Reduced Rank Mixed Effects Models for Spatially Correlated Hierarchical Functional Data

Lan Zhou, Jianhua Z. Huang, Josue G. Martinez,
Arnab Maity, Veerabhadran Baladandayuthapani and Raymond J. Carroll

Summary

Hierarchical functional data are widely seen in complex studies where sub-units are nested within units, which in turn are nested within treatment groups. We propose a general framework of functional mixed effects model for such data: within unit and within sub-unit variations are modeled through two separate sets of principal components; the sub-unit level functions are allowed to be correlated. Penalized splines are used to model both the mean functions and the principal components functions, where roughness penalties are used to regularize the spline fit. An EM algorithm is developed to fit the model, while the specific covariance structure of the model is utilized for computational efficiency to avoid storage and inversion of large matrices. Our dimension reduction with principal components provides an effective solution to the difficult tasks of modeling the covariance kernel of a random function and modeling the correlation between functions. The proposed methodology is illustrated using simulations and an empirical data set from a colon carcinogenesis study. Supplemental materials are available online.

Some Key Words: Correlated functions, Functional data, Longitudinal data, Mixed effects models, Penalized splines, Principal components, Reduced rank models.

Short Title: Correlated Hierarchical Functional Data

First page footnote:

Lan Zhou is Assistant Professor (Email: lzhou@stat.tamu.edu), Jianhua Huang is Professor (Email: jianhua@stat.tamu.edu), Josue G. Martinez is Research Assistant Professor (Email: jgmartinez@stat.tamu.edu), and Raymond Carroll is Distinguished Professor (Email: carroll@stat.tamu.edu), Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Arnab Maity is Postdoctoral Fellow (Email: amaity@hsph.harvard.edu), Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115. Veerabhadran Baladandayuthapani is Assistant Professor (Email: veera@mdanderson.org), Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 447, Houston, Texas 77030-4009. Zhou and Martinez were supported by a postdoctoral training grant from the National Cancer Institute (CA90301). Zhou was also supported by NSF grant DMS-0907170. Huang was partially supported by NSF grants DMS-0606580, DMS-0907170 and the NCI grant CA57030. Carroll was partially supported by NCI grants CA57030, CA104620. Huang and Carroll's work was also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

1 Introduction

The goal of this paper is to develop a new methodology for modeling hierarchical, spatially correlated functional data, where sub-units are nested within units, which in turn are nested within treatment groups, and the functions are allowed to be correlated at the sub-unit level. There is an extensive literature for modeling independent functional data with various levels of hierarchies; see for example, Shi, et al. (1996), Grambsch et al. (1995), Brumback and Rice (1998), Staniswallis and Lee (1998), Wang (1998), Wang and Wahba (1998), Rice and Wu (2001), Wu and Zhang (2002), Liang, et al. (2003), Morris et al. (2003), Wu and Liang (2004), Yao et al. (2005a; 2005b), Morris and Carroll (2006), Di et al. (2008), among many others. Recently, Baladandayuthapani, et al. (2008) proposed a Bayesian model for hierarchical, spatially correlated functional data. However, the methodology for dealing with such data is still in its infancy and further development is needed.

In modeling spatially correlated hierarchical functional data we face two major challenges. One is the specification of the covariance structure of random functions. Because of the high dimensionality, it is not feasible for statistical estimation to employ an unstructured covariance specification, which is flexible but requires a large number of unknown parameters. Therefore dimension reduction of some sort is necessary. Some useful covariance specifications with simpler structure have been considered in the literature. For example, diagonal covariance matrices have been used to specify part of the covariance structure by Morris et al. (2003), Morris and Carroll (2006), Baladandayuthapani, et al. (2008). However, such simple structure is often too restrictive in real data analysis. Another challenge is modeling spatial correlation of random functions. This is much less studied in the literature. In the only work we know, Baladandayuthapani, et al. (2008) used a space-time separable covariance structure: while their work is an important step forward, this assumption is hard to justify for real data.

To overcome these challenges, we propose to use functional principal components for both dimension reduction and modeling the spatial correlation of sub-unit level functions. In our functional mixed effects model, an observed functional data object is decomposed as the sum of a fixed treatment effect, a unit level random effect representing unit specific deviation from the treatment effect, a sub-unit level random effect representing sub-unit specific deviation from the unit effect and an error term. The covariance structures of the unit level and sub-unit level random effects are modeled using different sets of principal components and the spatial correlation of sub-unit level random functions is modeled through the spatial correlation of the principal component scores. We use penalized splines to model the mean functions and the principal components functions at both the unit and the sub-unit level. Our approach has two substantial methodological advantages over the existing approach of Baladandayuthapani, et al. (2008): First, a more flexible covariance structure is adapted in our model since principal components instead of pre-specified basis functions are used to represent functional data. Second, we allow a non-separable correlation structure when more than one sub-unit level principal components are used, see Section 2.5 for a detailed discussion. This paper also makes a contribution to modeling hierarchical independent functional data. When no spatial correlation is specified, our modeling framework provides a random effects model alternative to the fixed effects model of Brumback and Rice (1998). In particular, use of two sets of principal components for dimension reduction and modeling covariance structure is new in this context.

The idea of using functional principal components is not new, but most existing work is restricted to independent functions (e.g., Rice and Silverman 1991, Silverman 1996, James et al. 2000, Rice and Wu 2001, Yao, et al. 2005a, 2005b, Di et al. 2009). Recently, Zhou et al. (2008) used principal components to model the correlation of a pair of functions. This paper extends that work in several important aspects. First, the three-level hierarchical data structure presented here is much more complex. Secondly, instead of a pair of correlated functions, a collection of spatially correlated, possibly irregularly positioned functions are considered in this paper. In addition, the complexity of data and model structure introduces substantial computational challenges. Naive extension of Zhou et al. (2008) is theoretically sound but computationally impractical and causes problems with computer storage and computational time. In this paper, we have developed specific techniques to circumvent such difficulties by avoiding storage and computation of large matrices, see the supplemental materials for details.

Our work is motivated by analyzing data from an experiment using rodent models to investigate the role of p27, an important cell-cycle mediator, in early colon carcinogenesis. To investigate the mechanisms by which diet modulates colon tumor development, rats were fed particular diets of interest for specific periods, exposed to a carcinogen inducing colon cancer and subsequently euthanized for sample collection. The colon was then resected from these rats and colonic cells were examined for response of interest. Colonic cells replicate and grow completely within discrete units called crypts which are finger-like structures that grow into the wall of the colon. The need of a model for spatially correlated hierarchical functional data comes from the following three aspects of the data. First, the data are inherently functional in nature since responses from the cells within each crypt can be viewed as a function of the cell position. Second, the experimental data have a three-level hierarchy: crypts are nested within rats, and rats within treatment (diet) groups. In addition, although the rats were independent samples, there may exist spatial correlation at the deepest level of the hierarchy, since within a rat one crypt may behave in accordance with its neighboring crypts.

The rest of the paper is organized as follows. In Section 2, we introduce our model for hierarchical spatially correlated functional data. Section 3 deals with the estimation of model parameters and inference. Section 4 discusses some model specification issues including tuning parameter selection for penalized splines and choice of the number of principal components in modeling random functions. Section 5 gives a simulation study that compares our method with that of Baladandayuthapani, et al. (2008). In Section 6 we show the application of our proposed model and method to the colon carcinogenesis data. Section 7 concludes the paper.

2 The Model

Due to the complexity of the data structure, we divide our model specification into four parts: Section 2.1 presents the basic form of our hierarchical mixed effects model of functions; Section 2.2 introduces the dimension reduction for modeling the covariance structure of random functions; Section 2.3 describes how to model the correlation of functions at the sub-unit level of the hierarchy and summarizes the overall covariance structure of the data; Section 2.4 discusses spline models of functions and gives conditions for parameter identifiability. Section 2.5 compares the proposed approach with the Bayesian approach of Baladandayuthapani, et al. (2008).

2.1 Data Structure and Hierarchical Mixed Effects Model

We consider functional data that have a natural hierarchical structure. At the top level of the hierarchy, there are treatment groups; within treatment groups, there are experimental or sampling units and nested within these units are sub-units. For the colon carcinogenesis data we analyze in Section 6, the treatment groups correspond to different diets and time after exposure to the carcinogen, the experimental units are rats and sub-units are colon crypts.

A multilevel model that takes into account the hierarchy is the following:

$$\begin{aligned} Y_{abc}(t) &= \mu_{abc}(t) + \epsilon_{abc}(t), \\ \mu_{abc}(t) &= \mu_{ab}(t) + \eta_{abc}(t), \qquad (1) \\ \mu_{ab}(t) &= \mu_{a}(t) + \xi_{ab}(t), \end{aligned}$$

where t is a generic argument represents the evaluation point of the underlying function, such as the cell position on a crypt or time, and $Y_{abc}(t)$ is the observation of the quantity of interest at t for sub-unit c from unit b of treatment group a. The functions $\mu_{abc}(\cdot)$, $\mu_{ab}(\cdot)$, and $\mu_a(\cdot)$ represent true underlying functions of t for an individual sub-unit, unit and treatment level, respectively. We treat $\mu_a(\cdot)$ as a fixed effect, modeled as a fixed smooth function; $\xi_{ab}(\cdot)$ and $\eta_{abc}(\cdot)$ are random effects, modeled as realizations from zero mean random processes, and $\epsilon_{abc}(\cdot)$ are white noise processes. Model (1) can be written succinctly as

$$Y_{abc}(t) = \mu_a(t) + \xi_{ab}(t) + \eta_{abc}(t) + \epsilon_{abc}(t). \tag{2}$$

The functions $Y_{abc}(t)$ are usually observed on a finite number of points t, and the sets of observation points usually vary from sub-unit to sub-unit. The covariance kernels of the random processes $\xi_{ab}(\cdot)$ and $\eta_{abc}(\cdot)$ are bivariate functions, a direct nonparametric fit of which is difficult since $\xi_{ab}(\cdot)$ and $\eta_{abc}(\cdot)$ are latent processes and thus not observable.

