root tip angle of each seed, with respect to a horizontal line, was recorded by a camera every 3 minutes for a total duration of 3 hours. To study root bending rates, we convert each raw angle to change in angle from the previous time point. Figure 1 (b) shows these angle changes as a function of time for all maize seeds. As shown in the plot, root tip bending mostly occurs during the first 1.5 hours. Therefore, we focus on this most informative part of the data and only model the bending rate process within the first 1.5 hours. Figure 1 (c) illustrates the variation between two randomly selected genotypes, and panel (d) illustrates the variation between two files within the same genotype. The bending rate is a time-dependent process and hence naturally modeled as functional data (Ramsay and Silverman, 2005), but these functions have a nested correlation structure inherited from the experimental design. We model the genotype, file and seed effects as nested functional random effects, where each can be represented by a Karhunen-Loève expansion. This approach is often referred to as hierarchical or multi-level functional principal component analysis (FPCA).

Over the four-months duration of the study, the bending rates of root tips were examined on different lunar days corresponding to different moon phases. In many cultures throughout history, agricultural activities such as planting and harvesting have been scheduled according to moon phases, and the gravitational pull of the moon is known to vary throughout the lunar cycle. Thus, it is natural to investigate potential lunar effects on gravitropism by modeling and testing for lunar effects on root tip bending of maize seeds. We model the bending rate trajectories of the seeds as hierarchical functional data, allowing the mean function of the bending rate to depend on both the lunar day bending rates were measured and the time point during the measurement process.

In functional data analysis (FDA), data are usually curves or images. FPCA has become one of the most important modeling and dimension reduction tools in FDA. Some classic work on FPCA methodology includes Yao and Lee (2006) and James et al. (2000), and

the theoretical properties of FPCA are investigated by Hall et al. (2006) and Li and Hsing (2010). However, these papers only study independent data curves. As technology advances, hierarchical functional data are becoming increasingly available. Di et al. (2009) studies two-level hierarchical functional data from a sleep heart health study, where each subject yields multiple electroencephalographic (EEG) curves from multiple hospital visits. Li et al. (2015) analyzes three-level functional data from an exercise intervention trial where real time measurements on the activity level (measured by metabolic units or METs) have a subject-week-day three-level hierarchical structure. Other related papers include Zhou et al. (2008) on paired functional data, Zhou et al. (2010) on spatially correlated hierarchical functional data, Serban and Jiang (2012) on multilevel functional clustering analysis, Serban et al. (2013) and Goldsmith et al. (2015) on two-level binary functional data, and Shou et al. (2015) on structured functional principal component analysis.

Compared with existing methodology on analyzing hierarchical functional data, our main contributions are the following. First, we estimate the mean function by anisotropic bivariate penalized splines and adopt a fast hierarchical FPCA algorithm, which directly estimates the covariance functions of the functional random effects using a method-of-moment approach based on penalized tensor product B-spline smoothing (Ruppert et al., 2003). This approach can handle large functional data sets and does not involve computationally intensive EM iterations as in Li et al. (2015). Compared with Di et al. (2009), we provide more detailed smoothing strategies to eliminate measurement errors, and we estimate the principal component scores using the empirical best linear unbiased predictor (BLUP) method rather than time consuming Markov Chain Monte Carlo (MCMC). To improve the estimation efficiency of the mean function, we adopt an iterative decorrelation procedure similar to that of Yao and Lee (2006) for uni-level functional data.

Second, we propose a new method to choose the number of principal components based on a conditional Aikaike information criterion (AIC). Selecting the number of principal com-

ponents is one of the most important model selection issues in FPCA. The current literature on hierarchical FPCA, including Di et al. (2009) and Li et al. (2015), selects the number of components subjectively using an ad hoc "percentage of variation explained" (PVE) method. In contrast, our proposed method, which extends the recent work of Li et al. (2013) for independent functional data to the hierarchical setting, is completely data-driven and vastly outperforms the existing ad hoc methods in our simulations studies.

Third, and most importantly, we propose new test procedures on the mean function based on generalized likelihood ratio (GLR) test statistics (Fan et al., 2001). There is relatively little work on nonparametric inference for hierarchical functional data, with the exception of Li et al. (2015) which considers a Wald test on the mean parameters. In our model, in order to test the lunar effects, we compare two models: in the full model, the mean bending rate is a bivariate function that depends on both the recording time point during the measurement process and the lunar day of measurement; in the reduced model, the mean function is univariate and does not depend on the lunar day. To the best of our knowledge, nonparametric tests of univariate nulls versus bivariate alternatives have not been investigated in the literature. We propose three versions of GLR tests based on the marginal likelihood, conditional likelihood and working independence, respectively. We propose a simple permutation strategy to estimate the null distribution of these test statistics, and investigate the empirical size and power of the proposed test procedures.

The rest of paper is organized as follows. We describe the hierarchical functional data model in Section 2 and the estimation procedure in Section 3. We address the model selection and inference issues in Section 4, illustrate the proposed methods by simulation studies in Section 5, and analyze our motivating data set in Section 6. Some concluding remarks are provided in Section 7. The online supplementary materials contain additional simulation results and graphs.

2 Hierarchical Functional Data Modeling

Let $Y_i(t)$ be the bending rate of the ith seed at time $t \in \mathcal{T}$, where $\mathcal{T} = [0, 1.5]$ is the observation time period and $i = 1, 2, \ldots, n$. Denote s_i as the lunar day on which the ith seed was measured, where $s_i \in \mathcal{S}$ and $\mathcal{S} = [1, 30]$. For each seed, we also observe a covariate vector $\boldsymbol{X}_i$. In the RIS data, $\boldsymbol{X}_i$ consists of variables that indicate which of the 7 cameras was used to collect the data for the ith seed. The effects associated with these indicator variables are modeled as fixed effects that do not drift over time. Because the data are collected from a large number of genotypes, and seeds are measured in groups (i.e., files), the effects of these factors are modeled as random effects that evolve over time. There are a total of G genotypes documented in F files. We use g, f and i for indices of genotype, file and seed respectively. With a slight abuse of notation, we also use $g(\cdot)$ and $f(\cdot)$ as index functions; e.g., $g(i)$ and $f(i)$ are the genotype and file number of the ith seed respectively. Similar notation was used by Brumback and Rice (1998) for ANOVA modeling of nested curve data. For the gth genotype, define $n_g = \#\{i : g(i) = g\}$ as the number of seeds with genotype g and $F_g = \#\{f : g(f) = g\}$ as the number of files with genotype g.

We model the RIS data by the following hierarchical functional data model

$$Y_i(t) = \mu(s_i, t) + \boldsymbol{X}_i'\boldsymbol{\alpha} + Z_{1,g(i)}(t) + Z_{2,f(i)}(t) + Z_{3,i}(t) + \epsilon_i(t), \tag{1}$$

where $\mu(s_i, t)$ is the mean function of the bending rate under a baseline camera setup, $\boldsymbol{\alpha}$ represents fixed effects of cameras, $Z_{1,g(i)}(t)$, $Z_{2,f(i)}(t)$, and $Z_{3,i}(t)$ are random processes representing the functional random effects of genotype $g(i)$, file $f(i)$, and seed i, respectively, and $\epsilon_i(t)$ is a white noise measurement error with variance σ^2. We assume that $Z_l(t)$, $l = 1, 2, 3$, are zero-mean random processes in time with covariance functions

$$\mathcal{K}_l(t_1, t_2) = \text{cov}\{Z_l(t_1), Z_l(t_2)\}. \tag{2}$$

Furthermore, we assume that $Z_1(t)$, $Z_2(t)$, $Z_3(t)$ and $\epsilon(t)$ are mutually independent.

The covariance functions in (2) are positive semi-definite bivariate functions with spectral decomposition

$$\mathcal{K}_l(t_1, t_2) = \sum_{k=1}^{\infty} \omega_{l,k} \psi_{l,k}(t_1) \psi_{l,k}(t_2), \quad l = 1, 2, 3, \tag{3}$$

where $\omega_{l,k}$ are the eigenvalues of $\mathcal{K}_l$ in a descending order and $\psi_{l,k}$ are the corresponding eigenfunctions. The eigenfunctions are orthonormal in the sense that $\int_{\mathcal{T}} \psi_{l,k}(t)\psi_{l,k'}(t)\mathrm{d}t$ is 1 if $k = k'$ and is 0 otherwise. By the Karhunen-Loève expansion,

$$Z_l(t) = \sum_{k=1}^{\infty} \xi_{l,k} \psi_{l,k}(t), \tag{4}$$

where $\xi_{l,k}$ are zero-mean, uncorrelated random variables, known as the principal component scores of Z_l such that $\mathrm{var}(\xi_{l,k}) = \omega_{l,k}$. In practice, the Karhunen-Loève expansions in (4) need to be truncated at finite orders. Suppose that the process $Z_l(t)$ can be characterized by p_l principal components, $l = 1, 2, 3$. These numbers determine the model complexity of the longitudinal correlation structure, so selection of these numbers plays a key role in FPCA. We will propose a data-driven model selection method in Section 4.1 to choose these numbers.

3 Estimation Procedure

3.1 Estimating the mean and covariance functions

We first estimate the mean function $\mu(s, t)$ by penalized tensor product spline regression. In the real data, we have $m_T = 31$ observation points in $\mathcal{T}$ and $m_S = 30$ lunar days in $\mathcal{S}$. Because the lunar effects are periodic by nature with an approximate 30-day cycle but the effect in t does not have such a constraint, we use different basis functions in the two domains. Define B-spline basis functions $\boldsymbol{B}_T(t) = (B_{T1}, \ldots, B_{TK_T})'(t)$ on $\mathcal{T}$ with equally-spaced interior knots, and let $\boldsymbol{B}_L(s) = (B_{L1}, \ldots, B_{LK_L})'(s)$ be Fourier basis functions with a 30-day period on $\mathcal{S}$. Define the tensor product basis on $\mathcal{S} \times \mathcal{T}$ as $\boldsymbol{B}_\mu(s, t) = \boldsymbol{B}_L(s) \otimes \boldsymbol{B}_T(t)$

where $\otimes$ is the Kronecker product. Then we can approximate $\mu(s,t)$ by $\boldsymbol{B}'_\mu(s,t)\boldsymbol{\beta}_\mu$ and estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_\mu$ by minimizing the following penalized sum of squares

$$\sum_{i=1}^{n}\sum_{j=1}^{m_T}\{Y_i(t_j) - \boldsymbol{B}'_\mu(s_i,t_j)\boldsymbol{\beta}_\mu - \boldsymbol{X}'_i\boldsymbol{\alpha}\}^2 + \mathcal{P}(\boldsymbol{\beta}_\mu;\lambda_\mu,\varrho), \tag{5}$$

where $\mathcal{P}(\boldsymbol{\beta}_\mu;\lambda_\mu,\varrho)$ is a penalty on $\boldsymbol{\beta}_\mu$ with tuning parameters λ_μ and ϱ. To increase model flexibility, we allow μ to have different degrees of smoothness in s and t by introducing two tuning parameters. To mimic the anisotropic thin-plate spline (Wood, 2000), we put penalty on

$$\lambda_\mu \int_{\mathcal{S}}\int_{\mathcal{T}} [\{\mu^{(2,0)}(s,t)\}^2 + 2\varrho\{\mu^{(1,1)}(s,t)\}^2 + \varrho^3\{\mu^{(0,2)}(s,t)\}^2]\,\mathrm{d}t\mathrm{d}s$$

where $\mu^{(k_1,k_2)}$ is the (k_1,k_2)th partial derivative of μ. Using the basis function representation, the thin-plate penalty can be written as

$$\mathcal{P}(\boldsymbol{\beta}_\mu;\lambda_\mu,\varrho) = \lambda_\mu\boldsymbol{\beta}'_\mu \int_{\mathcal{S}}\int_{\mathcal{T}} [\{\boldsymbol{B}^{(2,0)}_\mu(s,t)\}^{\otimes 2} + 2\varrho\{\boldsymbol{B}^{(1,1)}_\mu(s,t)\}^{\otimes 2} + \varrho^3\{\boldsymbol{B}^{(0,2)}_\mu(s,t)\}^{\otimes 2}]\,\mathrm{d}t\mathrm{d}s\boldsymbol{\beta}_\mu,$$

where $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}'$ for any matrix $\boldsymbol{A}$. Following Ruppert et al. (2003), we set both K_T and K_L to be relatively large and let the smoothness of the estimated function controlled by the tuning parameters (λ_μ,ϱ), which can be selected by data-driven methods such as the generalized cross-validation (GCV) (Wahba, 1990). Denote the estimators for camera effects and the mean function as $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\mu}(s,t)$.

