ABSTRACT

TIDEMANN-MILLER, BETH A. Statistical Modeling of Multivariate Functional Data that Exhibit Complex Correlation Structures. (Under the direction of Brian Reich and Ana-Maria Staicu.)

Due to the large size of modern data sets, there is an ever-increasing need for computationally efficient inferential methods designed for realistic models of large observed functional data sets. The first part of this dissertation introduces an innovative modeling framework for the analysis of multivariate functional data, where each individual functional component exhibits multilevel and spatial structures. The proposed methodology uses a functional principal components based approach for multivariate functional data, which has important advantages in the dimensionality reduction of the data and brings considerable computational savings. Moreover, our approach quantifies the spatial auto- and cross-correlation between units at the lowest level of the hierarchy. The proposed procedure is illustrated through simulation studies and data from a colon carcinogenesis experimental study.

In the second part of the dissertation, we propose a Bayesian modeling framework for jointly analyzing multiple functional responses of different types (e.g. binary and continuous data). Our approach is based on a multivariate latent Gaussian process and models the dependence among the functional responses through the dependence of the latent process. Our framework easily accommodates additional covariates. We offer a way to estimate the multivariate latent covariance, allowing for implementation of multivariate functional principal components analysis to specify basis expansions and simplify computation. We demonstrate our method through both simulation studies and an application to real data from a periodontal study.

Statistical Modeling of Multivariate Functional Data that Exhibit Complex Correlation Structures

by
Beth A. Tidemann-Miller

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2014

APPROVED BY:

Brian Reich Co-chair of Advisory Committee	Ana-Maria Staicu Co-chair of Advisory Committee
Marie Davidian	Arnab Maity

Sukanta Basu

DEDICATION

For Jeff.

BIOGRAPHY

Beth A. Tidemann-Miller grew up in Sidney, Nebraska. She graduated from Sidney High School in May of 2003 and went on to pursue undergraduate studies at the University of Nebraska-Lincoln. She graduated with high distinction in May of 2008, earning a Bachelor of Arts with majors in Spanish, International Studies, and Latin American Studies, and a Bachelor of Science with a major in Mathematics & Statistics. She moved to Raleigh, North Carolina in August of 2008 to begin graduate studies in Statistics at North Carolina State University. She earned her Master of Statistics degree in May of 2010 and her PhD in December of 2014.

ACKNOWLEDGEMENTS

It is with the deepest gratitude that I acknowledge several key individuals who have accompanied me on this educational journey. Primarily, I wish to thank my co-advisors, Drs. Ana-Maria Staicu and Brian Reich, for their dedication to helping me see this work through to completion. They have molded me into a more thoughtful, critical, insightful, and technically sound statistician, and I will benefit from their instruction for years to come. To my committee members, Drs. Arnab Maity, Sukanta Basu, and Marie Davidian, I extend my appreciation for the valuable input they have provided for the research contained in this dissertation. I would like to acknowledge Dr. Davidian in particular for her support throughout my graduate school career. The chance to work on her training grant was the deciding factor in my choice of graduate programs, and she has shown continued commitment to my professional development through financial assistance and mentorship. I also would like to express my gratefulness to Dr. Montserrat Fuentes for her many kindnesses and sound professional guidance. To all of my professors at NCSU, I extend many thanks.

The members of the astounding departmental staff at NCSU have been my lifelines. There is no adequate way in which to express exactly how much dear, sweet, Alison McCoy has meant to me throughout this process. Her encouragement, hugs, smiles, and love always carried me through the rough patches, and I am eternally grateful for her friendship. Also, I am convinced my simulations would still be running if it weren't for the help of Terry Byron and Chris Waddell who are always willing to share their invaluable IT knowledge and accommodate my many and varied requests for assistance.

I owe so much of my statistical knowledge to my fellow graduate students who were always eager to help me understand difficult concepts, complete homework assignments, study for exams, and learn tips and tricks for computer programming. Moreover, I received endless encouragement and support from friends who earned their PhDs and helped me believe that I could, too. I formed numerous lifelong relationships at NCSU that I will always cherish. Thank you all for being such wonderful colleagues and friends. I must also thank my colleagues at Duke and the DCRI where I gained invaluable experience that helped prepare me for my chosen career path in Biostatistics.

My success in graduate school is due in large part to the solid background in Mathematics and Statistics that I received at the University of Nebraska-Lincoln. For their gentle nudges that led me to be a Math major and to explore Statistics, I thank Drs. Mohammad Rammaha and Gordon Woodward. For the high quality statistical instruction that prepared me well for graduate school, I thank Drs. Erin Blankenship, Walter Stroup, Tisha Hooks, Jacqueline Wroughton, and Chad Brassil. Drs. Blankenship and Brassil served as my undergraduate co-advisors, and it was under their instruction that I successfully completed my undergraduate honors thesis.

To my family and friends: you have continually given me your love, support, encouragement, patience, understanding, and kindness throughout this journey. To Mom and Dad, thank you for teaching me to be inquisitive, respectful, and perseverant, to do my best in all things, and to make school a priority. Darcia and Jeremy, as your little sister I have always looked up to you and had the privilege of following your example. Thank you for demonstrating such dedication to academics, motivating me to challenge myself, and believing in my potential from day one.

Finally, to my husband, Jeff: this PhD is *ours*. Your devotion, selflessness, humor, patience and unceasing demonstrations of love were instrumental in bringing us to this point. You make me better in every way, and I am blessed to share my life with you. From the bottom of my heart: Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 What is functional data analysis?

The area of Statistics known as functional data analysis (FDA) has undergone many methodological developments in the past two decades and is still experiencing intense growth. Many types of electronic devices can now observe functions at very fine increments and have become ubiquitous in a wide variety of disciplines. Such a rapid advancement of technology in recent years has elevated the level of interest in FDA as researchers seek innovative ways of handling large and increasingly complex data that come in the form of functions.

Unlike other statistical frameworks (e.g. longitudinal, multivariate, or time-series), FDA views an observed function as a single (functional) datum. For example, consider an electrocardiogram (ECG) machine that monitors the electrical activity of a patient's heart every fraction of a second for one hour. For the purposes of FDA, the entire function of ECG observations for the hour comprise one functional datum, contrary to more traditional approaches where the value of the ECG at a single time point would be an individual datum.

Although earlier research in FDA had appeared in the literature, the comprehensive monographs *Functional Data Analysis* (Ramsay and Silverman 1997) and *Applied*

functional data analysis: methods and case studies (Ramsay and Silverman 2002) helped to unify FDA methods to form a statistical sub-discipline. FDA was justing starting to gain momentum within the statistical community around the time the second edition of *Functional Data Analysis* was published (2005) and also the well regarded FDA monograph of Ferraty and Vieu (2006) appeared. Also around this time, special issues of several journals were devoted to exploring FDA in more depth, such as bridging the gap between longitudinal data analysis and FDA (see Davidian M. and Wang (2004); Rice (2004)), modeling functional data (see Valderrama (2007)), and other topics (see Gonzalez-Manteiga and Vieu (2007)).