2.2 Dimension Reduction with Principal Components

We assume that the important mode of variation of the processes $\xi_{ab}(\cdot)$ and $\eta_{abc}(t)$ can be summarized by a few principal components (PCs). Specifically, for a given treatment group a, we assume the variation of $\xi_{ab}(\cdot)$ among units within this treatment group is summarized by a set of K_ξ PCs $\{f_j(\cdot)\}$. This suggests the reduced rank model

$$\xi_{ab}(t) = \sum_{j=1}^{K_\xi} f_j(t)\alpha_{abj}, \tag{3}$$

where α_{abj} are the unit level PC scores, which are assumed to be components of a random vector from a multivariate normal distribution with mean 0 and diagonal covariance matrix $\mathbf{D}_{\alpha,a}$, $f_j(t)$ are PC functions subject to the orthogonality constraint $\int f_j f_l \, dt = \delta_{jl}$ for $j, l = 1, \cdots, K_\xi$, δ_{jl} is the Kronecker delta.

Similarly, for a fixed unit b in treatment group a, we assume the variation of $\eta_{abc}(\cdot)$ is summarized by a set of K_η PCs $\{g_j(\cdot)\}$ and have the model

$$\eta_{abc}(t) = \sum_{j=1}^{K_\eta} g_j(t)\beta_{abcj}, \tag{4}$$

where β_{abcj} are the sub-unit level PC scores, which are assumed to be components of a random vector from a multivariate normal distribution with mean 0 and diagonal covariance matrix $\mathbf{D}_{\beta,a}$. The PC functions $g_j(t)$ are subject to the orthogonality constraint $\int g_j g_l \, dt = \delta_{jl}$, $j, l = 1, \cdots, K_\eta$.

Use of a few PCs in (3) and (4) effectively reduces the dimensionality of the random effects processes. The difficult task of modeling the covariance kernel of a stochastic process is reduced to modeling the covariance matrix of a low dimensional vector. After dimension reduction with PCs, model (2) becomes

$$\begin{aligned} Y_{abc}(t) &= \mu_a(t) + \sum_{j=1}^{K_\xi} f_j(t)\alpha_{abj} + \sum_{j=1}^{K_\eta} g_j(t)\beta_{abcj} + \epsilon_{abc}(t) \\ &= \mu_a(t) + \boldsymbol{f}(t)^{\mathrm{T}}\boldsymbol{\alpha}_{ab} + \boldsymbol{g}(t)^{\mathrm{T}}\boldsymbol{\beta}_{abc} + \epsilon_{abc}(t), \end{aligned} \tag{5}$$

where $\boldsymbol{f} = (f_1, \ldots, f_{K_\xi})^{\mathrm{T}}$, $\boldsymbol{g} = (g_1, \ldots, g_{K_\eta})^{\mathrm{T}}$, $\boldsymbol{\alpha}_{ab} = (\alpha_{ab1}, \ldots, \alpha_{abK_\xi})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{abc} = (\beta_{abc1}, \ldots, \beta_{abcK_\eta})^{\mathrm{T}}$. Here, the mean functions $\mu_a(\cdot)$ and the PC functions $f_1(\cdot), \ldots, f_{K_\xi}(\cdot)$, $g_1(\cdot), \ldots, g_{K_\eta}(\cdot)$ are fixed but unknown and need to be estimated, the random effects $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{abc}$ are also unknown and will be treated as missing data when fitting the model. Specification of the joint distribution of the random effects will be given in Section 2.3.

Our formulation and methodology can be extended in a straightforward manner to allow the PC functions to vary among treatment groups. Such extension only involves some notational complication. We will not discuss such extension for simplicity of the presentation.

2.3 Modeling Correlations

With the assumption that all random components in (5) come from normal distributions, we only need to specify the covariance structures to complete the model specification. We assume the unit level random effects are independent and the sub-unit level random effect functions are spatially correlated through the correlation of PC scores. To be specific, we assume that the scores of each PC are realizations of a spatially stationary process. Let x_{abc} be the physical location of sub-unit c from unit b in treatment group a. It is assumed that for each j, $\mathrm{corr}(\beta_{abcj}, \beta_{abc'j}) = \rho(d_{cc'}; \theta_{aj})$, where $\rho(\cdot)$ is a correlation function with a parameter vector θ_{aj} and $d_{cc'} = |x_{abc} - x_{abc'}|$ is the Euclidean distance between the sub-units c and c'. Any parametric family of correlation functions can be used for $\rho(\cdot, \cdot)$ (Stein 1999).

One choice of correlation functions is the Matérn family (Handcock and Stein 1993; Stein 1999), which is used in our numerical examples. The Matérn isotropic autocorrelation function has the general form

$$\rho(d; \phi, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{2d\nu^{1/2}}{\phi} \right)^{\nu} K_\nu \left(\frac{2d\nu^{1/2}}{\phi} \right), \qquad \phi > 0, \quad \nu > 0, \tag{6}$$

where $K_\nu(\cdot)$ is the modified Bessel function of order ν. This Matérn function has two relatively independent parameters. The range parameter, $\phi > 0$, controls the rate of decay of the correlation between observations as distance d increases. Large values of ϕ indicate that sites that are relatively far from one another are moderately (positively) correlated. The order parameter ν basically controls the behavior of the autocorrelation function for observations that are separated by small distances.

Our assumptions on the correlation structure can be summarized as follows.

- The unit level PC scores α_{abj}'s are independent with mean 0 and variance $\sigma^2_{\alpha,aj}$ and the covariance $\mathrm{cov}(\alpha_{abj}, \alpha_{a'b'j'}) = 0$ if $a \neq a'$ or $b \neq b'$ or $j \neq j'$. In addition, the α_{abj}'s are independent of the sub-unit level PC scores β_{abcl} and the error $\epsilon_{abc}(t)$;

- The sub-unit level PC scores β_{abcj}'s are mean 0 and independent with the error $\epsilon_{abc}(t)$'s. The covariances are $\mathrm{cov}(\beta_{abcj}, \beta_{abc'j}) = \sigma^2_{\beta,aj}\rho(|x_{abc} - x_{abc'}|; \theta_{aj})$, where $\rho(\cdot)$ is a spatial correlation coefficient depending on the distance between x_{abc} and $x_{abc'}$, and $\mathrm{cov}(\beta_{abcj}, \beta_{a'b'c'j'}) = 0$ if $a \neq a'$ or $b \neq b'$ or $j \neq j'$. Note that the variances of

β_{abcj}'s do not depend on b.

- For the purpose of identifiability, we require that $\sigma^2_{\alpha,a1} > \cdots > \sigma^2_{\alpha,aK_\xi}$ and $\sigma^2_{\beta,a1} > \cdots > \sigma^2_{\beta,aK_\eta}$ for $a = 1$. More details about the identifiability issue will be given in Section 2.4.

- The errors $\epsilon_{abc}(t)$ are mutually independent with mean 0 and constant variance σ^2.

To see that our modeling framework allows non-separable correlation structure (Schabenberger and Gotway 2005, Section 9.2), consider two sub-units c and c' from unit b of treatment group a with physical distance $d_{cc'}$. The covariance of the corresponding sub-unit level functions $\eta_{abc}(t)$ and $\eta_{abc'}(t')$ is

$$\begin{aligned}\mathrm{cov}\{\boldsymbol{g}(t)^T\boldsymbol{\beta}_{abc}, \boldsymbol{g}(t')^T\boldsymbol{\beta}_{abc'}\} &= \boldsymbol{g}(t)^T\mathrm{cov}(\boldsymbol{\beta}_{abc}, \boldsymbol{\beta}_{abc'})\boldsymbol{g}(t') \\ &= \sum_{j=1}^{K_\eta} \sigma^2_{\beta,aj}\rho(d_{cc'}, \theta_{aj})g_j(t)g_j(t'),\end{aligned} \tag{7}$$

which is in general nonseparable, except when there is only one sub-unit PC (i.e., $K_\eta = 1$) or the θ_{aj}'s do not depend on j. Parsimonious sub-models of our general model can be obtained by removing the dependence of the variances of the random effects on the treatment groups and/or by removing the dependence of the correlation parameters on the treatment groups and the PCs. Such parsimonious sub-models are useful when we do not have enough data to support a flexible specification.