Under model (1), we can easily see

$$\mathcal{G}_1(t_1,t_2) \equiv \mathrm{cov}\{Y_i(t_1),Y_i(t_2)\} = \mathcal{K}_1(t_1,t_2) + \mathcal{K}_2(t_1,t_2) + \mathcal{K}_3(t_1,t_2),\ \text{if } t_1 \neq t_2;$$

$$\mathcal{G}_2(t_1,t_2) \equiv \mathrm{cov}\{Y_{i_1}(t_1),Y_{i_2}(t_2)\} = \mathcal{K}_1(t_1,t_2) + \mathcal{K}_2(t_1,t_2),\ \text{if } i_1 \neq i_2,\ g(i_1) = g(i_2),\ f(i_1) = f(i_2);$$

$$\mathcal{G}_3(t_1,t_2) \equiv \mathrm{cov}\{Y_{i_1}(t_1),Y_{i_2}(t_2)\} = \mathcal{K}_1(t_1,t_2),\ \text{if } i_1 \neq i_2,\ g(i_1) = g(i_2),\ f(i_1) \neq f(i_2);$$

$$\sigma^2_Y(t) \equiv \mathrm{var}\{Y_i(t)\} = \mathcal{K}_1(t,t) + \mathcal{K}_2(t,t) + \mathcal{K}_3(t,t) + \sigma^2.$$

We will first estimate $\mathcal{G}_l$, $l = 1,2,3$, and then use the above relationships to estimate $\mathcal{K}_l$.

Our strategy for covariance estimation is to perform bivariate smoothing on the product of the residuals. Many authors have demonstrated the necessity of smoothing for covariance estimation in functional data with a random design (i.e., random, irregular observation time points on individual curves), including Yao et al. (2005), Hall et al. (2006) and Li and Hsing (2010). For functional data following a fixed design (i.e., all functions observed on the same dense grid) and free of measurement errors, many authors (Di et al., 2009; Shou et al., 2015) estimate the covariance matrix of the observed data without smoothing, which can be considered as approximating the covariance function by a bivariate step function. However, in virtually every application, discrete observations on the curves are contaminated with measurement errors, and the empirical covariance matrix based on the error-contaminated data is biased with an additional nugget effect on the diagonal. Under such situations, the authors mentioned above also recommend smoothing. When the true covariance function is smooth, as assumed in most functional data literature, smoothing the covariance function can significantly reduce the effect of measurement error by borrowing information from neighboring points. Furthermore, spline smoothing can also help to reduce the dimensionality of the covariance matrix. For these reasons, we advocate smoothing even when the observations are on a regular grid.

Define the residuals $\mathcal{E}_{ij} = Y_i(t_j) - \widehat{\mu}(s_i, t_j) - \boldsymbol{X}_i'\widehat{\boldsymbol{\alpha}}$, then the equations above suggest that we could estimate $\mathcal{G}_1(t_{j_1}, t_{j_2})$ by $\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_{ij_1}\mathcal{E}_{ij_2}$ for $j_1 \neq j_2$. However, these estimates are only defined on discrete time points with 3-minute gaps between consecutive time points. To estimate $\mathcal{G}_1$ as a function, we apply penalized tensor-product spline smoothing to these empirical covariance estimators. Define the second tensor-product spline basis $\boldsymbol{B}_{\mathcal{G}}(t_1, t_2) = \boldsymbol{B}_T(t_1)\otimes\boldsymbol{B}_T(t_2)$ and approximate $\mathcal{G}_1(t_1, t_2)$ by $\widehat{\mathcal{G}}_1(t_1, t_2) = \boldsymbol{B}_{\mathcal{G}}'(t_1, t_2)\widehat{\boldsymbol{\beta}}_{\mathcal{G}1}$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{G}1}$ minimizes

$$\sum_{i=1}^{n}\sum_{j_1=1}^{m_T}\sum_{j_2\neq j_1}\{\mathcal{E}_{ij_1}\mathcal{E}_{ij_2} - \boldsymbol{B}_{\mathcal{G}}'(t_{j_1}, t_{j_2})\boldsymbol{\beta}_{\mathcal{G}1}\}^2 + \lambda_{\mathcal{G}1}\boldsymbol{\beta}_{\mathcal{G}1}'\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}1},$$

and $\lambda_{\mathcal{G}1}$ and $\boldsymbol{\Omega}_{\mathcal{G}}$ are the tuning parameter and penalty matrix respectively. Note that the

$j_1 = j_2$ terms are omitted in the penalized sum of squares above in order to the eliminate the nugget effect caused by measurement errors (Yao et al., 2005; Di et al., 2009).

Similarly, we estimate $\mathcal{G}_2$ and $\mathcal{G}_3$ by $\widehat{\mathcal{G}}_2 = \boldsymbol{B}'_{\mathcal{G}}\widehat{\boldsymbol{\beta}}_{\mathcal{G}2}$ and $\widehat{\mathcal{G}}_3 = \boldsymbol{B}'_{\mathcal{G}}\widehat{\boldsymbol{\beta}}_{\mathcal{G}3}$ where $\widehat{\boldsymbol{\beta}}_{\mathcal{G}2}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{G}3}$ minimize the following penalized sum of squares

$$\sum_{i_1=1}^{n}\sum_{i_2\neq i_1}\sum_{j_1=1}^{m_T}\sum_{j_2=1}^{m_T}\{\mathcal{E}_{i_1j_1}\mathcal{E}_{i_2j_2} - \boldsymbol{B}'_{\mathcal{G}}(t_{j_1},t_{j_2})\boldsymbol{\beta}_{\mathcal{G}2}\}^2 I\{g(i_1)=g(i_2), f(i_1)=f(i_2)\} + \lambda_{\mathcal{G}2}\boldsymbol{\beta}'_{\mathcal{G}2}\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}2},$$
$$\sum_{i_1=1}^{n}\sum_{i_2\neq i_1}\sum_{j_1=1}^{m_T}\sum_{j_2=1}^{m_T}\{\mathcal{E}_{i_1j_1}\mathcal{E}_{i_2j_2} - \boldsymbol{B}'_{\mathcal{G}}(t_{j_1},t_{j_2})\boldsymbol{\beta}_{\mathcal{G}3}\}^2 I\{g(i_1)=g(i_2), f(i_1)\neq f(i_2)\} + \lambda_{\mathcal{G}3}\boldsymbol{\beta}'_{\mathcal{G}3}\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}3}.$$

When $\mathcal{G}$ is spanned by a tensor-product spline basis $\boldsymbol{B}_{\mathcal{G}}$, the penalty matrix corresponding to a thin-plate spline penalty is

$$\boldsymbol{\Omega}_{\mathcal{G}} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{\boldsymbol{B}_{\mathcal{G}}^{(2,0)}(t_1,t_2)\}^{\otimes 2} + 2\{\boldsymbol{B}_{\mathcal{G}}^{(1,1)}(t_1,t_2)\}^{\otimes 2} + \{\boldsymbol{B}_{\mathcal{G}}^{(0,2)}(t_1,t_2)\}^{\otimes 2}\mathrm{d}t_1\mathrm{d}t_2.$$

We can also estimate $\sigma_Y^2(t)$ by $\widehat{\sigma}_Y^2(t) = \boldsymbol{B}'_T(t)\widehat{\boldsymbol{\beta}}_\sigma$, where $\widehat{\boldsymbol{\beta}}_\sigma$ minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{m_T}\{\mathcal{E}_{ij}^2 - \boldsymbol{B}_T(t_j)\boldsymbol{\beta}_\sigma\}^2 + \lambda_\sigma\boldsymbol{\beta}_\sigma\boldsymbol{\Omega}_T\boldsymbol{\beta}_\sigma,$$

and λ_σ and $\boldsymbol{\Omega}_T = \int\{\boldsymbol{B}_T^{(2)}(t)\}^{\otimes 2}\mathrm{d}t$ are the tuning parameter and penalty matrix respectively. All tuning parameters defined above can be chosen by GCV.

Next, we estimate the covariance functions $\mathcal{K}_l, l = 1,2,3$ and the error variance σ^2 by

$$\begin{aligned}
&\widehat{\mathcal{K}}_1(t_1,t_2) = \widehat{\mathcal{G}}_3(t_1,t_2), \quad \widehat{\mathcal{K}}_2(t_1,t_2) = \widehat{\mathcal{G}}_2(t_1,t_2) - \widehat{\mathcal{G}}_3(t_1,t_2),\\
&\widehat{\mathcal{K}}_3(t_1,t_2) = \widehat{\mathcal{G}}_1(t_1,t_2) - \widehat{\mathcal{G}}_2(t_1,t_2),\\
&\widehat{\sigma}_I^2 = |\mathcal{T}|^{-1}\int\{\widehat{\sigma}_Y^2(t) - \widehat{\mathcal{K}}_1(t,t) - \widehat{\mathcal{K}}_2(t,t) - \widehat{\mathcal{K}}_3(t,t)\}\mathrm{d}t. \qquad (6)
\end{aligned}$$

The eigenvalues $\omega_{l,k}$ and eigenfunctions $\psi_{l,k}$ in (3) can be estimated by an eigenvalue decomposition of $\widehat{\mathcal{K}}_l$ using the approach of Ramsay and Silverman (2005). Denote their estimates as $\widehat{\omega}_{l,k}$ and $\widehat{\psi}_{l,k}$, respectively. Since all functions above are approximated by finite-dimensional splines, the eigenvalue decomposition problem reduces to a multivariate problem. None of

the estimated covariance functions, $\widehat{\mathcal{K}}_l$, $l = 1, 2, 3$, are guaranteed to be positive semi-definite. However, since these covariance estimators are consistent for the true functions, their leading eigenvalues approach the truth and tend to be strictly positive, and any negative eigenvalues tend to be small and close to zero. In practice, any negative eigenvalues of $\widehat{\mathcal{K}}_l$ can be truncated at zero. Such an approach is widely used in the functional data literature (Yao et al., 2005; Li, 2011) and theoretically justified by Hall et al. (2008). In fact, we will select the number of principal components using data-driven methods and reconstruct covariance functions from their leading principal components; therefore, the possibility of $\widehat{\mathcal{K}}_l$ not being positive semi-definite is not a great concern.