Although FDA is a relatively new area of statistical analysis, methodology has developed to such an extent that topics of major interest such as regression, classification, and prediction, to name a few, have existing analogues within FDA. For example, functional linear regression incorporates a functional predictor to model either a scalar response (Cardot et al. 2007; Crambes et al. 2009; Ferraty and Vieu 2002; James 2002) or a functional response (James et al. 2009; Liang and Zeger 1986; Wu et al. 1998; Yao et al. 2005b). Importantly, the goals of FDA align with those of any other area of Statistics, for instance, describing central tendency and variation, and forming parsimonious models. Unique to FDA is the ability to use derivatives of the curves to inform an analysis. Sometimes, trends in the derivative itself are of interest. A good example of this comes from a growth curve study described in Chapter 1 of Ramsay and Silverman (2005) in which profiles of girls' height were collected from childhood to adolescence. For this study, one might be interested the speed (first derivative) and acceleration (second derivative) of the height profiles.

Functions on the time-domain, such as the ECG and growth curve examples above, naturally fit within the FDA framework. However, FDA readily applies to functions on any continuum. For instance, the data application in Chapter 2 is from a colon carcinogenesis experiment in which the data are functions of cell depth within fingerlike structures on the inner lining of the colon. In the periodontal data application of Chapter 3, measures of patients' oral health are observed at each of the patients' teeth, and we treat this as functional data due to the natural ordering of teeth in the mouth. Moreover, functional data applies to functions of two or more dimensions, for example, data from three-dimensional functional magnetic resonance imaging (fMRI) used to measure activity in the brain.

In FDA a function is considered as having an infinite-dimensional domain, although in reality only a finite number of realizations of the function can be collected. Many methods in FDA assume that the functional data are observed densely over the domain, meaning values of the curve are evaluated frequently. When data is observed sparsely, methods for densely observed functional data may no longer be applicable. For simplicity we mainly focus our attention to the densely observed case and only briefly touch upon methods for sparse data.

The goal of this introduction is to provide a primer for the content that follows in subsequent chapters. It is meant to introduce the reader to the lens through which FDA approaches data analysis and to familiarize the reader with concepts that are both common within FDA and also recurrent throughout the methodologies we propose later on. The information presented in this introduction is largely compiled from Ramsay and Silverman (2002), Ramsay and Silverman (2005), and the collection of works found in Ferraty and Romain (2011), and we suggest that the reader consult these or other texts for a more comprehensive review of FDA.

1.2 Smoothing

The concept of smoothness is an integral part of FDA since it is assumed that functional data emanate from a smooth underlying process. Saying a function is smooth usually means that one or more derivatives exist (Ramsay and Silverman (2005), Ch. 3). More simply, it means the function is devoid of abrupt changes and smoothly transitions from one value to the next.

To further understand the meaning of smooth functions, consider the functional datum $Y(t)$ collected at corresponding evaluation points $t = t_1, \ldots, t_{\mathrm{L}}$ within some interval $\mathcal{T}$. We can view the datum as L realizations (observed with or without error) of a smooth function $s(t)$, where the underlying function $s(t)$ is also defined for $t \in \mathcal{T}$. When the datum is observed without error, we can use the model $Y(t) = s(t)$, and recovering the underlying function $s(t)$ is called interpolation. More commonly, the datum suffers from observational error, making the observed curve rough or wiggly, and an appropriate model is $Y(t) = s(t) + \epsilon(t)$ where $\epsilon(t)$ is a random error process. In this case, recovering $s(t)$ is known as smoothing.

Smoothing can remove the observational error, and smoothed data can also be used in place of the raw data (Zhang and Chen 2007). Smoothing also enables the

researcher to evaluate $s(t)$ at any $t \in \mathcal{T}$, not just at the evaluation points t_ℓ, $\ell = 1, \ldots, \mathrm{L}$, where $Y(t)$ is observed. How dense the evaluation points t_ℓ are within the domain $\mathcal{T}$ along with the curvature of the underlying function determine what is known as the resolution of the curve.

If the resolution of the curve is low, meaning the observations are too sparsely observed to adequately capture the features of the underlying process, then smoothing individual curves may not be appropriate. Given a sample of curves $Y_i(t_{i\ell})$, $i = 1, \ldots, \mathrm{N}$, where each curve is observed at L_i evaluation points $t_{i\ell}$, one can smooth each curve by borrowing information across curves. Methods for this sparse case include mixed effects modeling and local smoothing, among others; we direct the interested reader to James (2011) for an excellent review of methods for sparse data related to principal components, clustering, classification, and regression.

Smoothing is primarily achieved using 1) global smoothing methods such as basis expansions with or without roughness penalties, and 2) local smoothing methods such as kernel smoothers. The basis expansion method uses a linear combination of basis functions to represent a smooth function. This method is described in Section 1.2.1 with particular attention given to B-splines (de Boor 1978) which Ullah and Finch (2013) found to be the most popular smoothing method implemented within the FDA literature. Basis functions can be used in conjunction with a roughness penalty approach finds the function that fits the data well but also does not exhibit too much local variation (Section 1.2.2). Kernel smoothing is described in Section 1.2.3.

1.2.1 Basis Expansions

Expressing a smooth function as an expansion of basis functions is commonplace within FDA and appears frequently in the methodologies presented in Chapters 2 and 3. The basis expansion approach has several uses such as smoothing the raw data, representing the data at any location in the domain (regardless of whether a function value was measured at that point), and unifying a sample of curves that were observed for different evaluation points in the domain.

A basis system is a set of basis functions in which the functions are independent of one another and span a particular function space. For example, the polynomial basis $\{t^{k-1} : k \geq 1\}$ spans the space of polynomial functions. This means that any

polynomial function $f(t)$ of degree $\mathrm{K}-1$ can be represented by a linear combination of these basis functions, that is, $f(t) = \sum_{k=1}^{\mathrm{K}} \beta_k t^{(\mathrm{K}-1)}$. In general, let $f(t)$ be a smooth function and let $\{\phi_k(t) : k \geq 1\}$ be a set of basis functions (considered known) for $t \in \mathcal{T}$. Using a basis expansion, the function $f(t)$ is approximated by

$$f(t) \approx \sum_{k=1}^{\mathrm{K}} \beta_k \phi_k(t), \tag{1.1}$$

where the parameter K is the number of basis functions included in the expansion and the parameters β_k are unknown but fixed coefficients that can be estimated through regression methods such as least squares.