2.4 Modeling Functions with Splines

We choose to fit the mean functions and the PC functions using polynomial splines. Use of polynomial splines can be justified by the good approximation properties of regression splines to smooth functions as well-studied in applied mathematics (de Boor 2001). Other basis expansion approximation of functions can also be used in our methodology. Let $\boldsymbol{b}(t) = \{b_1(t), \cdots, b_q(t)\}^{\mathrm{T}}$ be a spline basis with dimension q. Write $\mu_a(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{\mu,a}$, $\boldsymbol{f}(t)^{\mathrm{T}} =$

$\boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\Gamma}_{\xi}$ and $\boldsymbol{g}(t)^{\mathrm{T}} = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\Gamma}_{\eta}$, where $\boldsymbol{\gamma}_{\mu,a}$ is a q-dimensional vector, $a = 1, \cdots, A$, $\boldsymbol{\Gamma}_{\xi} = (\boldsymbol{\gamma}_{\xi,1}, \ldots, \boldsymbol{\gamma}_{\xi,K_{\xi}})$, $\boldsymbol{\Gamma}_{\eta} = (\boldsymbol{\gamma}_{\eta,1}, \ldots, \boldsymbol{\gamma}_{\eta,K_{\eta}})$ are, respectively, a $q \times K_{\xi}$ and a $q \times K_{\eta}$ matrix of spline coefficients. The reduced rank model (5) then takes the form

$$Y_{abc}(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{\mu,a} + \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\Gamma}_{\xi}\boldsymbol{\alpha}_{ab} + \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\Gamma}_{\eta}\boldsymbol{\beta}_{abc} + \epsilon_{abc}(t), \qquad a = 1, \ldots, A,\ b = 1, \ldots, B_a,\ c = 1, \ldots, C_{ab}. \tag{8}$$

For identifiability, we require that $\boldsymbol{b}(t)$, $\boldsymbol{\Gamma}_{\xi}$ and $\boldsymbol{\Gamma}_{\eta}$ satisfy the conditions

$$\int \boldsymbol{b}(t)\boldsymbol{b}(t)^{\mathrm{T}}\,dt = \mathbf{I},\ \boldsymbol{\Gamma}_{\xi}^{\mathrm{T}}\boldsymbol{\Gamma}_{\xi} = \mathbf{I},\ \boldsymbol{\Gamma}_{\eta}^{\mathrm{T}}\boldsymbol{\Gamma}_{\eta} = \mathbf{I}. \tag{9}$$

The equations in (9) imply orthogonal constraints on the PC functions such that

$$\int \boldsymbol{f}(t)\boldsymbol{f}(t)^{\mathrm{T}}\,dt = \boldsymbol{\Gamma}_{\xi}^{\mathrm{T}}\int \boldsymbol{b}(t)\boldsymbol{b}(t)^{\mathrm{T}}\,dt\,\boldsymbol{\Gamma}_{\xi} = \mathbf{I}, \tag{10}$$

and

$$\int \boldsymbol{g}(t)\boldsymbol{g}(t)^{\mathrm{T}}\,dt = \boldsymbol{\Gamma}_{\eta}^{\mathrm{T}}\int \boldsymbol{b}(t)\boldsymbol{b}(t)^{\mathrm{T}}\,dt\,\boldsymbol{\Gamma}_{\eta} = \mathbf{I}. \tag{11}$$

See Appendix 1 of Zhou, Huang and Carroll (2008) on how to construct a spline basis $\boldsymbol{b}(t)$ that satisfies the orthonormal constraints given in (9).

Denote $\mathbf{D}_{\alpha,a} = \mathrm{diag}(\sigma^2_{\alpha,a1}, \ldots, \sigma^2_{\alpha,aK_{\xi}})$ and $\mathbf{D}_{\beta,a} = \mathrm{diag}(\sigma^2_{\beta,a1}, \ldots, \sigma^2_{\beta,aK_{\eta}})$. Only the covariance matrices of $\boldsymbol{\Gamma}_{\xi}\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\Gamma}_{\eta}\boldsymbol{\beta}_{abc}$, which are $\boldsymbol{\Gamma}_{\xi}\mathbf{D}_{\alpha,a}\boldsymbol{\Gamma}_{\xi}^{\mathrm{T}}$ and $\boldsymbol{\Gamma}_{\eta}\mathbf{D}_{\beta,a}\boldsymbol{\Gamma}_{\eta}^{\mathrm{T}}$ respectively, can be identified. To identify $\boldsymbol{\Gamma}_{\xi}$, $\boldsymbol{\Gamma}_{\eta}$, $\mathbf{D}_{\alpha,a}$, and $\mathbf{D}_{\alpha,a}$, we need to impose some restrictions on these parameters, as detailed in the following proposition. The result follows directly from the uniqueness of eigen-decomposition of nonnegative definite matrices.

Proposition 1. Assume $\boldsymbol{\Gamma}_{\xi}^{\mathrm{T}}\boldsymbol{\Gamma}_{\xi} = \mathbf{I}$ and $\boldsymbol{\Gamma}_{\eta}^{\mathrm{T}}\boldsymbol{\Gamma}_{\eta} = \mathbf{I}$. In addition, assume that the first nonzero element of each column of $\boldsymbol{\Gamma}_{\xi}$ and $\boldsymbol{\Gamma}_{\eta}$ is positive. Suppose the variances of elements of $\boldsymbol{\alpha}_{1b}$ and $\boldsymbol{\beta}_{1bc}$ satisfy $\sigma^2_{\alpha,11} > \cdots > \sigma^2_{\alpha,1K_{\xi}}$ and $\sigma^2_{\beta,11} > \cdots > \sigma^2_{\beta,1K_{\eta}}$. Then the model specified by (8) and (9) is identifiable.

In this proposition, the first nonzero element of each column of $\boldsymbol{\Gamma}_{\xi}$ and $\boldsymbol{\Gamma}_{\eta}$ is used to determine the sign at the population level. To minimize the influence by finite sample random fluctuation, in our implementation we have used the elements of the largest magnitude in each column of $\mathbf{\Gamma}_\xi$ and $\mathbf{\Gamma}_\eta$ to determine the sign. In addition, we suggest to let the first group ($a = 1$) be the treatment group with the largest sample size to improve stability of the variance estimates.

2.5 Discussion

Baladandayuthapani, et al. (2008) developed a Bayesian approach for modeling the same kind of hierarchical functional data considered by this paper. In their modeling framework, the component functions $\mu_a(\cdot)$, $\mu_{ab}(\cdot)$ and $\mu_{abc}(\cdot)$ in (1) and (2) are represented using basis expansions

$$\mu_a(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_a, \quad \mu_{ab}(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{ab}, \quad \mu_{abc}(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{abc},$$

where $\boldsymbol{b}(\cdot)$ is a q-dimensional vector of basis functions, and $\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_{ab}, \boldsymbol{\gamma}_{abc}$ are the coefficients in basis expansion. The treatment effects $\boldsymbol{\gamma}_a$ are assumed to be fixed effects and are given a prior $\boldsymbol{\gamma}_a \sim N(0, \mathbf{\Sigma}_1)$. The coefficients $\boldsymbol{\gamma}_{ab}$ for the unit level functions are mutually independent for different units b and are assumed to have a $N(0, \mathbf{\Sigma}_{2a})$ distribution. The coefficients γ_{abc} for the sub-unit level functions are independent across different units b but may be spatially correlated for different sub-units c within the same sub-unit b. It is assumed that marginally $\boldsymbol{\gamma}_{abc} \sim N(0, \mathbf{\Sigma}_{3a})$ and, $\mathrm{cov}(\boldsymbol{\gamma}_{abc}, \boldsymbol{\gamma}_{abc'}) = \rho(d_{cc'}; \theta_a)\mathbf{\Sigma}_{3a}$ where $\rho(\cdot; \theta_a)$ is a correlation function with parameter θ_a, $d_{cc'}$ is the Euclidean distance between sub-units c and c'. The Matérn family of correlation functions is used in their empirical analysis.

If left unstructured, each of the covariance matrices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_{2a}$, and $\mathbf{\Sigma}_{3a}$ has $q(q+1)/2$ unique parameters where q equals the number of basis functions used in the basis expansion. Since q can be relatively large, there is an obvious need for dimension reduction. Baladandayuthapani, et al. (2008) proposed the following approach for dimension reduction. They focused on the truncated power basis of quadratic splines

$$\boldsymbol{b}(t)^{\mathrm{T}} = (1, m_1(t), m_2(t), (t-t_1)_+^2, \ldots, (t-t_k)_+^2), \tag{12}$$

where $(1, m_1(t), m_2(t))$ is an orthonormal basis of quadratic polynomials, and $t_1, \ldots, t_k$ are knots of the splines. The covariance matrices of the coefficient vectors are assumed to have a special block diagonal structure, i.e., $\boldsymbol{\Sigma}_1 = \text{diag}(c\boldsymbol{I}_3, \sigma_1^2 \boldsymbol{I}_k)$, $\boldsymbol{\Sigma}_{2a} = \text{diag}(\boldsymbol{\Sigma}_{2a}^*, \sigma_2^2 \boldsymbol{I}_k)$, $\boldsymbol{\Sigma}_{3a} = \text{diag}(\boldsymbol{\Sigma}_{3a}^*, \sigma_3^2 \boldsymbol{I}_k)$, where c is a large number serving as a non-informative vague prior, $\boldsymbol{I}_k$ is a $k \times k$ identity matrix, $\boldsymbol{\Sigma}_{2a}^*$ and $\boldsymbol{\Sigma}_{3a}^*$ are unstructured 3×3 matrices. Using the proposed structure reduces the number of parameters to 2 for $\boldsymbol{\Sigma}_1$ and to 7 for $\boldsymbol{\Sigma}_{2a}$ and $\boldsymbol{\Sigma}_{3a}$.