3.2 Estimating the principal component scores

For predetermined numbers of principal components (p_1, p_2, p_3) for the three levels of functional random effects, we estimate the principal component scores by best linear unbiased prediction (BLUP). This is an extension of the "PACE" method of Yao et al. (2005) to hierarchical functional data.

Let $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,m_T})'$, $\boldsymbol{\mu}_i = \{\mu(s_i, t_1), \ldots, \mu(s_i, t_{m_T})\}'$, $\boldsymbol{\psi}_{l,k} = \{\psi_{l,k}(t_1), \ldots, \psi_{l,k}(t_{m_T})\}'$, $\boldsymbol{\Psi}_l = (\boldsymbol{\psi}_{l,1}, \boldsymbol{\psi}_{l,2}, \ldots, \boldsymbol{\psi}_{l,p_l})$, $l = 1, 2, 3$. Define the vectors of FPCA scores $\boldsymbol{\xi}_{1,g} = \{\xi_{1,g,1}, \ldots, \xi_{1,g,p_1}\}'$ for $g = 1, \ldots, G$, $\boldsymbol{\xi}_{2,f} = \{\xi_{2,f,1}, \ldots, \xi_{2,f,p_2}\}'$ for $f = 1, \ldots, F$, and $\boldsymbol{\xi}_{3,i} = \{\xi_{3,i,1}, \ldots, \xi_{3,i,p_3}\}'$ for $i = 1, \ldots, n$. For any positive integer d, denote $\mathbf{1}_d$ as a d-dimensional vector of ones and $\boldsymbol{I}_d$ as a d by d identity matrix. For any index set $I = \{k_1, \ldots, k_d\} \subset \{1, \ldots, n\}$ and any sequence of vectors or matrices $\boldsymbol{A}_k$ of the same dimension(s), denote $(\boldsymbol{A}_k)_{k \in I}$ as $(\boldsymbol{A}'_{k_1}, \ldots, \boldsymbol{A}'_{k_d})'$. For $g = 1, \ldots, G$, define $\boldsymbol{Y}_g = (\boldsymbol{Y}_i)_{g(i)=g}$, $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_i)_{g(i)=g}$, $\boldsymbol{X}_g = (\mathbf{1}_{m_T} \otimes \boldsymbol{X}'_i)_{g(i)=g}$, $\boldsymbol{\xi}_g = \{\boldsymbol{\xi}'_{1,g}, (\boldsymbol{\xi}_{2,f})'_{g(f)=g}, (\boldsymbol{\xi}_{3,i})'_{g(i)=g}\}'$, $\boldsymbol{\Lambda}_g = \mathrm{diag}\{\mathrm{var}(\boldsymbol{\xi}_g)\}$, $\boldsymbol{\Phi}_g = (\mathbf{1}_{n_g} \otimes \boldsymbol{\Psi}_1, \boldsymbol{D}_g \otimes \boldsymbol{\Psi}_2, \boldsymbol{I}_{n_g} \otimes \boldsymbol{\Psi}_3)$, where $\boldsymbol{D}_g$ is a n_g by F_g matrix with element (k_1, k_2) equal to 1 if the k_1th seed in genotype g is recorded in the k_2th file and 0 otherwise. It is easy to see that $\boldsymbol{\Sigma}_g = \mathrm{cov}(\boldsymbol{Y}_g) = \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g \boldsymbol{\Phi}'_g + \sigma^2 \boldsymbol{I}_{m_T n_g}$. Under the assumption that $\boldsymbol{\xi}_g$ and $\epsilon(t)$ are jointly Gaus-

sian, $\mathrm{E}(\boldsymbol{\xi}_g|\boldsymbol{Y}_g, \boldsymbol{X}_g) = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{Y}_g - \boldsymbol{\mu}_g - \boldsymbol{X}_g\boldsymbol{\alpha})$. The estimator of $\boldsymbol{\xi}_g$ is its empirical BLUP

$$\widehat{\boldsymbol{\xi}}_g = \widehat{\boldsymbol{\Lambda}}_g\widehat{\boldsymbol{\Phi}}_g'\widehat{\boldsymbol{\Sigma}}_g^{-1}(\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\mu}}_g$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Lambda}}_g$, and $\widehat{\boldsymbol{\Phi}}_g$ are the estimates using the proposed FPCA method described in Section 3.1, $\widehat{\boldsymbol{\Sigma}}_g = \widehat{\boldsymbol{\Phi}}_g\widehat{\boldsymbol{\Lambda}}_g\widehat{\boldsymbol{\Phi}}_g' + \widehat{\sigma}_I^2\boldsymbol{I}_{m_T n_g}$, and $\widehat{\sigma}_I^2$ is a pilot estimator of σ^2 obtained by integration defined in (6). The matrix $\widehat{\boldsymbol{\Sigma}}_g$ is of dimension $m_T n_g \times m_T n_g$. When the number of observations per curve m_T is large, directly inverting this matrix can pose a numerical challenge. Instead, an equivalent formula $\widehat{\boldsymbol{\Sigma}}_g^{-1} = \widehat{\sigma}_I^{-2}\boldsymbol{I}_{m_T n_g} - \widehat{\sigma}_I^{-4}\widehat{\boldsymbol{\Phi}}_g(\widehat{\boldsymbol{\Lambda}}_g^{-1} + \widehat{\sigma}_I^{-2}\widehat{\boldsymbol{\Phi}}_g'\widehat{\boldsymbol{\Phi}}_g)^{-1}\widehat{\boldsymbol{\Phi}}_g'$ can be used in the computation, which only involves the inverse of a matrix of a much lower dimension $p_1 + F_g p_2 + n_g p_3$.

3.3 Iterative procedure to refine mean estimation

The algorithm we propose in Section 3.1 to estimate the mean and covariance functions is an extension of the method for two-level hierarchical functional data in Di et al. (2009) to the three-level setting but with more emphasis on smoothing. The benefit of this approach is that it does not involve computationally intensive EM algorithms as in Li et al. (2015) and can handle large functional data sets. However, the estimator in (5) is a working independence estimator (Lin and Carroll, 2001) which ignores correlation in the data, so it is not efficient. To improve the estimation efficiency and increase the power for statistical tests, we refine the mean estimator using an iterative decorrelation procedure similar to that of Yao and Lee (2006) for uni-level FPCA.

The iterative procedure is as follows.

Step 1: Use the procedures in Section 3.1 to obtain $\widehat{\boldsymbol{\mu}}_i$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\mathcal{K}}_l$, $l = 1, 2, 3$, and perform spectral decomposition of these covariance functions.

Step 2: Use AIC defined in Section 4.1 to choose the numbers of principal components $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$ for the three levels and obtain estimates of the eigenvalues $\widehat{\boldsymbol{\Lambda}}_g$, eigenfunctions

$\widehat{\boldsymbol{\Psi}}_g$ and the principal component scores $\widehat{\boldsymbol{\xi}}_g$, $g = 1, \ldots, G$.

Step 3: Re-estimate $\mu(s,t)$ and $\boldsymbol{\alpha}$ using (5) with $Y_i(t_j)$ replaced by

$$Y_i^*(t_j) = Y_i(t_j) - \sum_{k=1}^{\widehat{p}_1} \widehat{\xi}_{1,g(i),k}\widehat{\psi}_{1,k}(t_j) - \sum_{k=1}^{\widehat{p}_2} \widehat{\xi}_{2,f(i),k}\widehat{\psi}_{2,k}(t_j) - \sum_{k=1}^{\widehat{p}_3} \widehat{\xi}_{3,i,k}\widehat{\psi}_{3,k}(t_j). \tag{7}$$

Step 4: Update the covariance estimates using the updated mean function and parameters; update the estimates of the eigenvalues, eigenfunctions, and principal component scores; and if necessary, adjust the numbers of principal components using AIC.

Step 5: Repeat *Step 3* and *Step 4* until relative changes in $\widehat{\mu}$ and $\widehat{\boldsymbol{\alpha}}$ between adjacent iterations are smaller than a predetermined tolerance. The final numbers of principal components are those selected by AIC in the final step.

The decorrelation procedure in *Step 3* is an extension of the algorithm in Yao and Lee (2006) to hierarchical functional data, where we diminish the correlation in the response by subtracting the predicted functional random effects. The estimator in (5) is efficient for uncorrelated responses. The iterative procedure described above is also closely connected with the ECME algorithm described in Li et al. (2015), where the difference is in how parameters are updated in the M-step.

Based on our simulation studies, the numbers of principal components are usually chosen perfectly by AIC in *Step 2*. To save computation time, the subsequent updates of $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$ can be skipped in *Step 4*. Because our initial estimator in *Step 1* is already consistent, the full iterative procedure can usually converge in 5 iterations.

4 Model Selection and Statistical Inference

4.1 Selecting the number of principal components

All existing papers on hierarchical functional data analysis select the number of principal components using the ad hoc PVE method, where the estimated eigenvalue sequence of each

functional random effect is truncated at a subjectively chosen percentage of variation explained. More recently, Li et al. (2013) proposed a model selection method using conditional AIC to select the number of principal components for uni-level functional data, and we now extend their method to hierarchical functional data.

Assuming the measurement errors are Gaussian, the conditional log-likelihood of the observed data $\{\boldsymbol{Y}_i\}_{i=1}^n$ given the principal component scores is

$$\ell_{n,\mathcal{C}} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i - \boldsymbol{X}_i'\boldsymbol{\alpha} - \boldsymbol{\Psi}_1\boldsymbol{\xi}_{1,g(i)} - \boldsymbol{\Psi}_2\boldsymbol{\xi}_{2,f(i)} - \boldsymbol{\Psi}_3\boldsymbol{\xi}_{3,i}\|^2, \qquad (8)$$

where $\|\cdot\|$ denotes the L^2 norm and $N = nm_T$ is the total number of measurements.

For a given model specified by the numbers of components (p_1, p_2, p_3), we first predict the functional random effects using the procedure in Section 3.2 and evaluate the conditional likelihood of the observed data by replacing various fixed and random effects by their estimators or predictors. The AIC is the value of negative conditional log-likelihood plus a penalty on the complexity of the hierarchical functional data model. Motivated by Li et al. (2013), because the functional random processes for the three levels are mutually independent, we penalize the number of estimated random effects to obtain

$$\mathrm{AIC}(p_1, p_2, p_3) \;=\; -2\widehat{\ell}_{n,\mathcal{C}} + 2(Gp_1 + Fp_2 + np_3), \qquad (9)$$

where $\widehat{\ell}_{n,\mathcal{C}}$ is the estimated conditional likelihood as described above. The numbers of components are selected by minimizing (9) using a grid search method. To reduce computation burden, we set an upper bound for each p_l (e.g., 5 or 10). If the minimizer of AIC falls on an upper boundary for $l = 1, 2, \text{or } 3$, the searched region is expanded until AIC is minimized in the interior of the searched region.

An intuitive explanation for the penalty in (9) is as follows. The fixed effects including $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$ and $\boldsymbol{\Psi}_3$ are estimated with the highest accuracy by pooling all data together and hence can be deemed as known for the purpose of selecting p_1, p_2, and p_3. Then the

likelihood in (8) can be considered as a regression on $\boldsymbol{Y}_i - \boldsymbol{\mu}_i - \boldsymbol{X}_i'\boldsymbol{\alpha}$ against $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$ and $\boldsymbol{\Psi}_3$, where $\boldsymbol{\xi}_{1,g}$, $\boldsymbol{\xi}_{2,f}$ and $\boldsymbol{\xi}_{3,i}$ are the genotype-specific, file-specific and seed-specific regression coefficients. The total number of regression coefficients is $Gp_1 + Fp_2 + np_3$. This calculation is logical because we have dense observations on each curve so that there are enough data to fit a regression for each seed.