Given that an appropriate basis system is chosen, the quality of the approximation in (1.1) depends on the value of K. For observed function values $f(t_\ell)$, $\ell = 1, \ldots, \mathrm{L}$, it is possible for the basis expansion to interpolate the values $f(t_\ell)$ exactly if the number of basis functions is chosen to be equal to the number of evaluation points, that is, by setting $\mathrm{K} = \mathrm{L}$ (Ramsay and Silverman 2005). Even though specifying a large number of basis functions might offer an excellent fit to the data, overfitting is a pitfall one should avoid. In practice, the number of basis functions should be much smaller than the number of evaluation points but still large enough to provide a good fit. Indeed, the goal of using a basis function expansion is to specify a basis system that fits the data well for a relatively small value of K.

Since the set of basis functions is considered known in advance (pre-determined), it is important that the choice of basis system coincides with the features of the data and the goals of the analysis. For example, a Fourier basis is a good choice for data of a cyclic nature, and a B-spline basis is a common choice for non-cyclic data. Other types of bases include wavelets, exponential bases, and power bases. Since we use a B-spline basis system in Chapter 3, we describe it below. We refer the reader to texts such as Wood (2006) and Ramsay and Silverman (2005) for a thorough review of this and many other types of basis systems.

B-splines offer a very flexible way of representing a non-periodic function and are appealing due to their fast computation. In general, splines (Wahba 1990) involve smoothly joining segments of polynomial functions together at so-called knots, or points that break the domain into smaller intervals. The number and location of the knots are pre-specified. There is always a knot specified at the beginning and the end

of an interval, but the placement of interior knots depends on the application. For data observed densely, equally spaced knots may be appropriate, but one can also use quantiles of the evaluation points t_ℓ, especially for functional data that have not been observed on an equally spaced grid (Ruppert et al. 2003). Also, it can be useful to place knots more closely where the curvature of the function is more complex.

In addition to knots, one must specify the order of a B-spline basis, which defines the degree of the polynomial segments. The order of the B-spline is $m + 1$ for a polynomial of degree m, so that B-splines of order 2 join straight lines, of order 3 join quadratic functions, of order 4 join cubic functions, and so on. At each interior knot, segments are joined in such a way that not only the value of the two segments must be equal, but their derivatives (up to order $m - 2$) must also match. Cubic B-splines (order 4) are common choices as they have continuous second derivatives.

The number of basis functions needed for a B-spline fit is calculated as the order plus the number of interior knots, that is, $\mathrm{K} = (m + 1) + (c - 2) = m + c - 1$, where c is the total number of knots. Generally for a basis expansion, increasing the number of basis functions K will lead to a better fit, but this is not always the case in B-splines due to the interplay between the knots and the order on determining the number of basis functions. It is typically better to increase K by specifying more knots instead of increasing the order $m + 1$ of the B-splines (Marx and Eilers 1998; Wood 2000).

B-splines have very attractive computational characteristics. A B-spline of order $m + 1$ has basis functions that are only positive over at most $m + 1$ adjacent intervals, making them localized. This feature gives the B-spline system the advantage of behaving like an orthogonal basis system so that including a very large number of functions K only increases computation time linearly with K (Ramsay and Silverman (2005), Ch. 3).

1.2.2 Roughness Penalties

As mentioned previously, roughness penalties are used to control how well a function fits the data while forcing the fitted function to retain a smooth form. This is the classic bias-variance trade off, where a function that fits the data too well has small bias but a lot of local variation (it is rough or wiggly), and a function that is too smooth fits the data poorly but has small variation. The ideal fit balances the bias and the variance, which is the goal of smoothing with a roughness penalty.

In this approach, the fit is achieved through optimizing some criterion that measures closeness to the observed data (e.g. sum of squared error) while imposing a penalty to control smoothness. The second derivative $f''(t)$ of a function $f(t)$ gives an indication of the curvature of $f(t)$, where both positive or negative values of the second derivatives indicate some type of curvature, and a second derivative close to zero indicates small curvature (recall that $f(t)$ is straight line if $f''(t) = 0$). Thus, a typical penalty involves penalizing the square of the second derivative of the function, expressed symbolically as $\theta \int \{f''(t)\}^2 dt$. Then the function $f(t)$ is that which minimizes, for instance, the penalized least squares criterion $\int \{Y(t) - f(t)\}^2 dt + \theta \int \{f''(t)\}^2 dt$. The parameter $\theta \geq 0$ is known as the smoothing parameter (more commonly denoted as λ, but we change the notation to avoid confusion with the eigenvalues in Section 1.3). As θ increases, $f(t)$ becomes smoother, and as θ decreases to zero, $f(t)$ becomes rougher, corresponding to the unpenalized case. The smoothing parameter θ can be estimated using cross-validation among other techniques, and more information can be found in Chapter 4.5 of Wood (2006).

The roughness penalty approach can be used in conjunction with many other methods, for example, with the basis expansion approach of Section 1.1. So-called penalized splines or smoothing splines (Eilers and Marx 1996) use a spline expansion for the function with the additional smoothness constraint imposed by the roughness penalty. Let $f(t) \approx \sum_{k=1}^{K} \beta_k \phi_k(t)$ represent a B-spline basis expansion. If the B-splines (without the penalty) are fit by least squares, the coefficients β_k are those that minimize $\int \{Y(t) - f(t)\}^2 dt$. In penalized B-spline smoothing, β_k are those that minimize $\int \{Y(t) - f(t)\}^2 dt + \theta \int \{f''(t)\}^2 dt$.

1.2.3 Kernel Smoothing

In kernel smoothing (Fan and Gijbels 1996; Wand and Jones 1996) the function $f(t_0)$ at a target point t_0 is estimated by using a weighted combination of observations $Y(t_\ell)$ with evaluation points t_ℓ that are close to t_0. As opposed to the global smoothing approaches in Sections 1.1 and 1.2.2 where a smooth function $f(t)$ is approximated simultaneously for all values t in the domain, kernel smoothing is localized in that the fit is done separately for each target point t_0 and depends only on information from values in a neighborhood of that point. The influence of a point t_ℓ on the fit at t_0 is determined by the kernel function $K_h(t_0, t_\ell)$ which weights $Y(t_\ell)$

based on the proximity of t_ℓ to the target point t_0. The parameter h is known as the bandwidth and determines the width of the neighborhood around t_0.

Nearest neighbor estimation is one of the most basic forms of kernel smoothing where the estimate $\widehat{f}(t_0)$ is simply the average of all the observed values within the neighborhood of t_0. (We employ the concept of nearest neighbors in a more complicated setting in Chapter 2.3.1.) Nearest neighbor averages give equal weight to all values within the neighborhood regardless of their relative proximity to the target point t_0. The Nadaraya-Watson kernel-average (Nadaraya 1964; Watson 1964) improves upon this by using a weighted average and is given by

$$\widehat{f}(t_0) = \frac{\sum_{\ell=1}^{n} K_h(t_0, t_\ell) Y(t_\ell)}{\sum_{\ell=1}^{n} K_h(t_0, t_\ell)}.$$

It produces a continuous estimate for a continuous kernel $K_h(t_0, t_\ell)$ (see Sarda and Vieu (2012)). For instance, by choosing the Epanechnikov kernel $K_h(t_0, t_\ell) = D(|t_\ell - t_0|/h)$ where $D(x) = 3/4(1 - x^2)$ if $|x| \leq 1$ and $D(x) = 0$ otherwise, assigns weights that increase smoothly as t_ℓ approaches t_0, or equivalently, weights that die off as t_ℓ gets further from t_0 (see Hastie et al. (2009)).