There are two limitations of the approach by Baladandayuthapani, et al. (2008). First, the covariance structure is assumed separable, that is,

$$\text{cov}\{\mu_{abc}(t), \mu_{abc'}(t')\} = \boldsymbol{b}(t)^{\mathrm{T}} \text{cov}(\boldsymbol{\gamma}_{abc}, \boldsymbol{\gamma}_{abc'}) \boldsymbol{b}(t')^{\mathrm{T}} = \rho(d_{cc'}; \theta_a)\, \boldsymbol{b}(t)^{\mathrm{T}} \boldsymbol{\Sigma}_{3a} \boldsymbol{b}(t')^{\mathrm{T}}. \tag{13}$$

Separability is a strong assumption that may be hard to justify for a given data set. In comparison, our approach does not impose the separability assumption and thus can provide more reliable inference. Second, the marginal block diagonal covariance structure used by Baladandayuthapani, et al. (2008) is convenient to implement but imposes a usually unrealistic restriction on the data generating process. In contrast, our principal components based dimension reduction can model a broader class of covariance structures. The methodological advantage of our approach is confirmed by simulation results in Section 5.

3 Model Fitting and Inference

3.1 Maximum penalized likelihood

We use the method of penalized maximum likelihood for parameter estimation. Roughness penalties are introduced to regularize the spline fits of functions (Eilers and Marx 1996; Ruppert et al. 2003).

For $a = 1, \ldots, A$, $b = 1, \ldots, B_a$, and $c = 1, \ldots, C_{ab}$, suppose the observation points of $Y_{abc}(\cdot)$ are t_{abcs}, $s = 1, \ldots, n_{abc}$. Denote $Y_{abcs} = Y_{abc}(t_{abcs})$, $\boldsymbol{Y}_{abc} = (Y_{abc1}, \ldots, Y_{abcn_{abc}})^{\mathrm{T}}$, and $\boldsymbol{Y}_{ab} = (\boldsymbol{Y}_{ab1}^{\mathrm{T}}, \ldots, \boldsymbol{Y}_{abC_{ab}}^{\mathrm{T}})^{\mathrm{T}}$. Recall that $\boldsymbol{b}(\cdot)$ is the q-dimensional spline basis vector.

Let $\boldsymbol{b}_{abcs} = \boldsymbol{b}(t_{abcs})$, $\mathbf{B}_{abc} = (\boldsymbol{b}_{abc1}, \dots, \boldsymbol{b}_{abcn_{abc}})^{\mathrm{T}}$, $\mathbf{B}_{ab} = (\mathbf{B}_{ab1}^{\mathrm{T}}, \dots, \mathbf{B}_{abC_{ab}}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbb{B}_{ab} = \mathrm{diag}(\mathbf{B}_{ab1}, \dots, \mathbf{B}_{abC_{ab}})$. Denote $\boldsymbol{\beta}_{ab} = (\boldsymbol{\beta}_{ab1}^{\mathrm{T}}, \dots, \boldsymbol{\beta}_{abC_{ab}}^{\mathrm{T}})^{\mathrm{T}}$. Denote $\boldsymbol{\Gamma}_{\eta,ab} = \mathbf{I}_{C_{ab}} \otimes \boldsymbol{\Gamma}_{\eta}$ where $\mathbf{I}_{C_{ab}}$ is the identity matrix of rank C_{ab}. Model (8) for observed data then can be rewritten as

$$\boldsymbol{Y}_{ab} = \mathbf{B}_{ab}\boldsymbol{\gamma}_{\mu,a} + \mathbf{B}_{ab}\boldsymbol{\Gamma}_{\xi}\boldsymbol{\alpha}_{ab} + \mathbb{B}_{ab}\boldsymbol{\Gamma}_{\eta,ab}\boldsymbol{\beta}_{ab} + \boldsymbol{\epsilon}_{ab}, \tag{14}$$

where $\boldsymbol{\epsilon}_{ab} = (\epsilon_{ab1}(t_{ab11}), \dots, \epsilon_{ab1}(t_{ab1n_{abc}}), \dots, \epsilon_{abC_{ab}}(t_{abC_{ab}1}), \dots, \epsilon_{abC_{ab}}(t_{abC_{ab}n_{abC_{ab}}}))^{\mathrm{T}}$.

If $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{ab}$ were observable, the joint likelihood of $(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab})$ can be factored as

$$L(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab}) = L(\boldsymbol{Y}_{ab}|\boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab})\, L(\boldsymbol{\alpha}_{ab})\, L(\boldsymbol{\beta}_{ab})$$

because of the independence between $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{ab}$. The likelihood of the observed data $\boldsymbol{Y}_{ab}$ is

$$\int L(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab})\, d\boldsymbol{\alpha}_{ab}\, d\boldsymbol{\beta}_{ab}.$$

We estimate model parameters by minimizing the following penalized likelihood criterion

$$-2\sum_{a=1}^{A}\sum_{b=1}^{B_a} \log\left\{\int L(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab})\, d\boldsymbol{\alpha}_{ab}\, d\boldsymbol{\beta}_{ab}\right\} + \lambda_{\mu}\Omega_{\mu}(\{\boldsymbol{\gamma}_{\mu,a}\}) + \lambda_{\xi}\Omega_{\xi}(\boldsymbol{\Gamma}_{\xi}) + \lambda_{\eta}\Omega_{\eta}(\boldsymbol{\Gamma}_{\eta}), \tag{15}$$

where the Ω's are roughness penalties and the λ's are penalty parameters. The roughness penalties will enforce the smoothness of the fitted mean and PC functions.

We define the roughness penalties using the integrated second derivatives of the fitted functions. For the mean functions $\mu_a(t) = \boldsymbol{b}(t)^{\mathrm{T}}\boldsymbol{\gamma}_{\mu,a}$, we use the penalty

$$\Omega_{\mu}(\{\boldsymbol{\gamma}_{\mu,a}\}) = \sum_{a=1}^{A}\int \{\mu_a''(t)\}^2\, dt = \sum_{a=1}^{A} \boldsymbol{\gamma}_{\mu,a}^{\mathrm{T}} \int \boldsymbol{b}''(t)\boldsymbol{b}''(t)^{\mathrm{T}}\, dt\, \boldsymbol{\gamma}_{\mu,a}.$$

Similarly, the penalties for the unit and sub-unit level PCs are, respectively,

$$\Omega_{\xi}(\boldsymbol{\Gamma}_{\xi}) = \sum_{j=1}^{K_{\xi}}\int \{f_j''(t)\}^2\, dt = \sum_{j=1}^{K_{\xi}} \boldsymbol{\gamma}_{\xi,j}^{\mathrm{T}} \int \boldsymbol{b}''(t)\boldsymbol{b}''(t)^{\mathrm{T}}\, dt\, \boldsymbol{\gamma}_{\xi,j}$$

and

$$\Omega_{\eta}(\boldsymbol{\Gamma}_{\eta}) = \sum_{j=1}^{K_{\eta}}\int \{g_j''(t)\}^2\, dt = \sum_{j=1}^{K_{\eta}} \boldsymbol{\gamma}_{\eta,j}^{\mathrm{T}} \int \boldsymbol{b}''(t)\boldsymbol{b}''(t)^{\mathrm{T}}\, dt\, \boldsymbol{\gamma}_{\eta,j}.$$

Although more flexible specification is possible, we use three penalty parameters, one for all mean functions, one for all unit level PC functions, and one for all sub-unit level PC functions. Selection of these penalty parameters will be discussed in Section 4.1.

Minimization of (15) does not have a closed-form solution. To avoid the computation of the high-dimensional integral in (15), we treat $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{ab}$ as missing values and use the EM algorithm (Dempster et al. 1977; Laird and Ware 1982) to compute our parameter estimates. At each iteration of the algorithm, given a set of current guesses of the parameter values, the EM algorithm updates the parameter estimates by minimizing the conditional expectation of the $-2\times$ log likelihood, where the expectation is taken under the distribution whose parameters are set at their current guesses. Details of the algorithm are given in subsequent subsections. Under mild conditions the EM algorithm converges and each iteration of the EM decreases the negative log likelihood of observed data (Wu, 1983). Our adding roughness penalties to the negative log likelihood does not change the convergence properties of the EM algorithm.