4.2 Testing for Lunar Effects

To test for lunar effects, we consider a reduced version of model (1), where the mean of the response $Y_i(t)$ does not depend on the lunar day s, i.e. $\mu(s,t) \equiv \mu_R(t)$. We will test the hypothesis

$$H_0 : \mu(s,t) = \mu_R(t) \quad \text{vs.} \quad H_1 : \mu(s,t) \neq \mu_R(t), \tag{10}$$

by a generalized likelihood ratio (GLR) test (Fan et al., 2001). The classic GLR test was proposed for testing nonparametric hypotheses for independent data. Some recent reviews on this test include Fan and Jiang (2007), and González-Manteiga and Crujeiras (2013). The approach has also recently extended to uni-level sparse functional data by Tang et al. (2016), who build a GLR test based on working independence estimators. In our setting, we introduce three versions of GLR tests based on marginal likelihood, conditional likelihood and working independence (WI), respectively.

Under the full model and Gaussian assumptions, the estimated marginal likelihood is

$$\widehat{\ell}_{n,\mathcal{M}}(H_1) \propto -\frac{1}{2}\sum_{g=1}^{G}(\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\alpha}})'\widehat{\boldsymbol{\Sigma}}_g^{-1}(\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\alpha}}), \tag{11}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\alpha}}$ are the refined estimators in Section 3.3, $\widehat{\boldsymbol{\Sigma}}_g$ is the covariance matrix for the response variables within genotype g reconstructed from FPCA as in Section 3.2 using the selected numbers of components.

Because the reduced model is nested in the full model, the FPCA estimators (including

the eigenvalues, eigenfunctions and FPC scores) under the full model are still legitimate when H_0 is true. We fit $\mu_R(t)$ using univariate penalized splines to the decorrelated response in (7), where the principal component scores, eigenvalues, and eigenfunctions are obtained from the full model. The likelihood under the reduced model, denoted as $\widehat{\ell}_{n,\mathcal{M}}(H_0)$, is similarly defined as in (11), where the $\widehat{\boldsymbol{\Sigma}}_g$ matrices are obtained under the full model but $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are fitted under the reduced model. This strategy is beneficial because the likelihoods under the full and reduced models are comparable, nuisance from refitting FPCA is avoided, and the likelihood ratio is guaranteed to be positive. Our approach is also motivated by power considerations: when the null hypothesis is wrong, the reduced model covariance estimator can pick up some of the signals missed by the reduced model mean estimator in order to maximize the likelihood; if we allow different covariance estimators in the full and reduced model likelihoods, the magnitude of the GLR statistic is reduced and hence the power of detecting any lunar day effect. The marginal likelihood based test statistic (GLR-ML) is defined as

$$T_{n,\mathcal{M}} = \widehat{\ell}_{n,\mathcal{M}}(H_1) - \widehat{\ell}_{n,\mathcal{M}}(H_0).$$

Similarly, we can define GLR statistics based on conditional likelihood (GLR-CL)

$$T_{n,\mathcal{C}} = \widehat{\ell}_{n,\mathcal{C}}(H_1) - \widehat{\ell}_{n,\mathcal{C}}(H_0),$$

where both likelihoods are as defined in (8) using the same σ^2, eigenfunctions and FPC scores estimated under the full model, but with $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ estimated under the full and reduced models to obtain $\widehat{\ell}_{n,\mathcal{C}}(H_1)$ and $\widehat{\ell}_{n,\mathcal{C}}(H_0)$, respectively.

For comparison, we also define a test based on working independence which totally ignores the covariance structure among the response observations. In general, WI is a simple strategy in longitudinal data analysis that results in consistent estimation (Lin and Carroll, 2001) and legitimate test procedures (Tang et al., 2016). In fact, our initial mean estimators in Section 3.1, which we now denote as $\widehat{\boldsymbol{\mu}}_g^{\mathcal{W}}$ and $\widehat{\boldsymbol{\alpha}}^{\mathcal{W}}$, are WI estimators. The WI test statistic

(GLR-WI) is defined as

$$T_{n,\mathcal{W}} = \widehat{\ell}_{n,\mathcal{W}}(H_1) - \widehat{\ell}_{n,\mathcal{W}}(H_0),$$

where $\widehat{\ell}_{n,\mathcal{W}}(H_1) \propto -\frac{1}{2}\sum_{g=1}^{G} \|\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g^{\mathcal{W}} - \boldsymbol{X}_g\widehat{\boldsymbol{\alpha}}^{\mathcal{W}}\|^2$ and $\widehat{\ell}_{n,\mathcal{W}}(H_0)$ is defined similarly except with the mean estimators replaced by their reduced model counterparts. The WI test is easy to implement because neither FPCA, model selection, nor the refined estimation procedure in Section 3.3 are needed. However, we find in our numerical studies that the GLR tests based on $T_{n,\mathcal{M}}$ and $T_{n,\mathcal{C}}$, both of which rely on FPCA, yield higher power than the WI test.

We propose to estimate the null distribution of the proposed test statistics using a permutation strategy for the following reasons. First, the asymptotic distributions of the GLR tests of univariate null versus bivariate alternatives have not been investigated in the literature. Second, as pointed out by many authors (Mammen, 1993; Fan and Jiang, 2007; Tang et al., 2016), the asymptotic distribution of a nonparametric test statistic usually performs poorly under finite samples because of the slow convergence rate in nonparametric settings. Even for the cases where the asymptotic distribution of the test statistic is available, these authors favor resampling methods.

A simple permutation strategy is to break the association between the response $Y_i(t)$ and the lunar day s_i. The RIS data were acquired on a total of D days. Each day in the original data corresponds to a particular lunar day in $\{1, 2, \ldots, 30\}$. Under the null hypothesis of no lunar effects, all possible permutations of the lunar day labels in the observed data relative to the D days of data collection are equally likely. Each such permutation results in a new lunar day assignment for the ith observation that we denote as s_i^* for all $i = 1, \ldots, n$.

Step 1: Randomly select one permutation as described above by shuffling lunar day labels relative to data collection days and express the permuted data set as $\{Y_i(t), \boldsymbol{X}_i, s_i^*; i = 1, \ldots, n\}$.

Step 2: Calculated the GLR test statistic T_n^* based on the permuted data set.

Step 3: Independently repeat Steps 1 and 2 a large number of times and estimate the p-value by the empirical relative frequency that T_n^* is greater than T_n.

This procedure is applicable to all three versions of T_n proposed above.

5 Simulation Studies

5.1 Estimation Results

We evaluate the performance of the proposed methodology using simulation studies that mimic the real data. We generate data according to the following model:

$$Y_i(t_j) = \mu(s_i, t_j) + \boldsymbol{X}_i'\boldsymbol{\alpha} + \sum_{k=1}^{p_1} \xi_{1,g(i),k}\psi_{1,k}(t_j) + \sum_{k=1}^{p_2} \xi_{2,f(i),k}\psi_{2,k}(t_j) + \sum_{k=1}^{p_3} \xi_{3,i,k}\psi_{3,k}(t_j) + \epsilon_i(t_j),$$

where $\mu(s_i, t_j) = \{1 + \cos(2\pi s_i/10)/5\}\{-12(t_j - 1/2)^2 + 3\}$ is the mean function, $\boldsymbol{X}_i'$ is a vector of camera indicators, $\boldsymbol{\alpha} = (-0.3, -0.2, -0.1, 0.1, 0.2, 0.3)$, $\epsilon_i(t_j) \sim N(0, \sigma^2)$ with $\sigma = 1$, $s_i \in [1, 30]$, $t_j = j/(m_T - 1), j = 0, \ldots, (m_T - 1)$, and $m_T = 31$. The mean function μ, as shown in Figure 2 (a), is chosen to mimic the mean function estimated from the real data. We simulate data for $n = 1000$ seeds from $G = 100$ genotypes with two files for each genotype and five seeds in each file. Following the structure of the real data, we assume that all seeds within a genotype are observed on the same day with lunar days randomly assigned to genotypes. We consider the following two scenarios and conduct 200 simulations under each scenario.

Scenario I: Let $p_1 = p_2 = p_3 = 2$. The principal component scores $\xi_{l,k}$ are generated independently from $N(0, \omega_{l,k})$, $l = 1, 2, 3$. The eigenvalues are $(\omega_{1,1}, \omega_{1,2}) = (1, 1/4)$ at the genotype level, $(\omega_{2,1}, \omega_{2,2}) = (1/2, 1/4)$ at the file level, and $(\omega_{3,1}, \omega_{3,2}) = (5, 1/2)$ at the seed level. For the eigenfunctions, we set

$$\psi_{1,1}(t) = \sqrt{2}\sin(2\pi t), \quad \psi_{1,2}(t) = \sqrt{2}\cos(2\pi t), \quad \psi_{2,1}(t) = \sqrt{2}\sin(4\pi t),$$
$$\psi_{2,2}(t) = \sqrt{2}\cos(4\pi t), \quad \psi_{3,1}(t) = 1, \quad \psi_{3,2}(t) = \sqrt{12}(t - 1/2).$$

Note that the seed level eigenfunctions are not orthogonal to the genotype and file level eigenfunctions under this scenario.

Scenario II: Let $p_1 = 1, p_2 = 1, p_3 = 4$. To illustrate the robustness of our proposed procedure against violation of the normal assumption, we simulate the principal component scores from a skewed Gaussian mixture model. Specifically, for a principal component score ξ with mean zero and variance ω, we generate ξ with probability 1/3 from $N(2\sqrt{\omega/3}, \omega/3)$, and with probability 2/3 from $N(-\sqrt{\omega/3}, \omega/3)$. We set eigenvalues as $\omega_{1,1} = 1/4$ at the genotype level; $\omega_{2,1} = 1/2$ at the file level; and $(\omega_{3,1}, \omega_{3,2}, \omega_{3,3}, \omega_{3,4}) = (2, 1, 1/2, 1/4)$ at the seed level. For the eigenfunctions, we consider the following mutually orthogonal functions:

$$\psi_{1,1}(t) = \sqrt{2}\sin(2\pi t), \quad \psi_{2,1}(t) = \sqrt{2}\cos(2\pi t), \quad \psi_{3,1}(t) = \sqrt{2}\sin(4\pi t),$$
$$\psi_{3,2}(t) = \sqrt{2}\cos(4\pi t), \quad \psi_{3,3}(t) = \sqrt{2}\sin(6\pi t), \quad \psi_{3,4}(t) = \sqrt{2}\cos(6\pi t).$$

We first focus on the estimation results under Scenario I. To estimate $\mu(s,t)$, we use the tensor products between $K_T = 14$ cubic B-splines in t (10 interior knots and 4 boundary knots) and $K_L = 11$ Fourier functions in s as the basis. We also use the tensor product of 14 B-splines for all covariance estimation. All tuning parameters, including λ_μ and ϱ in (5) and $\lambda_{\mathcal{G}1}$, $\lambda_{\mathcal{G}2}$ and $\lambda_{\mathcal{G}3}$ for covariance estimation, are chosen by GCV. In Figure 2, we compare the true function $\mu(s,t)$ with the proposed spline estimator in a typical run and the mean of our estimator averaged over 200 runs. We also calculate the integrated mean squared error (IMSE) of the mean estimator, which is defined as $\int_{\mathcal{S}}\int_{\mathcal{T}}\{\widehat{\mu}(s,t) - \mu(s,t)\}^2 \mathrm{d}s\mathrm{d}t$. Compared with the working independence initial estimator (5), the refined mean estimator based on FPCA in Section 3.3 shows a significant reduction of IMSE (23.12% reduction under Scenario I and 28.28% reduction under Scenario II)

We also provide boxplots of the eigenvalues in Figure 3 and a graphical summary for the estimated eigenfunctions in Figure 4, where we compare in each panel the 2.5% and 97.5% point-wise percentiles of our eigenfunction estimator with the truth. As we can see, all of

these estimators perform quite well; the estimated eigenvalues are close to the true values, and the true eigenfunctions are always nicely covered by the percentile bands. We observe in Figure 3 that the estimated eigenvalues in the third level have less bias than the first two levels. Similarly, in Figure 4, the percentile bands for the third level eigenfunctions are tighter than the first two levels. An explanation of this phenomenon is that there are more repetitions for the third level functional random effects ($n = 1000$) than the first two levels ($G = 100$ and $F = 200$). Estimation results for Scenario II are similar to those in Scenario I and hence are relegated to the supplementary materials.