In kernel regression, one must specify the bandwidth h which controls the degree of smoothing, akin to the role of the smoothing parameter λ discussed in Section 1.2.2. With larger h, more values enter the neighborhood and the fit becomes smoother (less variability) but more biased. With smaller h, the fit is less smooth (higher variability) but less biased. The value of h can be chosen through cross validation, for example.

One downside to kernel smoothing is that the boundary estimates tend to exhibit more bias than interior estimates since there are fewer available points in the neighborhood around a t_0 close to the boundary. Local linear (more generally, polynomial) regression is an alternative to kernel smoothing that possesses better properties near the boundary and involves regressing a line (or polynomial) at each t_0. For more information on this method, we refer the reader to Chapter 6 of Hastie et al. (2009), for example.

1.3 Functional Principal Components Analysis

Functional principal components analysis (FPCA) is the functional equivalent of principal components analysis (PCA) from the usual multivariate framework. One of the main goals of both PCA and FPCA is to reduce the dimension of the data by finding the directions of the observation space that explain the majority of the variation within the data. FPCA has great utility within FDA; see, for instance, Besse and Ramsay (1986); Boente and Fraiman (2000); James et al. (2000); Ramsay and Dalzell (1991); Rice and Silverman (1991); Yao et al. (2005a), among many others. For an excellent review of FPCA literature, see Shang (2014). FPCA is central to the methodologies we present in later chapters where we offer extensions of FPCA for multivariate functional responses, a topic that to date has appeared only scarcely (Berrendero et al. 2011; Jacques and Preda 2014; Ramsay and Silverman 2005) within the FDA literature.

FPCA can be used to specify what Ramsay and Silverman (2005) call *designer bases* that account for a large amount of variation in the data (Rice and Silverman 1991). We capitalize on this feature in Chapter 3 where we perform FPCA and use the resulting principal component functions as a data-driven basis. Since much of the variation is captured by the basis functions, it alleviates the computational burden within our Bayesian framework of sampling from the conditional posterior distribution of an unstructured covariance matrix by allowing for specification of a much simpler diagonal covariance matrix.

We now offer a brief overview of FPCA for univariate functional data, following closely to the exposition found in Hall (2011). Let $W(t)$ be a square-integrable random function, that is, $\int_{\mathcal{T}} \mathrm{E}\{W(t)\}^2 dt < \infty$, where $t \in \mathcal{T}$ for some compact interval $\mathcal{T}$. Let $K(t,t') = \mathrm{Cov}\{W(t), W(t')\}$ denote the covariance operator of the process. Since the covariance operator is a symmetric and non-negative kernel, we can use Mercer's Theorem to represent $K(t,t')$ in terms of its spectral decomposition,

$$K(t,t') = \sum_{k=1}^{\infty} \lambda_k e_k(t) e_k(t'), \tag{1.2}$$

where $\lambda_k = \mathrm{Var}(\xi_k)$ for $\xi_k = \int_{\mathcal{T}} e_k(t) W(t) dt$ and $\{e_k(t) : k \geq 1\}$ is an orthonormal basis in $\mathcal{T}$, that is, the basis satisfies $\int_{\mathcal{T}} e_j(t) e_k(t) dt = \delta_{jk}$ and $\delta_{jk} = 1$ if $j = k$, $\delta_{jk} = 0$

otherwise. The Karhunen-Loève expansion (Karhunen 1947; Loève 1945) allows us to write the function $W(t)$ in terms of the principal component representation given in 1.2, specifically, $W(t) = \sum_{k=1}^{\infty} \xi_k e_k(t)$.

In the nomenclature of FPCA, $e_k(t)$ is called an eigenfunction or principal component function, λ_k is an eigenvalue, and ξ_k is called an FPC score. Typically, the eigenvalues and corresponding eigenfunctions are considered to be ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, so that the first FPC explains the most variation in the data, followed by the second, and so on. When in this order, the eigenfunctions can be defined sequentially, where the first eigenfunction $e_1(t)$ maximizes the variance of $\xi_k = \int_{\mathcal{T}} e(t)W(t)dt$ subject to the constraint that $\int_{\mathcal{T}} e(t)^2 dt = 1$. For $k > 1$, maximization with the additional constraint $\int_{\mathcal{T}} e_1(t)e_k(t)dt = \ldots = \int_{\mathcal{T}} e_{k-1}(t)e_k(t)dt = 0$ is needed to ensure orthogonality of the eigenfunction basis. Note the if $W(t)$ has not been standardized to have zero-mean, then the first FPC will be the mean function.

In order to estimate the eigenfunctions and eigenvalues, one begins by finding an estimator of the covariance operator $K(t, t')$. Given independent curves $W_i(t)$, $i = 1, \ldots, \mathrm{N}$, one can use the sample covariance

$$\widehat{K}(t, t') = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \{W_i(t) - \overline{W}(t)\}\{W_i(t') - \overline{W}(t')\} \tag{1.3}$$

where $\overline{W}(t) = 1/\mathrm{N} \sum_{i=1}^{\mathrm{N}} W_i(t)$. For a positive definite and symmetric estimator $\widehat{K}$, by Mercer's Theorem there exist eigenvalues $\widehat{\lambda}_k$, $k \geq 1$, and an orthonormal basis of eigenfunctions $\{\widehat{e}_k(t) : k \geq 1\}$ such that

$$\widehat{K}(t, t') = \sum_{k=1}^{\infty} \widehat{\lambda}_k \widehat{e}_k(t) \widehat{e}_k(t'). \tag{1.4}$$

This leads to a very useful result: performing an eigen-analysis of the estimated covariance (or another positive definite and symmetric estimator) leads to estimates of the functional principal components. Of course, we desire the functional principal components to be smooth. This can be achieved by smoothing the data prior to FPCA (Ramsay and Dalzell 1991) or via roughness penalty as in Section 1.2.2 by incorporating penalties into the FPCA constraints above (Pezzulli and Silverman 1993; Silverman 1996). Alternatively, prior to finding the eigen-decomposition, one can use a bivariate smoother like tensor products to smooth the estimated covariance (Di

et al. 2009; Staicu et al. 2010; Yao et al. 1993, 2005a).