3.2 The EM algorithm

Let $N_{ab} = \sum_{c=1}^{C_{ab}} n_{abc}$ and denote $\mathbf{V}_{ab} = \text{cov}(\boldsymbol{\beta}_{ab})$. Then $-2\times$ joint log likelihood of the "complete data" $(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab})$ from sub-unit b of unit a is

$$\begin{aligned}
&-2l(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab}; \{\boldsymbol{\gamma}_{\mu,a}\}, \boldsymbol{\Gamma}_{\xi}, \boldsymbol{\Gamma}_{\eta}, \sigma^2, \{\sigma^2_{\alpha,aj}\}, \{\sigma^2_{\beta,aj}\}, \theta_{aj}) \\
&\quad = N_{ab}\log(2\pi) + N_{ab}\log(\sigma^2) \\
&\qquad + \sigma^{-2}(\boldsymbol{Y}_{ab} - \mathbf{B}_{ab}\boldsymbol{\gamma}_{\mu,a} - \mathbf{B}_{ab}\boldsymbol{\Gamma}_{\xi}\boldsymbol{\alpha}_{ab} - \mathbb{B}_{ab}\boldsymbol{\Gamma}_{\eta,ab}\boldsymbol{\beta}_{ab})^{\mathrm{T}}\times \\
&\qquad\quad (\boldsymbol{Y}_{ab} - \mathbf{B}_{ab}\boldsymbol{\gamma}_{\mu,a} - \mathbf{B}_{ab}\boldsymbol{\Gamma}_{\xi}\boldsymbol{\alpha}_{ab} - \mathbb{B}_{ab}\boldsymbol{\Gamma}_{\eta,ab}\boldsymbol{\beta}_{ab}) \\
&\qquad + K_{\xi}\log(2\pi) + \log(|\mathbf{D}_{\alpha,a}|) + \boldsymbol{\alpha}_{ab}^{\mathrm{T}}\mathbf{D}_{\alpha,a}^{-1}\boldsymbol{\alpha}_{ab} \\
&\qquad + \{C_{ab}K_{\eta}\log(2\pi) + \log(|\mathbf{V}_{ab}|) + \boldsymbol{\beta}_{ab}^{\mathrm{T}}\mathbf{V}_{ab}^{-1}\boldsymbol{\beta}_{ab}\}.
\end{aligned} \tag{16}$$

The correlation parameters θ_{aj} enter the likelihood through $\mathbf{V}_{ab}$. Since sub-units are independent, the $-2\times$ joint log likelihood of the "complete data" $(\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is

$$\begin{aligned} &-2l(\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}; \{\boldsymbol{\gamma}_{\mu,a}\}, \boldsymbol{\Gamma}_{\xi}, \boldsymbol{\Gamma}_{\eta}, \sigma^2, \{\sigma^2_{\alpha,aj}\}, \{\sigma^2_{\beta,aj}\}, \theta_{aj}) \\ &\quad = \sum_{a=1}^{A} \sum_{b=1}^{B_a} \{-2l(\boldsymbol{Y}_{ab}, \boldsymbol{\alpha}_{ab}, \boldsymbol{\beta}_{ab}; \{\boldsymbol{\gamma}_{\mu,a}\}, \boldsymbol{\Gamma}_{\xi}, \boldsymbol{\Gamma}_{\eta}, \sigma^2, \{\sigma^2_{\alpha,aj}\}, \{\sigma^2_{\beta,aj}\}, \theta_{aj})\}. \end{aligned} \tag{17}$$

The E-step of the EM algorithm consists of finding the conditional expectation of (17) given $\boldsymbol{Y}_{ab}$ and the current parameter values. Since the log likelihood is a quadratic form of the random effects $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{ab}$, only their conditional first two moments need to be computed. Details on computing these conditional moments are provided in supplemental materials. Direct calculation of the conditional moments requires the inversion of the matrix $\mathrm{cov}(\boldsymbol{Y}_{ab})$. This matrix has size $N_{ab} \times N_{ab}$ with $N_{ab} = \sum_{c=1}^{C_{ab}} n_{abc}$, which is often very large. In our empirical example, for a typical rat, there are about $C_{ab} = 20$ crypts and $n_{abc} = 30$ observations on each crypt, so the matrix to be inverted is of size 600×600. By repeatedly applying the Sherman-Morrison-Woodbury formula, we developed the computational devices to circumvent the inversion of this matrix; see Supplemental materials for details.

The M-step of the EM algorithm updates the parameter estimates by minimizing the objective function which is the conditional expectation of $-2\times$ log likelihood given in (17), or by reducing the value of this objective function as an application of the generalized EM algorithm. The parameters are well separated in the expression of the conditional log-likelihood, so we can update the parameter estimates sequentially given their current values. We update according to the following order: (1) σ^2; (2) $\sigma^2_{\alpha,ai}$ and $\sigma^2_{\beta,aj}$, $a = 1, \ldots, A$, $i = 1, \ldots, K_\xi$ and $j = 1, \ldots, K_\eta$; (3) $\boldsymbol{\gamma}_{\mu,a}, a = 1, \ldots, A$; (4) $\boldsymbol{\Gamma}_\xi$; (5) $\boldsymbol{\Gamma}_\eta$; (6) the correlation parameter θ_{aj}, $a = 1, \ldots, A$, $j = 1, \ldots, K_\eta$. Details of the updating formula are given in supplemental materials. Note that when updating $\boldsymbol{\Gamma}_\xi$ and $\boldsymbol{\Gamma}_\eta$, some care is needed to enforce the orthonormal constraints.

In our implementation of the EM algorithm, the initial values are obtained using the following procedure. We sequentially fit fixed effects models to obtain raw estimates of the treatment group level, unit level and sub-unit level functions appeared in (2). We first obtain estimates of the treatment effects $\eta_a(t)$ by fitting fixed effects models with working independent covariance structures and use these raw estimates as the initial values of the treatment effects. Next, we remove the treatment effects from the data and obtain the unit level effects $\xi_{ab}(t)$ by fitting separate fixed effects models for each unit. Then, removing both the treatment effects and unit level effects we obtain the sub-unit level effects $\eta_{abc}(t)$ by fitting separate fixed effects models for each sub-unit. After raw estimates of unit level and sub-unit level functions are obtained, the standard principal components analysis is applied to get the initial estimates of functional principal components and associated variances. When obtaining the unit and sub-unit level effects, the ridge regression with a small ridge parameter is used to deal with the singularity problem caused by small sample size in fixed effects regression. The initial value of the error variance is obtained using the sample variance of the residuals after removing the treatment effects, the unit level and sub-unit level functions from the data. The initial values of the correlation parameters are chosen to be the mid-point of a prespecified interval. With starting values generated by this procedure, the EM algorithm works well in our simulation and real data analysis; it usually converges within twenty steps.

3.3 Prediction of random effects functions and inference

Using estimated parameters, the best linear unbiased predictors (BLUP, Henderson 1950) of the random effects $\boldsymbol{\alpha}_{ab}$ and $\boldsymbol{\beta}_{ab}$ are given by $\widehat{\boldsymbol{\alpha}}_{ab} = E(\boldsymbol{\alpha}_{ab}|\boldsymbol{Y}_{ab})$ and $\widehat{\boldsymbol{\beta}}_{ab} = E(\boldsymbol{\beta}_{ab}|\boldsymbol{Y}_{ab})$, whose closed-form expressions are available in the supplemental materials. The BLUP of the unit level random function $\xi_{ab}(t)$ in (2) is $\widehat{\boldsymbol{f}}(t)^T\widehat{\boldsymbol{\alpha}}_{ab}$ and the BLUP of the sub-unit level random function $\eta_{abc}(t)$ is $\widehat{\boldsymbol{g}}(t)^T\widehat{\boldsymbol{\beta}}_{abc}$, where $\widehat{\boldsymbol{f}}(t)$ and $\widehat{\boldsymbol{g}}(t)$ are vectors of estimated unit level and sub-unit level PCs respectively.

The bootstrap (Efron 1979) can be used to obtain standard errors of the parameter estimates. In a nonparametric bootstrap, we resample units without replacement to ensure that the covariance structure in the original sample is preserved in the bootstrap sample. In a parametric bootstrap, we draw bootstrap samples from the model with fitted parameters. After bootstrap samples are drawn, we estimate the model parameters for each bootstrap sample and obtain a collection of resampled estimates. The sample standard deviations of these resampled estimates provide estimates of the desired standard errors. Alternatively, the sample quantiles of the resampled estimates can be used directly to construct bootstrap confidence intervals.

4 Model Selection and Assessment

4.1 Specification of Splines and Penalty Parameters

When using penalized splines to fit smooth functions, degree two or three splines are commonly used, and the placement and number of knots is generally not crucial since the penalty takes care of overfitting (Ruppert 2002). For a typical application, 10–20 knots equally spaced over the data range is often sufficient and fewer knots can be used when the sample size is small or noise level is high.

To choose penalty parameters, one can calculate the crossvalidated score, defined as the crossvalidated $-2\times$ log likelihood, and select the parameters corresponding to the minimum. Details on computing the observed data log likelihood without forming the large covariance matrix $\mathrm{cov}(\boldsymbol{Y}_{ab})$ are given in supplemental materials. To preserve the covariance structure in the data, the data from a whole unit needs to be deleted and serve as a validation set in the crossvalidation. When there are a large number of units, K-fold crossvalidation can be used for computational efficiency. There are three penalty parameters in our method, so we need to search over a three dimensional space for a good choice of these parameters. A multidimensional optimization algorithm can be used to speed up the search. We applied the downhill simplex method of Nelder and Mead (1965) in our implementation of the method.

4.2 Selection of the Number of Significant PCs

To determine the number of important PCs at both the unit level and the sub-unit level, first note that the available data usually set an upper bound on the feasible number of PCs we can fit in the model. If too many PCs are acquired from the EM algorithm, numerical problems may occur such as inversion of singular matrices and the algorithm failing to converge. This is precisely the reason dimension reduction through use of principal components is needed. See James et al. (2000) for a similar discussion in the case of functional data without spatial correlation. Leave-one-unit-out crossvalidation or its K-fold version can be used to decide on the number of significant PCs. This strategy works well in our simulation studies: it can correctly identify the number of PCs in the data generating model most of times, see Section 5.

In actual data analysis, however, the crossvalidation score may keep on decreasing as more PCs are included in the model. This phenomenon, also observed in application of principal components analysis in multivariate analysis, is not surprising since a reduced rank model with a finite number of PCs is typically only an approximation. We can still use crossvalidation, not to identify the "correct" model, but to identify the most parsimonious model that fits the data well. If one arranges the crossvalidation scores according to the increasing complexity of the model, one usually sees a quick drop of the crossvalidation scores, followed by much slower decrease; the turning point suggests the suitable number of PCs. See Section 6 for an illustration of this method.