We show in Figure 5 scatterplots of the predicted principal component scores versus the true scores in a typical run under Scenario II. Even though the true distribution of the FPC scores are non-Gaussian, points in all panels of Figure 5 are remarkably close to the 45 degree reference lines, which also shows that the BLUP of the FPC scores is quite robust against mild violation of the Gaussian assumption.

5.2 Model selection results

To evaluate the finite sample performance of our model selection procedure based on conditional AIC, we compare our method with the widely-used PVE method (Di et al., 2009; Li et al., 2015). The threshold percentage in the PVE method is usually chosen subjectively (e.g. Di et al. (2009) use 90% and Li et al. (2015) use 85%). For completeness, we consider four commonly used thresholds, 85%, 90%, 95%, and 99%. For each simulation, we select the value of (p_1, p_2, p_3) that minimizes the conditional AIC over a simple grid of candidates. The model selection results are summarized in Table 1, where we present the empirical distribution of the selected number of principal components in each level of the hierarchy by each method. The modal frequency of each estimator is marked in bold. Under Scenario I, our AIC picked the correct number of principal component 100%, 97.5% and 100% of the time for the three hierarchies respectively; the PVE method with thresholds 85% and 90%

frequently selected the wrong model for level 2 and level 3. The PVE method with thresholds 95% and 99% frequently selected the wrong model for level 1 and level 2. Overall, our proposed method chose the correct number of principal components in all levels for 97.5% of the simulation replications, while PVE methods seldom selected the correct number of components in all levels. Under Scenario II with the correct model $(p_1, p_2, p_3) = (1, 1, 4)$, the contrast between our method and the PVE methods is even more striking: our method selected the correct number of components in all levels of the hierarchy 100% of the time, while the PVE methods never selected the true model.

Even though our conditional AIC is motivated from a conditional Gaussian likelihood and the principal component scores are estimated using BLUP under a Gaussian assumption, the procedure performs remarkably well under Scenario II, where the data are non-Gaussian. These results again show the robustness of the BLUP for the PC scores and the AIC against mild violation of Gaussian assumptions.

5.3 Hypothesis test results

We now demonstrate the performance of the proposed GLR tests on the hypotheses in (10). We first investigate whether the proposed permutation procedure retains the nominal size of the tests and compare the powers of the three versions of GLR tests. Because both GLR-ML and GLR-CL depend on specification of the numbers of principal components in FPCA, we also investigate the sensitivity of our GLR test statistics to these choices.

We adopt the same simulation setting in Scenario I described in Section 5.1 except that we set the mean function to be $\mu(s_i, t_j) = \{1+\delta\cos(2\pi s_i/10)/5\}\{-12(t_j-1/2)^2+3\}$ for some constant δ. It is easy to see that the null hypothesis $\mu(s,t) \equiv \mu(t)$ is true when $\delta = 0$, and that larger values of δ indicate further deviations from the null. We simulate 200 data sets for each $\delta \in \{0, 0.5, 1, 1.5, 2\}$ so that data are generated under both the null and alternative hypotheses.

We first assume the numbers of principal components are correctly specified, and perform level $\alpha = 0.05$ tests on the null hypothesis (10) for each simulated data set, where the decision is made using the permutation procedure described in Section 4.2. The empirical powers of the three GLR tests, as functions of δ, are shown in Figure 6 (a). As we can see, the permutation method estimates the null distributions remarkably well and all three GLR tests hold their nominal sizes. By comparing the three power curves, it seems that the GLR-ML is most powerful, followed by GLR-CL and then GLR-WI. The gap in estimated power among the three tests is the largest when $\delta = 1$. To check for significant differences in power at $\delta = 1$, we use McNemar's test. The p-values are 3×10^{-8} for GLR-ML vs. GLR-CL, 4×10^{-13} for GLR-ML vs. GLR-WI, and 0.02 for GLR-CL vs. GLR-WI. Even with a conservative Bonferroni correction for multiple testing across up to four δ values with three pairwise comparisons at each value of δ, the GLR-ML advantage in power over GLR-CL and GLR-WI is statistically significant. This finding supports our intuition that modeling the covariance structure in functional data should increase the power of tests.

Next, we investigate the sensitivity of the test statistics to the selected number of principal components. Figure 6 (b) shows the powers of the three GLR tests when the numbers of principal components are incorrectly specified as $(p_1, p_2, p_3) = (1, 1, 1)$; and Figure 6 (c) shows the powers when the numbers of principal components are incorrectly specified as $(p_1, p_2, p_3) = (3, 3, 3)$. The permutation method maintains the nominal sizes for all tests even under model mis-specification. Since GLR-WI does not depend on FPCA, its power curve is identical across the three panels of Figure 6. When the numbers of principal components are incorrectly specified as $(p_1, p_2, p_3) = (1, 1, 1)$, GLR-ML loses a lot of power and can fall beneath GLR-WI (the McNemar p-value is 3×10^{-6} at $\delta = 1.5$); GLR-CL is relatively robust and still more powerful than GLR-WI (the McNemar p-value is 6×10^{-4} at $\delta = 1$). On the other hand, panel (c) of Figure 6 is almost identical to panel (a), showing both GLR-ML and GLR-CL are relatively robust when the numbers of principal components

are incorrectly specified as $(p_1, p_2, p_3) = (3, 3, 3)$. One simple explanation is that, when extra principal components are included, the additional eigenvalues are close to zero and the variances of the extra FPC scores are too small to have any influence on the GLR test statistics.

These sensitivity analysis results show that using too few principal components affects power more seriously, especially for the test based on a marginal likelihood. On the other hand, even though using too many principal components may not lead to power loss, it can reduce the interpretability of the model.

6 Data Analysis

We now apply the proposed methodology to our motivating data, the RIS data. We use similar basis functions as in the simulation study except that the time domain is $[0, 1.5]$ rather than $[0, 1]$. All tuning parameters are selected by GCV. Figure 7 (a) shows the empirical mean surface by averaging bending rates over each combination of a lunar day and an observation time. As we can see from this plot, the design is not balanced in lunar days: the two missing stripes correspond to the 7th and the 21st lunar day when no experiments were conducted; the empirical mean curve shows substantially more variation on some lunar days because there are fewer observations on those days.

Figure 7 (b) shows the estimated mean surface using the proposed iterative procedure in Section 3.3. On a given lunar day, the mean curve looks somewhat parabolic across all but the beginning and end of the observation times with the highest bending rate around the mid point of the time span. The shape of the mean surface seems to suggest that the mean bending rate curve changes smoothly across lunar days and reaches its highest peak around the new moon. Using the 7th camera as the baseline, the estimated camera effects are $\widehat{\boldsymbol{\alpha}} = (-0.092, -0.041, -0.017, -0.028, -0.068, -0.016)'$ with corresponding standard errors $(0.019, 0.018, 0.018, 0.019, 0.019, 0.020)'$.

Next, we use the procedure in Section 3.1 to estimate the covariance functions and the error variance. The estimated error variance is $\widehat{\sigma}^2 = 1.034$. We apply the proposed conditional AIC to choose the numbers of principal components, which yields a model with $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (1, 1, 4)$. In contrast, the PVE method chooses $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (2, 2, 3)$ when the threshold is set at 85% of variation explained and $(3, 3, 3)$ at 90%. Table 2 shows the selected nonzero eigenvalues and their percentage of variation explained within their own levels. We find that the three levels of functional random effects do not contribute equally to the total variation in the response. A natural measure of the proportion of variability explained by level l is $\sum_{k=1}^{\widetilde{p}_l} \widehat{\omega}_{l,k} / \sum_{l=1}^{3} \sum_{k=1}^{\widetilde{p}_l} \widehat{\omega}_{l,k}$, where $\widetilde{p}_l = \max_k\{\widehat{\omega}_{l,k} > 0\}$, $l = 1, 2, 3$ Using this formula, the genotype, file, and seed effects account for 26.9%, 9.3%, and 63.8% of the total variation in the bending rate process. In light of this observation, the model chosen by our method seems to be more reasonable because it chooses more principal components in the third level that contributes more to the total variation. In contrast, the PVE method chooses the numbers of principal components independently in each level using the relative percentage within that level.

We show in Figure 8 the estimated eigenfunctions for all principal components selected by AIC. Ramsay-Silverman-style effect plots (Ramsay and Silverman, 2005) for the eigenfunctions are shown in Figure 9, where the effect of an eigenfunction is depicted as perturbations to the mean function. The sold curve in each panel of Figure 9 is the mean function averaged over lunar days, $\bar{\mu}(t) = |\mathcal{S}|^{-1} \int \mu(s,t) \mathrm{d}s$, and the other two curves are $\bar{\mu}(t) \pm 0.4 c \psi(t)$, where c is the standard deviation of $\bar{\mu}$.

The genotype eigenfunction, shown in Figure 8 (a), is generally positive with the highest value at about 1 hour. It indicates that seeds with positive scores on the genotype component tend to have higher bending rate than the population average and this tendency is strongest at about 1 hour after the beginning of the imaging process. The shape of the eigenfunction suggests that genetic effects are small at the beginning and the end of the time span and

are the largest in the middle of the observation period. The file-level eigenfunction, shown in Figure 8 (b), is similar to a sinusoidal function, the effect of which is shown in Figure 9 (b). It seems the file effect is a change in time scale: seeds with high loadings on the file eigenfunction have a prolonged bending process, and those with a negative loading finish the bending process in a shorter period of time. The leading principal component in the seed level, shown in Figure 8 (c), has a very similar shape as the file level eigenfunction and has a similar effect as well. The higher order seed level principal components can be interpreted in a similar fashion.

We show in Figure S.4 of the supplementary materials Normal Q-Q plots of the estimated principal component scores, where the empirical distributions of the FPC scores show various degrees of heavy-tail and/or skewness. It is comforting that our simulation results in Section 5 show that our methods are robust against deviation from normality.