Furthermore, it can be shown that the eigenvalues $\widehat{\lambda}_k$ vanish for $k \geq \mathrm{N}+1$, meaning $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_{\mathrm{N}} \geq \widehat{\lambda}_{\mathrm{N}+1} = \widehat{\lambda}_{\mathrm{N}+2} = \ldots = 0$ and the expansion in (1.4) can be truncated at N. If $\widehat{\lambda}_{\mathrm{N}}$ is non-zero, then the eigenfunctions $\{\widehat{e}_k(t) : k = 1, \ldots, \mathrm{N}\}$ are uniquely determined up to a sign (for more details, see Hall (2011)). Typically, the values of $\widehat{\lambda}_k$ diminish quickly as k increases, and (1.4) can be truncated at some $\mathrm{K} \leq \mathrm{N}$ such that cumulatively the first K eigenvalues explain the majority (e.g. 95% or 99%) of the variation. This is very useful if dimension reduction is a primary goal of implementing FPCA.

In both Chapters 2 and 3 we encounter situations where we do not have an independent sample of curves. Hence, obtaining estimators of the covariance is not as straightforward as in (1.3), and the process of finding suitable covariance estimators becomes a key focus of the modeling framework of each chapter.

CHAPTER 2

Modeling Multivariate Spatial Functional Data

2.1 Background

Functional data analysis is a rapidly maturing area of statistical inquiry, particularly due to its ability to handle increasingly large datasets which have become common with the fast pace of technological advancement. In particular, multivariate modeling of functional responses is undergoing intense methodological development. We propose a flexible framework for jointly modeling multiple real-valued functional responses nested within a hierarchy where the functions are observed on a spatial (or temporal) grid and are assumed to exhibit spatial (serial) auto- and cross-correlations. Our methods are applied to data from a colon carcinogenesis experiment, though they are applicable to any data with similar structure.

Our proposed methodology uses functional principal components analysis (FPCA) for multivariate functional data which has important advantages in dimensionality reduction and results in considerable computational savings. FPCA extends PCA from the usual multivariate framework to functional data (see, for example, Ramsay and Silverman (2005), Ch. 8), and recent improvements to FPCA have incorporated more complicated settings. For instance, Di et al. (2009) introduced FPCA for data that are observed in a nested design, i.e. multilevel data. Greven et al. (2010)

presented functional principal components for univariate functions observed longitudinally, and Berrendero et al. (2011) presented FPCA for multivariate functional data.

To our knowledge, this is the first method that allows for a complex spatial correlation structure among curves in the multivariate setting. Baladandayuthapani et al. (2008) presented a functional approach for spatially correlated univariate functional data. Morris and Carroll (2006) developed a wavelets-based approach for functional mixed models and Zhou et al. (2008) developed multivariate methods involving functional principal components analysis, but neither approach can handle the complex spatial correlation structure found in our model.

Staicu et al. (2010) presented methods for univariate multilevel functional data that are spatially correlated, and the methodology we present here shares several similarities. The differences between our method and theirs stem from the difficulties introduced by performing joint modeling of a bivariate response. In particular, the problem of estimating a bivariate spatial covariance is one of continuing research and requires the development of an entirely new and innovative framework compared to the univariate spatial estimation presented in Staicu et al. (2010). One reason for this difficulty is that the entire multivariate spatial covariance matrix must be nonnegative definite.

Two recent approaches to multivariate spatial modeling using the well known and widely used Matérn class of parametric covariance models (Guttorp and Gneiting 2006; Handcock and Stein 1993; Matérn 1986) have appeared in the literature. Gneiting et al. (2010) introduced a valid class of parametric covariances for multivariate spatial random fields, where the component covariance matrices and cross-correlation matrices take the form of a Matérn process. They present constraints on the parameters which ensure a valid covariance structure (nonnegative definiteness) for the bivariate case. Apanasovich et al. (2012) extended this class by relaxing some of the parametric conditions and allowing for more flexible modeling of a multivariate vector with any number of components. The bivariate Matérn model allows for different smoothness parameters that govern the differentiability of the auto- and cross-covariograms. Moreover, the parameters of the Matérn function have meaningful interpretations that can be useful for inference.

Jointly modeling multiple functional responses can offer several advantages over

univariate methods, including improvements in parameter estimates and prediction as well as a better understanding of the relationship between responses. For example, consider the motivating application discussed in Section 2.5 involving a rodent experiment designed to investigate how fish and corn oil diets affect colon carcinogenesis. The responses on which we focus our attention are apoptosis, or programmed cell death, and a cell cycle inhibitor protein called p27. These two responses are closely related in that p27 contributes to the regulation of apoptosis, and novel multivariate methods are essential to gaining a better understanding of the dynamics of their cross-dependence. Analyses of similar colon cancer rodent experiments have appeared in several works, including Morris et al. (2003, 2002, 2001), Morris and Carroll (2006), Baladandayuthapani et al. (2008), and Staicu et al. (2010), to name a few. However, these methods do not offer insight into response dynamics, which is one of our main objectives.

2.2 Model for multivariate functional response

For simplicity of exposition we now describe the modeling framework for bivariate functional data. Consider hierarchical data (groups-subjects-units) where within each subject are several units on which two response curves are observed. Without loss of generality, assume that the units within a subject are aligned on a one-dimensional grid and define the relative spatial location $s_{ij} \in \mathbb{R}$ to be the distance of unit $j = 1, \ldots, \mathrm{M}_i$ from the first unit within subject $i = 1, \ldots, \mathrm{N}$. Let $\boldsymbol{Y}_{ij}(t, s_{ij}) = [Y^1_{ij}(t, s_{ij}), Y^2_{ij}(t, s_{ij})]^{\mathrm{T}}$ be the continuous bivariate response measured at subunit $t \in \mathcal{T}$ within unit j at spatial location s_{ij} within subject i of group $G(i) = 1, \ldots, \mathrm{D}$. Our model is

$$\boldsymbol{Y}_{ij}(t, s_{ij}) = \boldsymbol{\mu}_{G(i)}(t) + \boldsymbol{Z}_i(t) + \boldsymbol{\mathcal{Q}}_i(t, s_{ij}) + \boldsymbol{\epsilon}_{ij}(t) \tag{2.1}$$

where $\boldsymbol{\mu}_{G(i)}(\cdot)$ is the fixed bivariate population-level group mean function and $\boldsymbol{Z}_i(\cdot)$ is the level 1 random bivariate subject-specific deviation from the mean. For a unit with spatial location s_{ij}, the level 2 random bivariate unit-specific deviation from the mean is $\boldsymbol{\mathcal{Q}}_i(t, s_{ij})$, and $\boldsymbol{\epsilon}_{ij}(t)$ is noise. As in Staicu et al. (2010), we further decompose $\boldsymbol{\mathcal{Q}}_i(t, s_{ij})$ into two parts, one part that only depends on the unit spatial location s_{ij}, and another part that only depends on the subunit location t. We write