4.3 Model assessment

Various diagnostic plots can be used to assess whether the number of PCs selected by crossvalidation is sufficient and how well the model fits the data. At the unit level, using too few

PCs will force some unit level effects to be included into the estimates of the sub-unit level effects. Thus deviation from zero of the unit mean of fitted sub-unit level effects for some units indicates an insufficient number of PCs at the unit level. At the sub-unit level, if too few PCs are fitted, a pattern will appear in the unit-wise residual plots. Thus examination of the unit-wise residual plots can help assess whether the number of PCs obtained from crossvalidation is sufficient at the sub-unit level. The normal quantile-quantile plots of fitted random effects and of residuals can be used to assess the distributional assumptions in the model.

5 Simulation

In this section we illustrate the performance of our method using simulated data and compare it with the Bayesian method of Baladandayuthapani, et al. (2008). The two methods differ mainly in the specification of the covariance structure. When setting up the simulation studies, we considered not only generating data from our reduced rank model but also from the model of Baladandayuthapani, et al. (2008). We used the existing computer code from the original paper when applying the Bayesian method.

The errors of estimating the treatment group means are measured using the following integrated absolute errors

$$\int |\widehat{\mu}_a(t) - \mu_a(t)| \, dt,$$

where a denotes the treatment group and the hat indicates estimated values. The errors from all treatment groups are then averaged to get a summary measure. To facilitate comparison of the estimated covariance structure from the two methods, we examined the predictions of the unit level and sub-unit level random effects. These predictions are given by the posterior means of the relevant quantities using the estimated covariance structure. Using the notation in (2) of Section 2.1, the errors of predicting the random effects are measured using the integrated absolute errors

$$\int |\widehat{\xi}_{ab}(t) - \xi_{ab}(t)|\, dt \quad \text{and} \quad \int |\widehat{\eta}_{abc}(t) - \eta_{abc}(t)|\, dt$$

for treatment a, unit b, and sub-unit c, where the hats indicate predicted values. All unit level errors are then averaged to get a summary measure at the unit level. Similarly, all sub-unit level errors are averaged to get a summary measure at the sub-unit level. In our calculations, the integrals were approximated using a Riemann sum with a grid of 20 equally spaced points.

In all simulations, there were two treatment groups, twelve units within each treatment group, twenty sub-units within each unit and twenty observations on each sub-unit. For each unit, the twenty sub-units were located on a line segment with locations independently generated from the uniform distribution on $[0, 14]$. For each sub-unit, the functional response $Y(t)$ was evaluated at twenty points randomly generated from the uniform distribution on $[0, 1]$. Details of setting up the mean and principal components functions, variance and correlation parameters are given below. We ran the simulation 100 times for each setup and used the measures described above to assess/compare the performance of the two methods. When applying our method to the simulated data, we used cubic splines with five interior knots to fit the functions and used crossvalidation to select the penalty parameters.

We first considered three different setups from our reduced rank model where we fixed the forms of the mean and PC functions but varied other parameters of the model. The mean curves for the two treatment groups were

$$\mu_1(t) = 7 - 16t + 30t^2 - 15t^3 \quad \text{and} \quad \mu_2(t) = 8 - 13t + 14t^2 - t^3.$$

The unit level PC functions were

$$f_1(t) = 1.414 \sin(2\pi t) \quad \text{and} \quad f_2(t) = -1.485 + 2.970 \sin(\pi t) \tag{18}$$

and the sub-unit level PC functions were

$$g_1(t) = 1 \quad \text{and} \quad g_2(t) = -1.118 + 3.354\, t^2. \tag{19}$$

The five-fold crossvalidation identified the correct number of PCs in more than 95% of the cases, in each setup.

Setup 1. We used a model with one unit level PC $f_1(t)$ given in (18) and one sub-unit level PC $g_1(t)$ given in (19). At the unit level, the PC score variances $\sigma^2_{\alpha,11} = 0.64$ for treatment group 1 and $\sigma^2_{\alpha,21} = 0.16$ for treatment group 2; at the sub-unit level, the PC score variances $\sigma^2_{\beta,11} = 0.36$ for treatment group 1 and $\sigma^2_{\beta,21} = 0.16$ for treatment group 2. The spatial correlation structure for modeling sub-unit level dependence was the Matérn family (6) with parameters $\phi = 8$ and $\nu = 0.1$. The error variance was set to be $\sigma^2 = 0.01$.

Setup 2. We used a model with two unit level and two sub-unit level PCs given in (18) and (19) respectively. The unit level PC score variances were $\sigma^2_{\alpha,11} = 0.64$ and $\sigma^2_{\alpha,12} = 0.25$ for treatment group 1 and $\sigma^2_{\alpha,21} = 0.16$ and $\sigma^2_{\alpha,22} = 0.04$ for treatment group 2. The sub-unit level PC score variances were $\sigma^2_{\beta,11} = 0.36$ and $\sigma^2_{\beta,12} = 0.09$ for treatment group 1 and $\sigma^2_{\beta,21} = 0.16$ and $\sigma^2_{\beta,22} = 0.04$ for treatment group 2. The spatial correlation structure was the Matérn family (6) with parameters $\phi_1 = 8$, $\phi_2 = 4$, $\nu_1 = 0.1$ and $\nu_2 = 0.3$. The error variance was set to be $\sigma^2 = 0.01$.

Setup 3. This is the same as Setup 2 except that all variance parameters, including PC variances and the error variance, were halved. This setup is equivalent to doubling the sample size of Setup 2. We did not simulate data by doubling the sample size because that would have created a serious computational burden for the Bayesian method.

Next we considered a slight modification of our model where functional data are independent.

Setup 4. This is the same as Setup 2 except that there was no spatial correlation among sub-unit level PCs. We applied our program to this setup to test its performance in a situation it is not designed to handle.

We also considered setups where data were generated from the Bayesian hierarchical models of Baladandayuthapani, et al. (2008). One setup is presented below. The results for two other setups are presented in the supplemental materials.

Setup 5. We used the notation introduced in Section 2.5 where a brief summary of the Bayesian hierarchical model was given. The basis functions corresponding to a three-knot quadratic splines were as in (12) with $m_1(t) = \sqrt{12}(t-0.5)$, $m_2(t) = \sqrt{180}\{(t-0.5)^2-1/12\}$ and knot locations $t_1 = 0.25$, $t_2 = 0.5$ and $t_3 = 0.75$. The fixed treatment effects were $\boldsymbol{\gamma}_1 = (1, 0.5, 1.5, 1, -0.8, 0.3)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_2 = (2, -0.5, -0.3, 1.2, -.4, -0.7)^{\mathrm{T}}$. To specify the unit level random effects distribution, we set $\sigma_2^2 = 0.1$ and

$$\boldsymbol{\Sigma}_{21}^* = \begin{pmatrix} 1.00 & 0.84 & 0.18 \\ 0.84 & 1.44 & 0.54 \\ 0.18 & 0.54 & 0.81 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{22}^* = \begin{pmatrix} 1.440 & 0.756 & 0.240 \\ 0.756 & 0.810 & 0.450 \\ 0.240 & 0.450 & 1.00 \end{pmatrix}.$$

and to specify the sub-unit level random effects distribution, we set $\sigma_3^2 = 0.1$ and

$$\boldsymbol{\Sigma}_{31}^* = \begin{pmatrix} 1.000 & 0.240 & 0.180 \\ 0.240 & 1.440 & 0.648 \\ 0.180 & 0.648 & 0.810 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{32}^* = \begin{pmatrix} 1.000 & 0.180 & 0.260 \\ 0.180 & 0.810 & 0.702 \\ 0.260 & 0.702 & 1.690 \end{pmatrix}.$$

The parameters of the Matérn family (6) were set to $\phi = 0.57$ and $\nu = 0.11$. The noise variance was set to be $\sigma^2 = 0.01$. We ran our method with three unit and sub-unit level principal components, a specification of our model that provided a reasonable approximation to the simulation model. The Bayesian method encountered some numerical problems and could only run on 78 out of 100 simulated data sets. More serious numerical problems were experienced by the Bayesian method in two setups presented in supplemental materials.

Table 1 shows the results of comparing our reduced rank method with the Bayesian method. For setups 1–4, our reduced rank method consistently did a much better job in estimating the mean functions and in predicting the random effects. The inferior performance of the Bayesian method might be due to its use of oversimplified covariance matrices and/or of a separable covariance structure, it might also be that the Baysian model is overfitting the data. For data generated from the Bayesian model, our reduced rank method is comparable

Table 1: Comparison of two methods based on 100 simulation runs. Mean (SE) of the integrated absolute errors of estimating the mean functions and predicting the unit and sub-unit level random effects. Numbers shown are the actual numbers multiplied by 10. “Reduced rank” refers to our method; “Bayesian” refers to the Bayesian method of Baladandayuthapani, et al. (2008).

Setup	Method	Mean	Unit	Sub-Unit
1	Reduced rank	1.332 (0.061)	1.361 (0.072)	0.861 (0.051)
	Bayesian	1.945 (0.041)	3.512 (0.046)	1.637 (0.030)
2	Reduced rank	1.633 (0.057)	2.152 (0.097)	1.628 (0.100)
	Bayesian	2.161 (0.046)	3.797 (0.052)	2.051 (0.023)
3	Reduced rank	1.151 (0.040)	1.487 (0.059)	1.117 (0.062)
	Bayesian	1.886 (0.023)	2.681 (0.036)	1.502 (0.016)
4	Reduced rank	1.492 (0.056)	1.517 (0.055)	0.543 (0.014)
	Bayesian	2.049 (0.034)	3.366 (0.046)	1.055 (0.011)
5	Reduced rank	4.170 (0.128)	5.571 (0.094)	4.063 (0.039)
	Bayesian	4.086 (0.146)	5.518 (0.108)	4.104 (0.046)

to the Bayesian method. We do not compare the parameter estimates here since parameters from the two models have different interpretations. Some results of parameter estimation for our method are presented in the online supplementary materials.