The most important scientific question that motivates our research is whether there are lunar effects on root bending. The mean surface in Figure 7 (b) seems to support the existence of such an effect, but a more formal conclusion needs to be drawn from a hypotheses test. We apply the proposed GLR tests described in Section 4.2 to the RIS data and approximate the null distributions of the GLR statistics by 1,000 permutation repetitions. The GLR-WI and GLR-CL yield p-values of 0.025 and 0.043 respectively, which indicate significant lunar effects. The GLR-ML on the other hand yields a p-value of 0.204 which is insignificant. Based on these results and our experience from the simulation studies, lunar effects may exist because the two robust versions of the GLR test, GLR-WI and GLR-CL, suggest so. Even though GLR-ML is marginally most powerful under the correctly specified model, it is not uncommon in our simulation study that both GLR-WI and GLR-CL reject the null hypothesis in a specific sample while GLR-ML fails.

7 Discussion

Motivated by the RIS data, we propose methodology for estimation, model selection, and testing in the presence of 3-level nested hierarchical functional data. These methods can be easily extended to more complicated multi-level functional data. We have shown that modeling the covariance structure in the data by multi-level FPCA can increase both estimation efficiency for the mean function and test power. To apply the FPCA methodology, selecting the number of principal components becomes a critical model selection problem, and we propose a conditional AIC for this task. Our simulation studies show that the BLUP of the principal component scores and the conditional AIC are robust against deviation from the Gaussian assumptions. The conditional AIC also vastly out-performed the widely used PVE method. Not only is the PVE method subjective, it may not make sense to decide the number of FPC in each level using the within-level percentage of variation explained when the three levels of functional random effects do not contributed equally to the overall variation in the response as in our RIS data.

Nonparametric hypothesis testing is a relatively underdeveloped area in functional data analysis, especially when testing a one-dimensional null model against a two-dimensional alternative model as in our motivating data. Li et al. (2015) proposed Wald tests for the mean components in 3-level hierarchical functional data analysis, but as they mentioned in their paper, their tests do not hold the nominal size. We propose three versions of GLR test statistics, based on working independence, conditional likelihood and marginal likelihood respectively, that can be used to test for lunar effects in our data. Our simulation results show that our simple permutation procedure holds the nominal size of the test remarkably well for all three versions of GLR tests. Among the three GLR tests, GLR-ML is most powerful if the numbers of FPC's are correctly specified, but it is also the least robust approach; the GLR-CL might be the best compromise between power and robustness; and the GLR-WI is most robust but least powerful.

Theoretical properties for GLR tests in functional data are largely unknown and will be the subject of our future work. The recent result of Tang et al. (2016) suggests, for sparse functional data, a GLR-ML test based on a working independence estimator like that in (5) does not necessarily have better power than GLR-WI. Therefore, it is important to improve the efficiency of the mean estimator using the FPCA-based iterative procedure in Section 3.3 before performing a GLR-ML test.

As a referee correctly pointed out, our model for the RIS data could potentially include effects for months and days nested in months to arrive at a model with five random factors instead of just three. Although such an extension could be valuable for our application and other complex experiments, it is, unfortunately, beyond the scope of our current paper and will be pursued as future work. Following another referee's suggestion, we also provide an additional simulation study in the supplementary material, where the eigenvalues and eigenfunctions are set to be the estimates from the real data and the true numbers of principal components are set to explain 99% of the variation in each level of the hierarchical model for the real data. In this setting, the true underlying functional random effects have long FPCA series with 4, 6 and 5 components respectively, and some nonzero eigenvalues are extremely small. Under such an extreme setting, the PVE method picks the right model with slightly higher chance than AIC when the threshold is set at the ideal level 99%. In practice, when there is no prior knowledge on the best choice for the percentage-of-variation threshold, we believe AIC is the better method. We also compare the prediction performance of the models picked by the two methods in independently generated test data sets. On average, the model selected by AIC has smaller integrated mean squared error when predicting the underlying genotype, file and seed random processes in the test data.

Supplementary Materials

The supplementary materials contain additional figures for the simulation study and data analysis. An additional simulation study with long FPCA series is also presented in the supplementary material.

Acknowledgment

This paper is based on part of Xu's Ph.D. dissertation under Li's supervision. Li's research was supported by the National Science Foundation award DMS-1317118. Nettleton's research was partially supported by Iowa State University Plant Sciences Institute Scholars Program. We thank Dr. Edgar Spalding of University of Wisconsin at Madison for providing the Root Image Study data. We also thank the Associate Editor and three anonymous referees for their helpful and constructive comments, which lead to significant improvement in this paper.

References

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93:961–976.

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

Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests (with discussion). *Test*, 16:409–478.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, 29:153–193.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71:344–353.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22:361–411.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34:1493–1517.

Hall, P., Müller, H. G., and Yao, F. (2008). Modeling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society. Series B*, 70:703–723.

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

Li, H., Keadle, S. K., J., S., Assaad, H., Huang, J. Z., and Carroll, R. J. (2015). Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, 16:754–771.

Li, Y. (2011). Efficient semiparametric regression in longitudinal data with nonparametric covariance estimation. *Biometrika*, 98:355–370.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 38:3321–3351.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108:1284–1294.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96:1045–1056.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21:255–285.

Noh, B., Bandyopadhyay, A., Peer, W. A., Spalding, E. P., and Murphy, A. S. (2003). Enhanced gravi-and phototropism in plant mdr mutants mislocalizing the auxin efflux protein PIN1. *Nature*, 423:999–1002.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis (2nd ed.)*. Springer-Verlag, New York.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Serban, N. and Jiang, H. (2012). Multilevel functional clustering analysis. *Biometrics*, 68:805–814.

Serban, N., Staicu, A., and Carroll, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics*, 69:903–913.

Shou, H., Zipunnikov, V., Crainiceanu, C. M., and Greven, S. (2015). Structured functional principal component analysis. *Biometrics*, 71:247–257.

Tang, J., Li, Y., and Guan, Y. (2016). Generalized quasi-likelihood ratio tests for semiparametric analysis of covariance models in longitudinal data. *Journal of the American Statistical Association*, 111:736–747.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B*, 62:413–428.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society. Series B*, 68:3–25.

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

Zhou, L., Huang, J., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105:390–400.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95:601–619.

Table 1: Empirical distributions for the number of selected principal components $\widehat{p}$ for the three levels of hierarchy using various methods.

Criteria	Level	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}=5$	All levels
Scenario I: $(p_1, p_2, p_3) = (2, 2, 2)$							
AIC	1	0.000	**1.000**	0.000	0.000	0.000	**0.975**
	2	0.000	**0.975**	0.025	0.000	0.000	
	3	0.000	**1.000**	0.000	0.000	0.000	
PVE 85%	1	0.005	**0.970**	0.025	0.000	0.000	0.000
	2	0.000	**0.815**	0.185	0.000	0.000	
	3	**1.000**	0.000	0.000	0.000	0.000	
PVE 90%	1	0.000	**0.930**	0.070	0.000	0.000	0.095
	2	0.000	**0.685**	0.315	0.000	0.000	
	3	**0.840**	0.160	0.000	0.000	0.000	
PVE 95%	1	0.000	**0.570**	0.430	0.000	0.000	0.225
	2	0.000	**0.525**	0.475	0.000	0.000	
	3	0.000	**1.000**	0.000	0.000	0.000	
PVE 99%	1	0.000	0.080	**0.695**	0.225	0.000	0.000
	2	0.000	0.025	**0.510**	0.450	0.015	
	3	0.000	**1.000**	0.000	0.000	0.000	
Scenario II: $(p_1, p_2, p_3) = (1, 1, 4)$							
AIC	1	**1.000**	0.000	0.000	0.000	0.000	**1.000**
	2	**1.000**	0.000	0.000	0.000	0.000	
	3	0.000	0.000	0.000	**1.000**	0.000	
PVE 85%	1	0.370	**0.625**	0.005	0.000	0.000	0.000
	2	**0.640**	0.360	0.000	0.000	0.000	
	3	0.000	0.000	**1.000**	0.000	0.000	
PVE 90%	1	0.145	**0.785**	0.070	0.000	0.000	0.000
	2	0.345	**0.645**	0.010	0.000	0.000	
	3	0.000	0.000	**1.000**	0.000	0.000	
PVE 95%	1	0.010	**0.555**	0.430	0.005	0.000	0.000
	2	0.050	**0.760**	0.190	0.000	0.000	
	3	0.000	0.000	0.000	**1.000**	0.000	
PVE 99%	1	0.000	0.005	0.270	**0.660**	0.065	0.000
	2	0.000	0.035	0.460	**0.495**	0.010	
	3	0.000	0.000	0.000	**1.000**	0.000	

Note: The empirical distributions above are based on 200 simulation runs. The last column contains the frequency of choosing the correct numbers of principal components in all levels.

Table 2: Estimated eigenvalues for the three levels for the RIS data. "percent var" means the percentage of variance explained within the level of hierarchy, and "cum percent var" means the cumulative percentage of variance explained within the level.

	Level 1	Level 2	Level 3			
Component	1	1	1	2	3	4
Eigenvalue	0.324	0.101	0.699	0.409	0.132	0.079
Percent var	56.515	50.854	51.226	29.996	9.702	5.804
Cum percent var	56.515	50.854	51.226	81.222	90.924	96.728