$\boldsymbol{\mathcal{Q}}_i(t, s_{ij}) = \boldsymbol{W}_{ij}(t) + \boldsymbol{U}_i(s_{ij})$ where $\boldsymbol{W}_{ij}(t)$ is a square-integrable, bivariate random process depending only on the subunit t; $\boldsymbol{U}_i(s_{ij})$ is a bivariate random spatial process depending only on the unit spatial location, s_{ij}, whose correlation structure models the spatial dependence between units within subject i. This leads to our parsimonious model:

$$\boldsymbol{Y}_{ij}(t, s_{ij}) = \boldsymbol{\mu}_{G(i)}(t) + \boldsymbol{Z}_i(t) + \boldsymbol{W}_{ij}(t) + \boldsymbol{U}_i(s_{ij}) + \boldsymbol{\epsilon}_{ij}(t). \tag{2.2}$$

We assume that $\boldsymbol{\mu}_{G(i)}(t)$ are modeled parametrically, and for identifiability we assume that $\boldsymbol{Z}_i(\cdot)$, $\boldsymbol{W}_{ij}(\cdot)$, $\boldsymbol{U}_i(\cdot)$ and $\boldsymbol{\epsilon}_{ij}(\cdot)$ are mean zero, uncorrelated bivariate random processes.

2.2.1 Model Assumptions

Henceforth, we use superscripts $p = 1, 2$ to denote the components of bivariate vectors. For example, the superscript $p = 1$ identifies the first component, $Y_{ij}^1(t, s_{ij})$, of the bivariate response $\boldsymbol{Y}_{ij}(t, s_{ij})$, and $p = 2$ identifies the second component $Y_{ij}^2(t, s_{ij})$. It is assumed that the functions are observed at a dense, balanced design, that is, $t_{ij\ell} = t_\ell$ for $\ell = 1, \ldots, \mathrm{L}$ and that there is an equal number of curves across subjects, that is, $\mathrm{M}_i \equiv \mathrm{M}$.

In model (2.2), $\epsilon_{ij}^p(t)$ are assumed to be white noise processes with $\epsilon_{ij}^p(t) \overset{i.i.d.}{\sim} N(0, \tau_p^2)$ and $\epsilon_{ij}^1(t)$ uncorrelated with $\epsilon_{ij}^2(t)$. The bivariate random processes $\boldsymbol{Z}(t)$ and $\boldsymbol{W}(t)$ have covariance operators $K_{pp'}^Z(t, t') = \mathrm{Cov}\{Z_i^p(t), Z_i^{p'}(t')\}$ and $K_{pp'}^W(t, t') = \mathrm{Cov}\{W_{ij}^p(t), W_{ij}^{p'}(t')\}$, respectively, for $p, p' = 1, 2$ leading to the 2×2 covariance matrices $\boldsymbol{K}^Z(t, t') = \{K_{pp'}^Z(t, t')\}_{p,p' \in \{1,2\}}$ and $\boldsymbol{K}^W(t, t') = \{K_{pp'}^W(t, t')\}_{p,p' \in \{1,2\}}$. The specific form of these covariance matrices is presented in Section 2.2.3.

For modeling the spatial dependence between units, we assume that $\boldsymbol{U}_i(\cdot) = [U_i^1(\cdot), U_i^2(\cdot)]^{\mathrm{T}}$ is a mean-zero and second-order stationary, isotropic random bivariate process that is measured at locations $s_{i1}, \ldots, s_{i\mathrm{M}} \in [0, \mathrm{H}]$ for $s_{ij} \in \mathbb{R}$. The assumption of stationarity is important because it means that the spatial covariance between units within a subject only depend on the distance between the unit locations and not on the unit locations themselves. Notationally, we represent the spatial covariance function as $C_{pp'}(\Delta_{ijj'}) = \mathrm{Cov}\{U_i^p(s_{ij}), U_i^{p'}(s_{ij'})\}$ where $p, p' = 1, 2$ and $\Delta_{ijj'} = |s_{ij} - s_{ij'}|$ is the distance between two units within the same subject. Thus,

the covariance between $\boldsymbol{U}_i(s_{ij})$ and $\boldsymbol{U}_i(s_{ij'})$ is the 2×2 spatial covariance matrix $\boldsymbol{C}(\Delta_{ijj'}) = \{C_{pp'}(\Delta_{ijj'})\}_{p,p' \in \{1,2\}}$ and its specific form is presented in Section 2.2.2. Moreover, we assume that the spatial covariance approaches zero as the distance between units increases, $C_{pp'}(\Delta) \to 0$ as $\Delta \to \infty$, an assumption essential to the estimation procedure in Section 2.3.

2.2.2 Bivariate Matérn structure for spatial covariance

Due to its flexibility, we consider the Matérn class to model the spatial covariance. Specifically, assume that the covariance matrix $\boldsymbol{C}(\cdot)$ has a bivariate Matérn structure. If $C_{pp'}(\Delta)$ is the spatial cross covariance function for $U^p(s_1)$ and $U^{p'}(s_2)$ that are measured at locations $s_1, s_2 \in \mathcal{D} \subset \mathbb{R}$ where $\Delta = s_1 - s_2$, then for $p, p' = 1, 2$, the bivariate Matérn class of cross covariance functions defines $C_{pp'}(\Delta) = \sigma_{pp'} M(\Delta | \nu_{pp'}, a_{pp'})$ with Matérn correlation function $M(\Delta|\nu, a) = \{2^{1-\nu}/\Gamma(\nu)\}(a|\Delta|)^\nu K_\nu(a|\Delta|)$, where K_ν is a modified Bessel function of the second kind and $\nu > 0$ and $a > 0$ are smoothness and scale parameters, respectively (Apanasovich et al. 2012; Gneiting et al. 2010; Matérn 1986). (For extension to d-dimensional locations, one only needs to replace the absolute value in the correlation function with the Euclidean norm.)

The smoothness parameter ν of the Matérn correlation function governs the differentiability of the process, where larger values of ν indicate a smoother process. (Special cases of the Matérn correlation function include the exponential ($\nu = 1/2$) and Whittle ($\nu = 1$) models, as well as the Gaussian when $\nu = \infty$.) With ν fixed, a governs how fast the correlation decays with distance. Larger values of a indicate a faster decay, and $1/a$ is sometimes called the correlation length. The auto-covariance components $C_{11}(\Delta)$ and $C_{22}(\Delta)$ are common Matérn covariance functions. For example σ_{11} represents the spatial variance of the process and is known as the partial sill in the spatial literature. The spatial cross covariance, represented by σ_{12}, is a function of the cross-correlation parameter ρ_{12} and each of the marginal spatial variances: $\sigma_{12} = \sigma_{21} = \rho_{12}\sqrt{\sigma_{11}\sigma_{22}}$.