6 Application to Colon Carcinogenesis Data

In this section we apply our model to the rodent experiment introduced in Section 1. We will focus on studying the cell expression level of the p27 protein, a cyclin-dependent kinase inhibitor in normal and neoplastic cells. These data were analyzed previously by Baladandayuthapani, et al. (2008) using their Bayesian hierarchical model. The data were also used in Li et al. (2008) to illustrate a method for nonparametric estimation of correlation functions.

In the experiment, 48 rats were randomized to 4 diet groups: corn oil with butyrate (CO+B), corn oil without butyrate (CO–B), fish oil with butyrate (FO+B) and fish oil without butyrate (FO–B). After being fed these diets for 2 weeks, each rat was exposed to the carcinogen azoxymethane (AOM) and euthanized at one of four randomly chosen time points: 0, 12 hours, 24 hours and 48 hours. From each rat, 20-30 crypts were selected from its colon and the physical locations of the crypts were measured in microns. Within rat crypt distance ranges from 5 microns to about $14,000$ microns. There are about 20-40 cells on each crypt and their relative cell positions t were calculated such that the bottom of each crypt has $t = 0$ and the top has $t = 1$, with positions in between coded proportionally. The expression level of p27 was determined for each cell. For details on the measuring process, see, Hong, et al. (1997). Our goal is to study the phenomenon of crypt signaling, that the level of p27 in the cells in a given crypt is affected by neighboring crypts and the effect is a function of crypt distances.

As an illustration of our methods, we use data from rats euthanized at 24 hours only. We have diet groups as the treatment groups, and each rat is a unit with its crypts as subunits. In the notation of our model (8) given in Section 2.4, denote Y for the logarithm transformed p27 levels which are standardized to have mean 0 and variance 1, denote X for the crypt location and t for relative cell position. Based on the biological nature of our data and the short length of the assayed colon slice, it is reasonable to assume that cells on nearby crypts would have similar p27 responses and the correlation coefficient function between the crypt level PC scores is stationary in the sense that it only depends on the relative distance between crypts. As mentioned in Section 2.3, we used the Matérn family of isotropic autocorrelation function in our model. For simplicity, the same parameters ϕ and ν are used for all diet groups.

For modeling the diet, rat and crypt level functions of relative cell positions, we used quadratic B-splines with 5 equally spaced interior knots. Three-fold crossvalidation, with rats from each treatment group evenly distributed to each fold, was used both to select the penalty parameters and to determine the number of PCs (Sections 4.1 and 4.2). Crossvalidation scores of some feasible number of PCs are given in Table 2. We decided to use one rat level and two crypt level PC functions because, the crossvalidation score increases significantly if fewer PCs are used and it does not change much if more PCs are used. Various diagnostic plots (Section 4.3, not shown) also confirmed that our choice is reasonable.

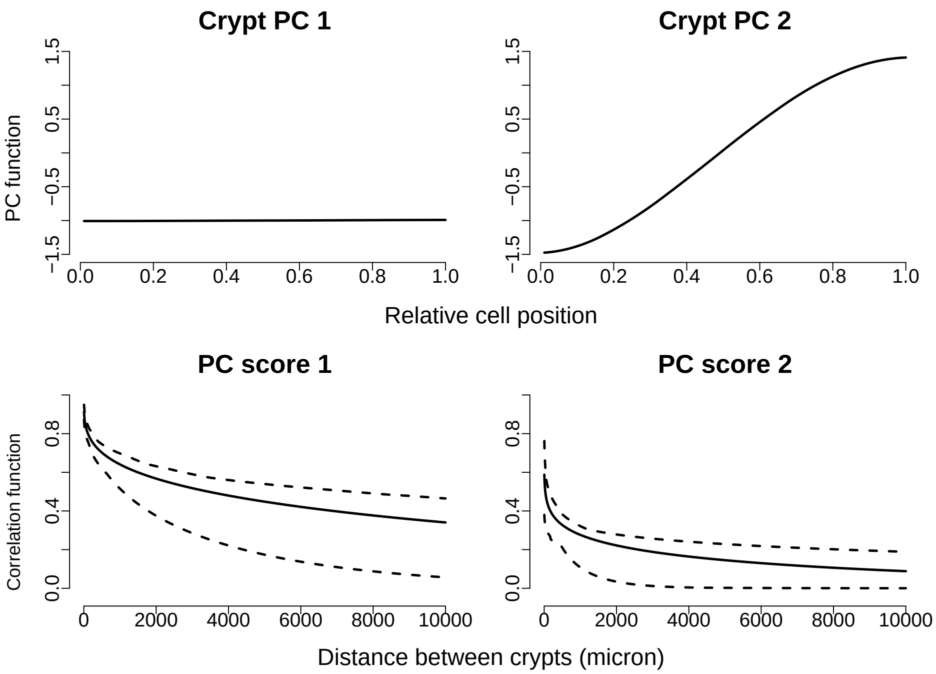
Next we report our analysis results on crypt signaling and diet effects. The parametric bootstrap method was applied with 500 bootstrap samples to estimate the distribution of parameter estimates. When applying the bootstrap, the penalty parameters are reestimated for each bootstrap sample while the number of PC's is fixed. We fixed the number of PC's to ensure that model parameters are the same for all bootstrap samples. This may lead to underestimation of the variation, however the effect will only be slight because the method of choosing the number of PCs is almost always correct in our simulations.

Table 2: The crossvalidation scores for some candidate models.

	# of crypt level PCs			
# of rat level PCs	1	2	3	4
1	−608.00	−702.84	−721.70	−725.65
2	170.46	−703.45	−722.52	−728.32

The estimated rat level PC is almost a constant (not shown). The estimated crypt level PC functions are shown on the top panels of Figure 1. For the spatial correlation parameters of the Matérn family, the range parameter ϕ is estimated to be 29.39 and 8.58 with 95% CIs $(4.73, 147.39)$ and $(0.74, 76.08)$ for the two PC scores respectively; the estimated Matérn order ν is 0.13 and 0.05 with 95% CIs $(0.10, 0.20)$ and $(0.02, 0.12)$ respectively. The bottom panels of Figure 1 show the estimated correlation function of crypt level PC scores along with corresponding 95% CIs for crypt distances from 5 to $10,000$ microns. The estimated

Figure 1: Colon carcinogenesis data. Top panels: estimated crypt level PC functions. Bottom panels: estimated correlation functions of PC scores over crypt distance and corresponding 90% pointwise confidence intervals.

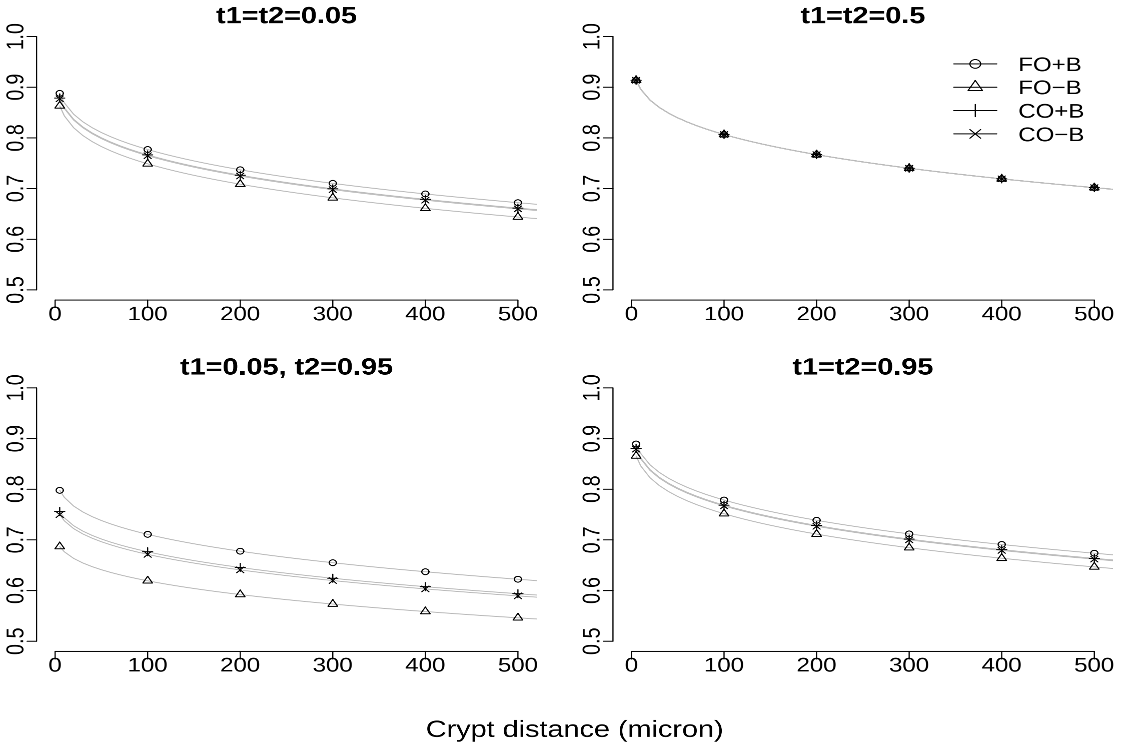


correlation functions strongly suggest the existence of crypt signaling: cells in crypts that are located close together tend to have similar p27 expression levels; the similarity decreases as the distance between crypts increases. The correlation function of the first PC scores decay slower than that of the second PC scores.