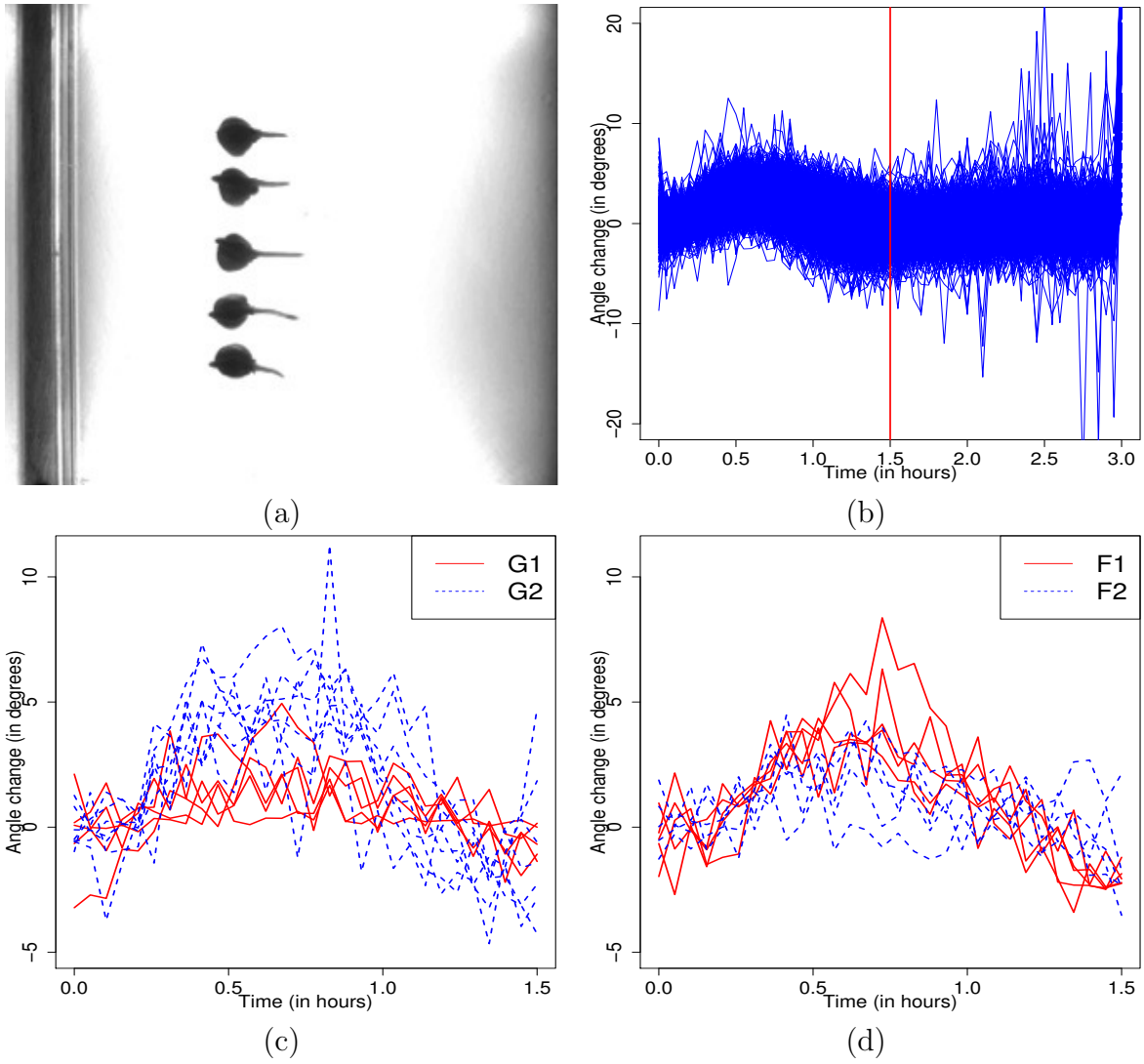
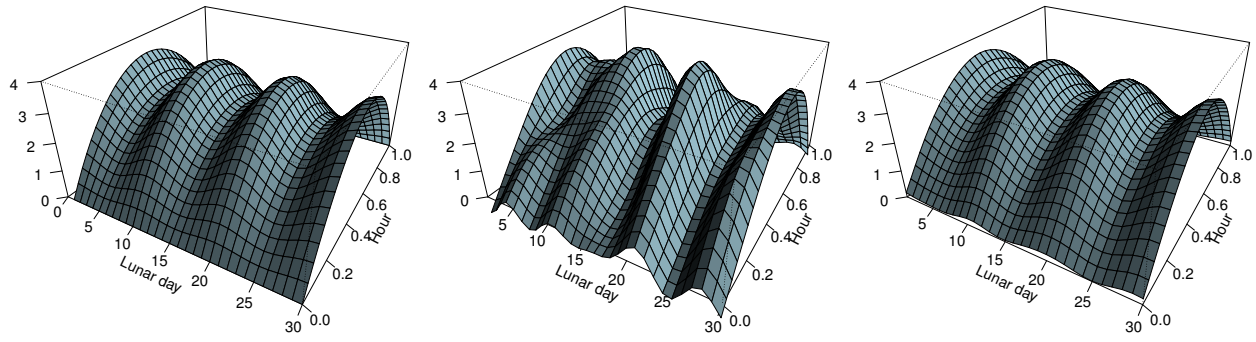
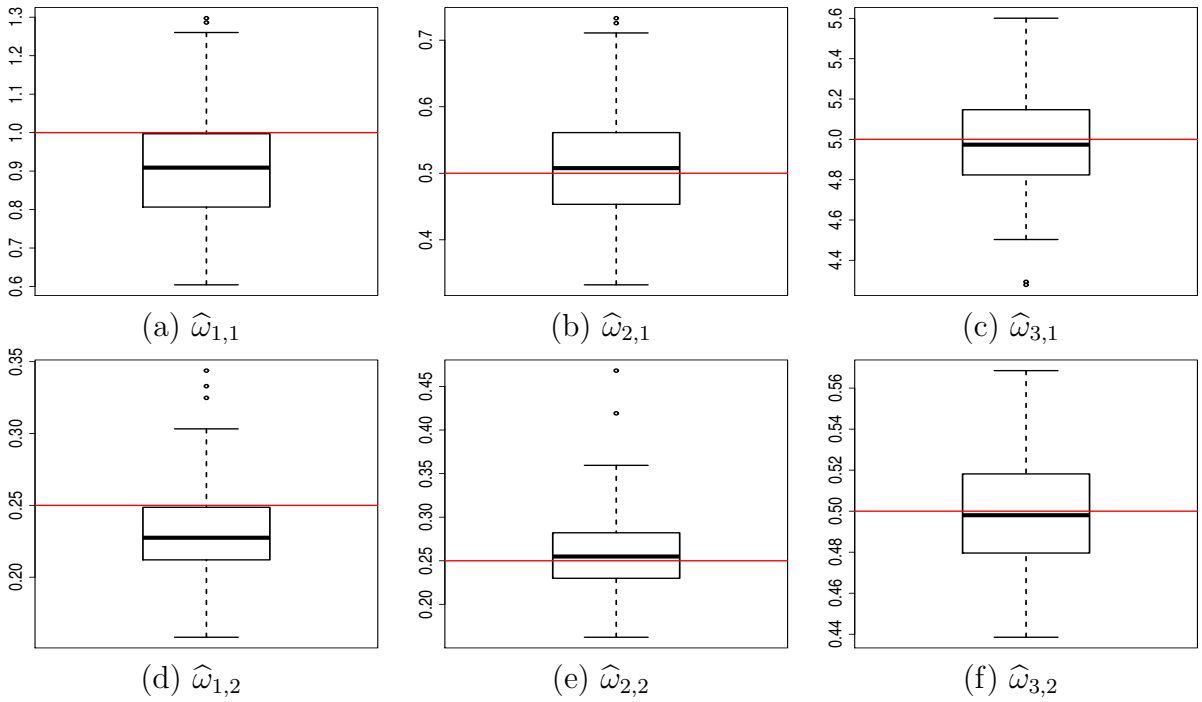


Figure 1: The root gravitropism data. Panel (a) shows an image of a group of seeds under an experimental setup; panel (b) shows the bending rate process for all seeds during the 3-hour experiment time; panel (c) shows the process in the first 1.5 hours for seeds from two randomly selected genotypes; panel (d) shows the process in the first 1.5 hours for seeds from two files within a particular same genotype.



(a) (b) (c)

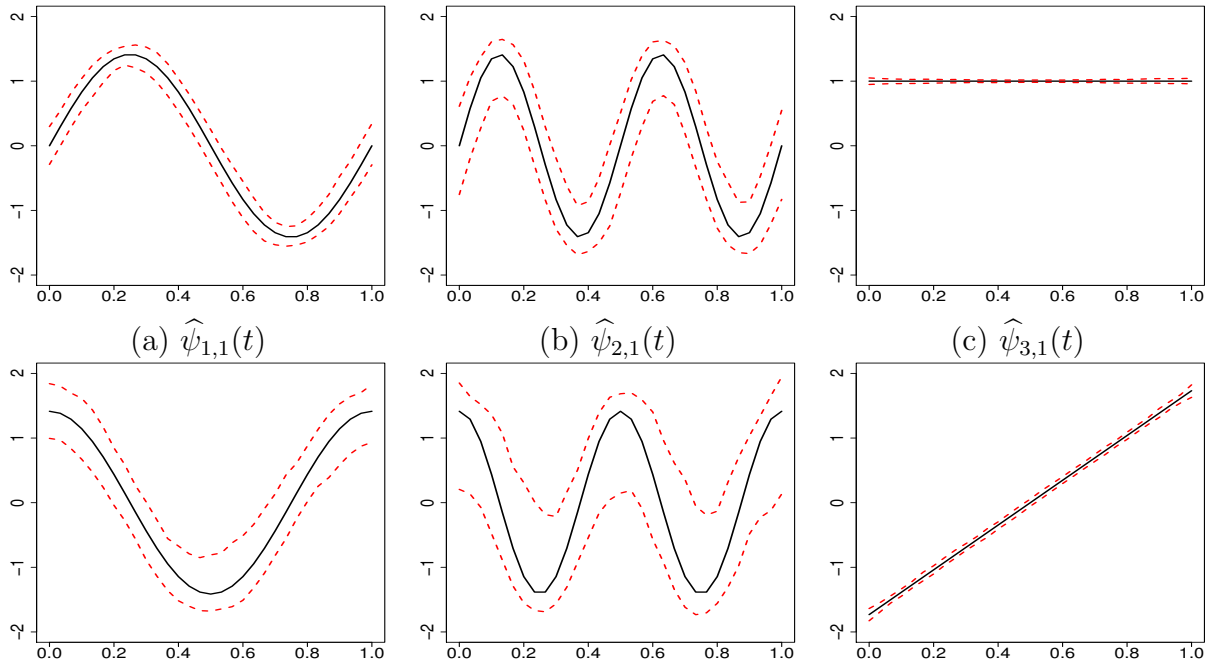
Figure 2: Estimation of the mean surface based on 200 simulations under Scenario I. Panel (a) is the true mean surface; Panel (b) is the estimated mean surface based on one simulation; Panel (c) is the estimated mean surface based on the average of 200 simulations.



(a) $\widehat{\omega}_{1,1}$ (b) $\widehat{\omega}_{2,1}$ (c) $\widehat{\omega}_{3,1}$

(d) $\widehat{\omega}_{1,2}$ (e) $\widehat{\omega}_{2,2}$ (f) $\widehat{\omega}_{3,2}$

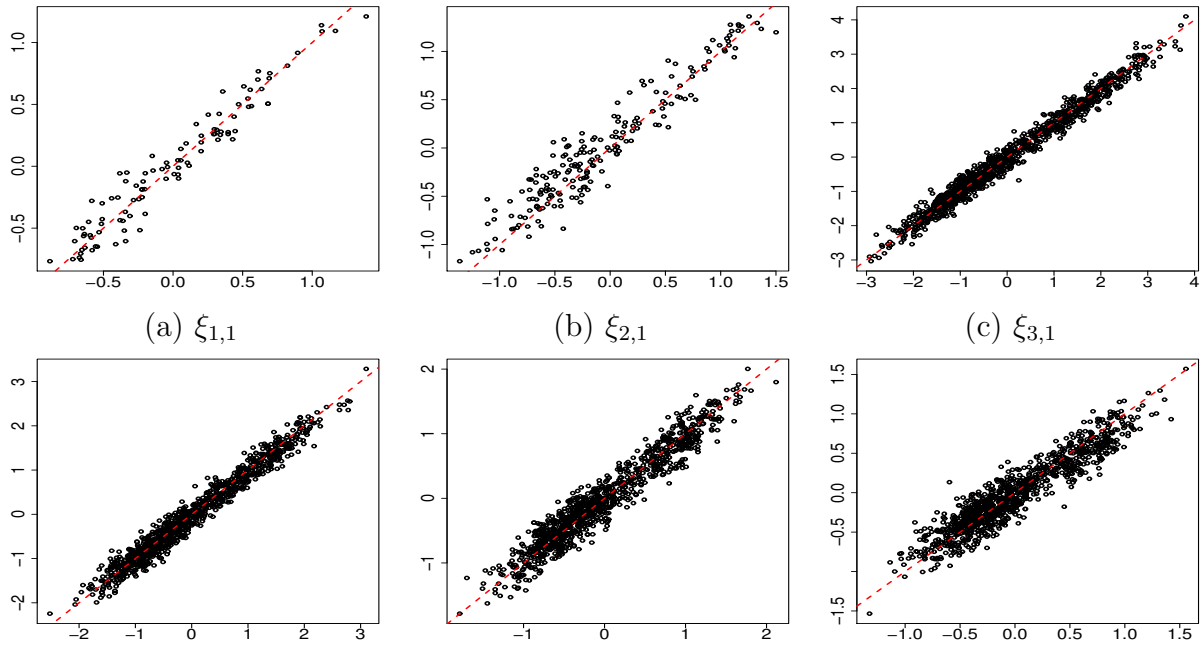
Figure 3: Boxplots of the estimated eigenvalues based on the 200 simulations under Scenario I. The horizontal lines mark the true eigenvalues.