2.2.3 Modeling of functional processes

In keeping with the multilevel terminology used in Di et al. (2009) and Staicu et al. (2010), we call $\boldsymbol{Z}_i(t)$ level 1 functions and $\boldsymbol{W}_{ij}(t)$ level 2 functions. Let $\boldsymbol{Z}_i(t)$ and $\boldsymbol{W}_{ij}(t)$ be processes in $\mathcal{L}^2[0,1] \times \mathcal{L}^2[0,1]$, and let $\{\boldsymbol{\phi}_k^Z(t) = [\phi_{k1}^Z(t), \phi_{k2}^Z(t)]^{\mathrm{T}}$:

$k \geq 1\}$ and $\{\boldsymbol{\phi}_\ell^W(t) = [\phi_{\ell 1}^W(t), \phi_{\ell 2}^W(t)]^{\mathrm{T}} : \ell \geq 1\}$ be two sets of orthogonal bivariate basis functions in $\mathcal{L}^2[0,1] \times \mathcal{L}^2[0,1]$ with respect to the norm induced by the inner product $< (f_1, g_1), (f_2, g_2) >= \int f_1 f_2 + \int g_1 g_2$. The functional processes can be expanded as $\boldsymbol{Z}_i(t) = \sum_{k=1}^{\infty} \xi_{i,k} \boldsymbol{\phi}_k^Z(t)$ and $\boldsymbol{W}_{ij}(t) = \sum_{\ell=1}^{\infty} \zeta_{ij,\ell} \boldsymbol{\phi}_\ell^W(t)$ where $\xi_{i,k}$ and $\zeta_{ij,\ell}$ are random basis coefficients calculated as $< \boldsymbol{Z}_i(t), \boldsymbol{\phi}_k^Z(t) >$ and $< \boldsymbol{W}_{ij}(t), \boldsymbol{\phi}_\ell^W(t) >$, respectively. Using this expansion, we can rewrite model (2.2): $\boldsymbol{Y}_{ij}(t, s_{ij}) = \boldsymbol{\mu}_{G(i)}(t) + \sum_{k=1}^{\infty} \xi_{i,k} \boldsymbol{\phi}_k^Z(t) + \sum_{\ell=1}^{\infty} \zeta_{ij,\ell} \boldsymbol{\phi}_\ell^W(t) + \boldsymbol{U}_i(s_{ij}) + \boldsymbol{\epsilon}_{ijt}$. In practice, finite truncations are used instead; let N_Z and N_W be the truncation values for $\boldsymbol{Z}_i(\cdot)$ and $\boldsymbol{W}_{ij}(\cdot)$, respectively, leading to the simplified model

$$\boldsymbol{Y}_{ij}(t, s_{ij}) = \boldsymbol{\mu}_{G(i)}(t) + \sum_{k=1}^{\mathrm{N}_Z} \xi_{i,k} \boldsymbol{\phi}_k^Z(t) + \sum_{\ell=1}^{\mathrm{N}_W} \zeta_{ij,\ell} \boldsymbol{\phi}_\ell^W(t) + \boldsymbol{U}_i(s_{ij}) + \boldsymbol{\epsilon}_{ijt}. \tag{2.3}$$

There are several ways to chose the basis functions by either using predetermined bases such as Fourier or wavelets or using the basis formed by the eigenfunctions of the covariance operator. We opt for the latter choice; if $\boldsymbol{K}^Z$ and $\boldsymbol{K}^W$ are the covariance operators of $\boldsymbol{Z}$ and $\boldsymbol{W}$ respectively, then by applying Mercer's Theorem (Indritz 1963) to multivariate data the eigenfunctions are obtained from the spectral decompositions of the respective covariance functions. In particular, $\boldsymbol{K}^Z(t, t') = \sum_{k=1}^{\infty} \lambda_k^Z \boldsymbol{\phi}_k^Z(t) \{\boldsymbol{\phi}_k^Z(t')\}^{\mathrm{T}}$ and $\boldsymbol{K}^W(t, t') = \sum_{\ell=1}^{\infty} \lambda_\ell^W \boldsymbol{\phi}_\ell^W(t) \{\boldsymbol{\phi}_\ell^W(t')\}^{\mathrm{T}}$. In this case, the function expansions leading to (2.3) are known as Karhunen-Loève expansions (Karhunen 1947; Loève 1945) and the random coefficients ξ_{ik} and $\zeta_{ij,\ell}$ as functional principal components (FPC) scores. Furthermore the FPC scores ξ_{ik} and $\zeta_{ij,\ell}$ are assumed to be uncorrelated over k and ℓ respectively, are zero-mean, and have variances λ_k^Z and λ_ℓ^W, respectively. Additionally, following the assumption that $\boldsymbol{Z}_i$ and $\boldsymbol{W}_{ij}$ are uncorrelated, it is assumed that $\{\xi_{i,k} : k = 1, 2, \ldots\}$ are uncorrelated with $\{\zeta_{ij,\ell} : \ell = 1, 2, \ldots\}$.

2.2.4 Notation for balanced design

To facilitate exposition of the estimation section, we rewrite in matrix form the covariance operators in Section 2.2.1 for a dense, balanced design. Let $\boldsymbol{t} = [t_1, \ldots, t_{\mathrm{L}}]^{\mathrm{T}}$ be the $\mathrm{L} \times 1$ vector of subunits at which each process is measured within a unit. The $2\mathrm{L} \times 1$ random vector $\boldsymbol{Z}_i = [Z_i^{1,\mathrm{T}}(\boldsymbol{t}), Z_i^{2,\mathrm{T}}(\boldsymbol{t})]^{\mathrm{T}}$ has the block covariance matrix $\mathrm{Var}(\boldsymbol{Z}_i) = \boldsymbol{K}^Z$ with blocks $\boldsymbol{K}_{pp'}^Z = \{K_{pp'}^Z(t_h, t_k)\}_{h,k \in \{1,\ldots,\mathrm{L}\}}$. The $2\mathrm{L} \times 1$ random vec-

tor $\boldsymbol{W}_{ij} = [W_{ij}^{1,\mathrm{T}}(\boldsymbol{t}), W_{ij}^{2,\mathrm{T}}(\boldsymbol{t})]^{\mathrm{T}}$ has covariance matrix $\mathrm{Var}(\boldsymbol{W}_{ij}) = \boldsymbol{K}^{W}$ defined analogously.

For the spatial process, let $\boldsymbol{s}_i = [s_{i1}, s_{i2}, \ldots, s_{i\mathrm{M}}]^{\mathrm{T}}$ be the $\mathrm{M} \times 1$ vector of unit locations for subject i with the convention that $s_{i1} = 0$, where s_{ij} is the relative distance of unit j from the first unit location. The $\mathrm{M} \times \mathrm{M}$ matrix $\boldsymbol{\Delta}_i = \{\Delta_{ijj'}\}_{j,j'=1}^{\mathrm{M}}$ is that which is formed from every pairwise distance between the M units located in subject $i = 1, \ldots, \mathrm{N}$. The (matrix-valued) block covariance matrix of $\boldsymbol{U}_i(\boldsymbol{s}_i) = [U_i^{1,\mathrm{T}}(\boldsymbol{s}_i), U_i^{2,\mathrm{T}}(\boldsymbol{s}_i)]^{\mathrm{T}}$ is $\boldsymbol{C}(\boldsymbol{\Delta}_i)$ which has blocks $C_{pp'}(\boldsymbol{\Delta}_i)$. Furthermore, assuming the bivariate Matérn covariance structure in Section (2.2.2) gives the parametric form $C_{pp'}(\boldsymbol{\Delta}_i) = \sigma_{pp'} M(\boldsymbol{\Delta}_i | \nu_{pp'}, a_{pp'})$.