Baladandayuthapani, et al. (2008) used a separable space-time covariance structure to analyze the same data. Our fitted model suggests a nonseparable covariance structure, because it uses two crypt level PC functions with different correlation parameters and in particular, the 95% CI for $\nu_1 - \nu_2$ is $(0.02, 0.15)$. More specifically, following (7) in Section 2.3, for two crypts c and c' with physical distance $d_{cc'}$ from diet group a, rat b, the covariance of the corresponding crypt level functions $\eta_{abc}(t)$ and $\eta_{abc'}(t')$ is

$$\mathrm{cov}(\eta_{abc}(t), \eta_{abc'}(t')) = \sigma^2_{\beta,a1}\rho(d_{cc'}; \phi_1, \nu_1) g_1(t) g_1(t') + \sigma^2_{\beta,a2}\rho(d_{cc'}; \phi_2, \nu_2) g_2(t) g_2(t'),$$

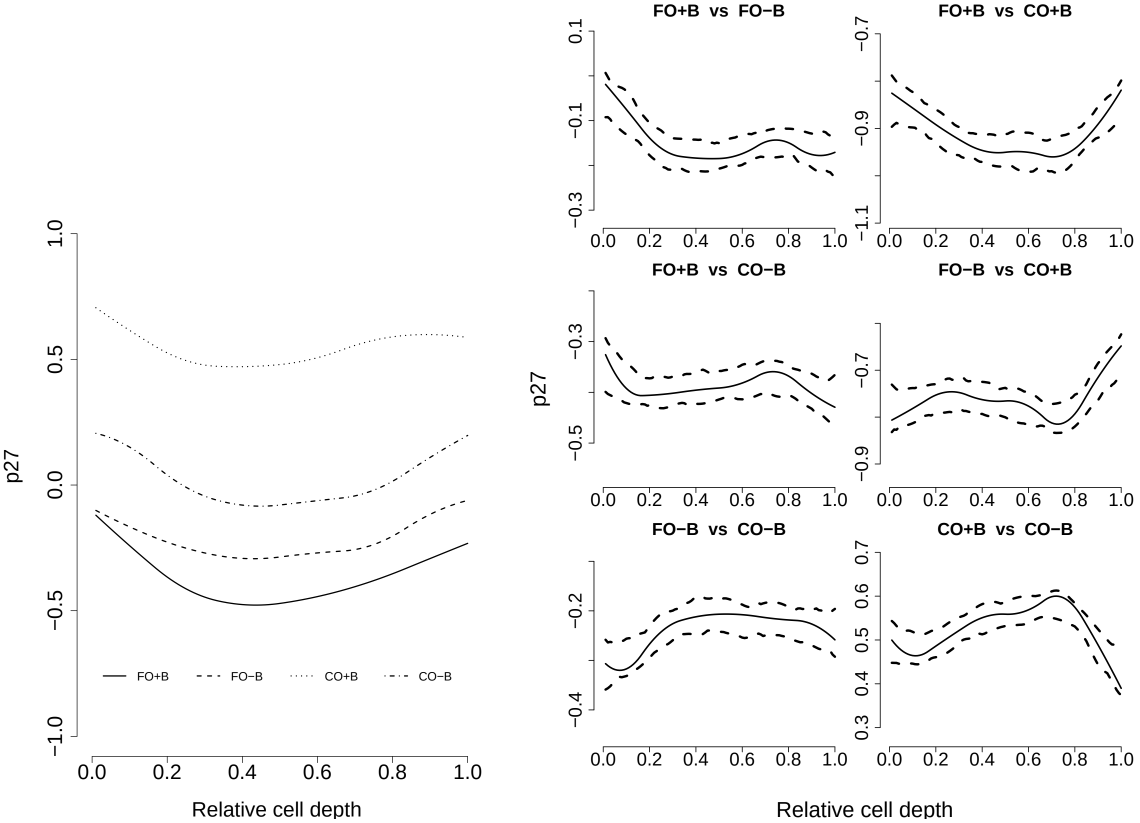
Figure 2: Estimated correlation functions over crypt distance for selected combinations of relative cell depth (denoted as t_1 and t_2).



which is nonseparable since $\phi_1 \neq \phi_2$ and $\nu_1 \neq \nu_2$. Figure 2 plots the fitted correlation functions as a function of physical distance $d_{cc'}$ for selected pairs of t and t'. In addition to suggesting crypt signaling, Figure 2 also suggests that the covariance structure is treatment dependent and that the correlation at locations t and t' from two different crypts is higher when t and t' are closer.

Figure 3(a) shows the mean diet level functions for the four diet groups. There seem to be some diet differences especially between the CO+B diet versus the rest of the diets. To investigate this further, Figure 3(b) shows all pairwise differences between the mean diet level functions of two diet groups, together with corresponding 95% pointwise confidence intervals. All diet groups show significant differences in mean (logarithm transformed) p27 level. Using their Bayesian hierarchical model and pointwise Bayesian credible intervals, Baladandayuthapani, et al. (2008) also found significant difference between the CO+B diet and the other three diets, but not among the latter three. Residual plots available in the supplemental materials clearly indicate that our reduced rank model provided better fit to the data than the Bayesian model.

Figure 3: Estimated mean log p27 level over relative cell depth for the four diet groups. CO is Corn Oil, FO is Fish Oil and with or without (±) Butyrate supplement. The left figure gives the fitted group mean functions. The right figure compares the mean functions with 95% pointwise confidence intervals.



7 Conclusion

In this paper we have proposed mixed effects models for spatially correlated hierarchical functional data. Dimension reduction by principal components plays an important role both in modeling the covariance structure of random functions and in modeling spatial correlation.

Existing work on modeling covariance structure applied diagonal correlation matrices on coefficients in certain fixed basis expansions; Baladandayuthapani, et al. (2008) used truncated power (spline) basis expansions. Our approach also applies diagonal correlation in basis expansion, but the major difference with existing work is that our basis system is determined by data. Note that relaxation of the diagonal restriction to the correlation matrices in the fixed basis approach introduces too many parameters and thus poses substantial statistical and computational challenges. Our use of principal components basis functions provides a flexible, yet feasible approach of modeling the covariance structure. Moreover, instead of modeling the correlation between sub-unit level random effects directly, we model spatial correlation through principal component scores and therefore relax the space-time separability assumption employed by Baladandayuthapani, et al. (2008).

8 Supplemental Materials

Supplemental materials contain the description of the computational methods for implementing the proposed methodology and additional simulation results. (pdf file)

References

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 200–23.

Baladandayuthapani, V., Mallick, B., Turner, N., Hong, M., Chapkin, R., Lupton, J. and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64**, 64–73.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion) *Journal of the American Statistical Association* **93**, 961–76.

de Boor, C. (2001). *A Practical Guide to Splines*, Revised Edition. Springer, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S. and Punjabi N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.*, **3**, 458–488.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **89**, 89–121.

Grambsch, P. M., Randall, B. L., Bostick, R. M., Potter, J. D. and Louis, T. A. (1995). Modeling the labeling index distribution: an application of functional data analysis. *Journal of the American Statistical Association* **90**, 813–21.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics* **35**, 403–10.

Handcock M. S. and Wallis J. R. (1993). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association* **89**, 368–78.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* **21**, 309–10.

Hong, M. Y., Chang, W. L., Chapkin, R. S., and Lupton, J. R. (1997). Relationship among colonocyte proliferation, differentiation, and apoptosis as a function of diet and carcinogen. *Nutrition and Cancer* **28**, 20–29.

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

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

Li, Y., Wang, N., Hong, M. Y., Turner N. D., Lupton, J. R., Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *Ann Statist.* **35**, 1608–1643.

Liang, H., Wu, H. and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics* **4**, 297–312.

Morris J. S. and Carroll R. J. (2006). Wavelet-Based Functional Mixed Models. *Journal of the Royal Statistical Society, Series B* **68**, 179–99.

Morris J. S., Vannucci M., Brown P. J. and Carroll R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association* **98**, 573–83.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–13.

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

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53**, 233–43.

Rice, J. A. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–59.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–57.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). Semiparametric regression. Cambridge University Press.

Schabenberger, O. and Gotway, C. A. (2005). Statistical Methods for Spatial Data Analysis. Boca Raton, Chapman & Hall/CRC.

Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996). An analysis of pediatric CD4+ counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* **45**, 151–63.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics* **24**, 1–24.

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **93**, 1403–1418.

Stein, M. L. (1999). *Interpolation of Spatial Data.* Springer: New York.

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B* **60**, 159–74.

Wang, Y. and Wahba, G. (1998) Discussion of "Smoothing spline models for the analysis of nested and crossed samples of curves" by Brumback and Rice. *Journal of the American Statistical Association* **93**, 976–80.

Wu, C.F.J. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics* **11**, 95–103.

Wu, H. and Liang, H. (2004) Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics* **31**, 3–20.

Wu, H. and Zhang, J. T. (2002) Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association* **97**, 883–97.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–90.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-903.

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