(a) $\widehat{\psi}_{1,1}(t)$ (b) $\widehat{\psi}_{2,1}(t)$ (c) $\widehat{\psi}_{3,1}(t)$

(d) $\widehat{\psi}_{1,2}(t)$ (e) $\widehat{\psi}_{2,2}(t)$ (f) $\widehat{\psi}_{3,2}(t)$

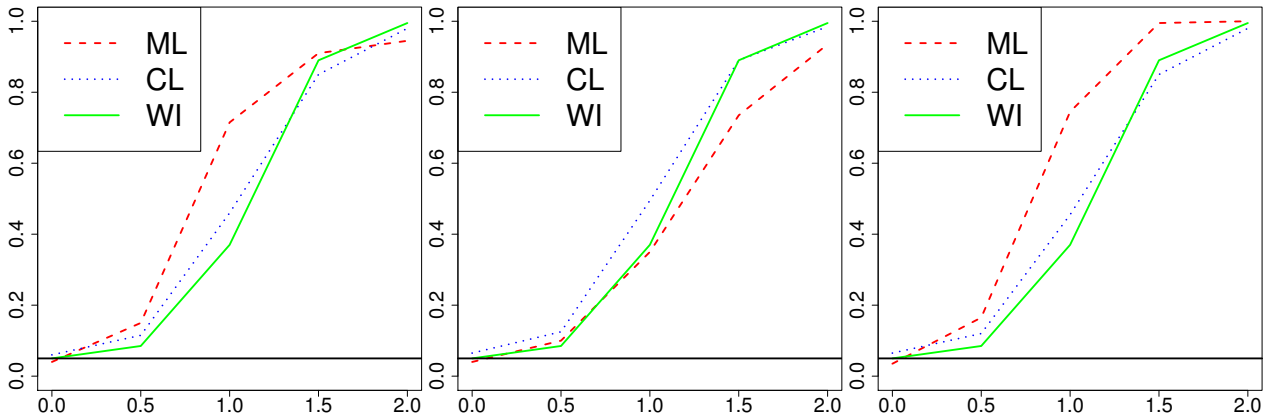
Figure 4: Performance of estimated eigenfunctions under Scenario I. In each panel the solid curve is the true eigenfunction, and the two dashed curves are the 2.5% and 97.5% point-wise quantiles of the estimated eigenfunction based on the 200 simulations.



(a) $\xi_{1,1}$ (b) $\xi_{2,1}$ (c) $\xi_{3,1}$

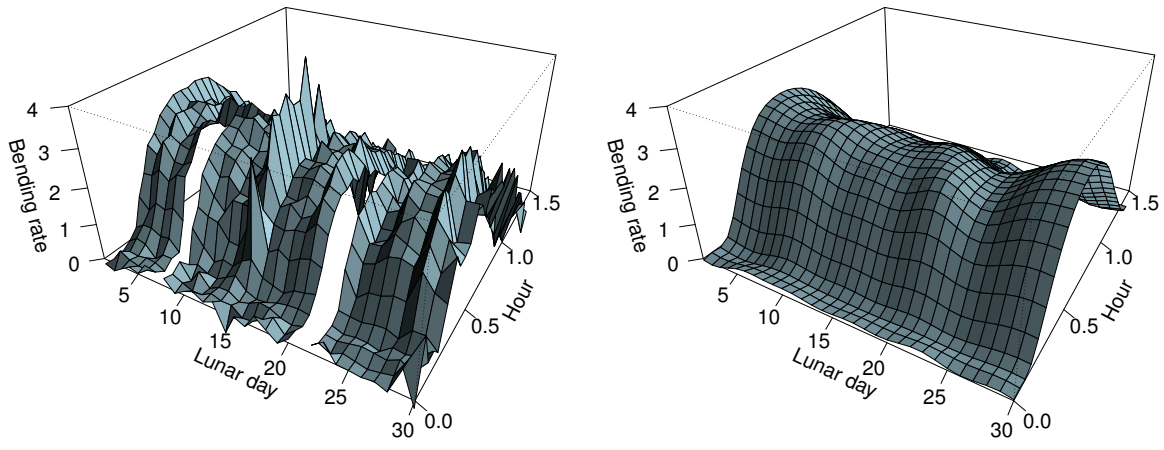
(d) $\xi_{3,2}$ (e) $\xi_{3,3}$ (f) $\xi_{3,4}$

Figure 5: Predicted principal component scores against true principal component scores for the first simulated data set under Scenario II. The dashed lines are 45 degree reference lines.



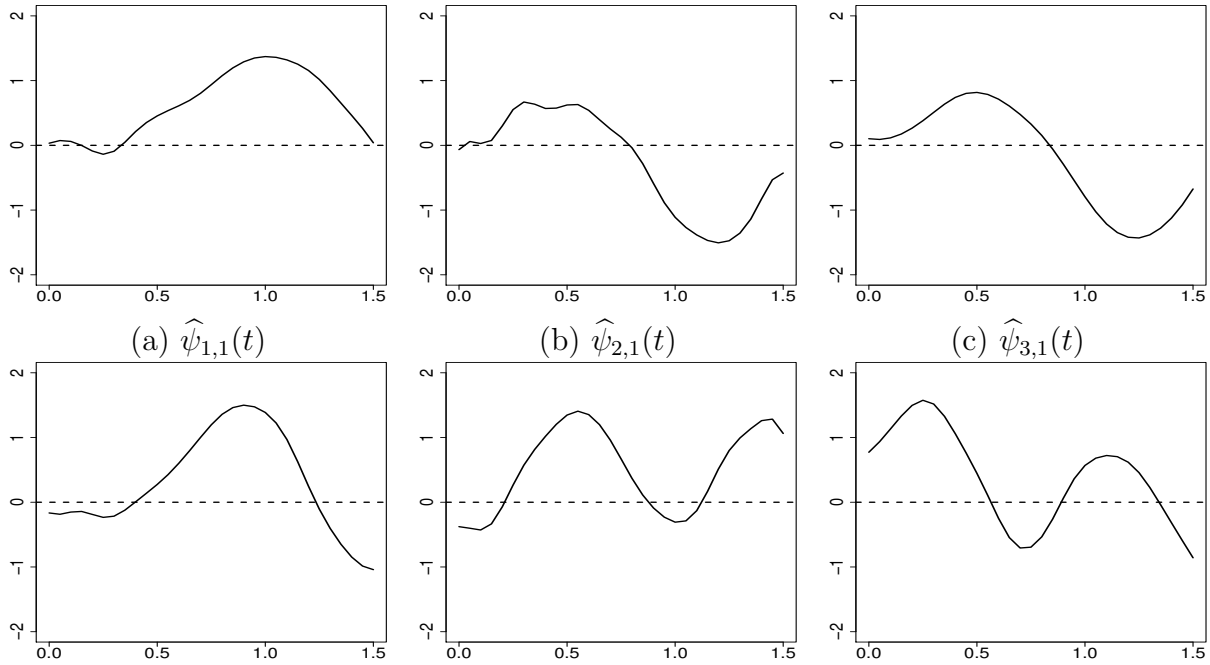
(a) (b) (c)

Figure 6: Performance and sensitivity of the GLR tests. X-axis: δ; y-axis: empirical power based on 200 simulations; solid black line: the reference line at 0.05; dashed red line: marginal likelihood (ML) based GLR test; dotted blue line: conditional likelihood (CL) based GLR test; solid green line: working independent (WI) GLR test. Panel (a): model is correctly specified with $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (2, 2, 2)$; Panel (b): the numbers of principal components are incorrectly specified as $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (1, 1, 1)$; Panel (c): the numbers of principal components are incorrectly specified as $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (3, 3, 3)$.



(a) (b)

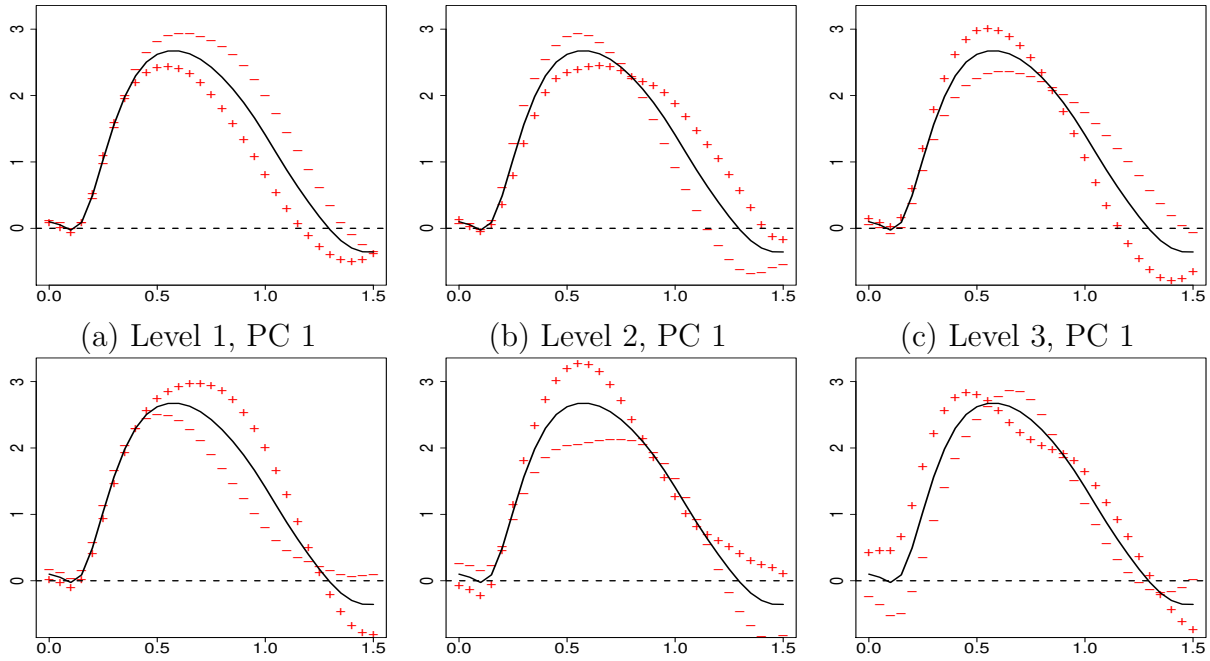
Figure 7: The mean surface estimates based on the RIS data. (a) is the empirical mean surface; (b) is the estimated mean surface using our proposed method.



(a) $\widehat{\psi}_{1,1}(t)$ (b) $\widehat{\psi}_{2,1}(t)$ (c) $\widehat{\psi}_{3,1}(t)$

(d) $\widehat{\psi}_{3,2}(t)$ (e) $\widehat{\psi}_{3,3}(t)$ (f) $\widehat{\psi}_{3,4}(t)$

Figure 8: Estimated eigenfunctions in the RIS data.



(a) Level 1, PC 1 (b) Level 2, PC 1 (c) Level 3, PC 1

(d) Level 3, PC 2 (e) Level 3, PC 3 (f) Level 3, PC 4

Figure 9: Effect plots of the eigenfunctions. Each plot is the mean bending rate curve plus ("+") or minus ("−") a suitable multiple of the eigenfunction.