2.3 Estimation

To address our primary objectives of estimating the group means and understanding how the distance between units affects the spatial correlation, we have developed an estimation procedure that identifies the key components that account for variation at each hierarchy level. Estimates of the Matérn parameters will identify the spatial signal across units as well as the strength and direction of the spatial correlation between the two response curves that we model jointly. Moreover, once estimates of the covariances $\boldsymbol{K}^{W}$, $\boldsymbol{K}^{Z}$, $\boldsymbol{C}_i$ and the error variances τ_1^2 and τ_2^2 are obtained, one can estimate the group mean functions $\boldsymbol{\mu}_{G(i)}(t)$ using generalized least squares regression with estimated covariance. The outline of our estimation procedure is:

1. Estimate the bivariate Matérn parameters for the spatial covariance (Section 2.3.1);
2. Estimate $\boldsymbol{K}^{Z}$ and $\boldsymbol{K}^{W}$ using method of moments (Section 2.3.2);
3. Obtain eigenfunctions and eigenvalues for $\boldsymbol{K}^{Z}$ and $\boldsymbol{K}^{W}$ through MFPCA and also estimate the error variance τ_p^2 for $p = 1, 2$ (Section 2.3.3);
4. Estimate group mean functions $\boldsymbol{\mu}_{G(i)}(t)$ using generalized least squares (GLS) regression with estimated covariance matrix (Section 2.3.4).

2.3.1 Spatial covariance estimation

Estimation of the spatial covariance matrix is done in two parts: 1) we define a raw estimator based on method of moments and 2) fit a parametric bivariate covariance structure that leads to a positive semi-definite, smoothed covariance estimator. Essential to part one is the assumption that the spatial correlation approaches zero as the distance between observations increases. Although the correlation will never be exactly zero, using this we can assume that there exists some correlation range Δ^* (chosen based on scientific or expert knowledge) for which units can be considered uncorrelated if the distance between them exceeds the correlation range: $C_{pp'}(\Delta) \approx 0$ if $\Delta \geq \Delta^*$.

The preferred moment-based estimation method in the spatial literature is based on the (cross) semivariogram (Cressie 1993), defined as half of the variance of the difference in residuals for observations separated by a given distance. Our setting requires a generalization of this standard approach because spatial variation is just one piece of the complex model given in (2.2). For spatial lag $\delta < \Delta^*$, denote by $\mathcal{N}(\delta, \epsilon)$ the set (across all subjects and all groups) of unit-pairs within the same subject whose distance from one another is within a tolerance, ϵ, of δ. We select ϵ so that at least 30 distinct unit-pairs are in $\mathcal{N}(\delta, \epsilon)$ (see Journel and Huijbregts 1978; Cressie 1993, Ch. 2). Define $\mathcal{N}(\delta, \epsilon)$ to be the same for all spatial lags $\delta > \Delta^*$. In summary, for $\Delta_{ijj'} = |s_{ij} - s_{ij'}|$,

$$\mathcal{N}(\delta, \epsilon) = \begin{cases} \{(i, j, j') : j \neq j' \ \& \ \Delta_{ijj'} \in [\delta - \epsilon, \delta + \epsilon]\} & \text{if } 0 < \delta < \Delta^* \\ \{(i, j, j') : j \neq j' \ \& \ \Delta_{ijj'} \geq \Delta^*\} & \text{if } \delta \geq \Delta^*. \end{cases} \tag{2.4}$$

Define $G_{ijj'}^{pp'}(t, t') = \frac{1}{2}\{Y_{ij}^{p}(t, s_{ij}) - Y_{ij'}^{p'}(t', s_{ij'})\}^2 - \frac{1}{2}\{Y_{ij}^{p}(t, s_{ij}) - Y_{ij}^{p'}(t', s_{ij})\}^2$, on which the following method of moments estimators are based. $G_{ijj'}^{pp'}(t, t')$ is only useful for observations from two different units since $G_{ijj'}^{pp'}(t, t') = 0$ if $j = j'$, and it is not unbiased for the (cross) semi-variogram $\gamma_{pp'}(\Delta_{ijj'}) = C_{pp'}^{U}(0) - C_{pp'}^{U}(\Delta_{ijj'})$, but instead is inflated by the term $\eta_{pp'}(t, t') = K_{pp'}^{W}(t, t') + \tau_p^2 \mathrm{I}(p = p', t = t')$. When $p = p'$ and $t = t'$, $G_{ijj'}^{pp'}(t, t')$ = $1/2\{Y^p(t, s_{ij}) - Y^p(t, s_{ij'})\}^2$ and is analogous to the classical semivariogram estimator. Define the moments-based nearest neighbor estimator of

the (cross) semivariogram for spatial lag $\delta > 0$ as

$$\widetilde{\gamma}_{pp'}(\delta, \epsilon) = \frac{1}{\mathrm{L}^2 |\mathcal{N}(\delta, \epsilon)|} \sum_{(i,j,j') \in \mathcal{N}(\delta,\epsilon)} \sum_{t,t'} G_{ijj'}^{pp'}(t, t'), \tag{2.5}$$

where $|\cdot|$ indicates cardinality of a set. For $\delta = 0$, define

$$\widetilde{\gamma}^0_{pp',\epsilon} = \frac{1}{\mathrm{L} |\mathcal{N}(\Delta^*, \epsilon)|} \sum_{(i,j,j') \in \mathcal{N}(\Delta^*,\epsilon)} \sum_{t=t'} G_{ijj'}^{pp'}(t, t'). \tag{2.6}$$

Let $\bar{\eta}_{a,pp'} = \mathrm{L}^{-1} \sum_{t=t'} \eta_{pp'}(t, t')$ and $\bar{\eta}_{b,pp'} = \mathrm{L}^{-2} \sum_{t,t'} \eta_{pp'}(t, t')$. Then (2.5) has expectation $\mathrm{E}\{\widetilde{\gamma}_{pp'}(\delta)\} = C_{pp'}(0) - C_{pp'}(\delta) + \bar{\eta}_{b,pp'}$, and (2.6) has expectation $\mathrm{E}\{\widetilde{\gamma}^0_{pp'}\} = C_{pp'}(0) - C_{pp'}(\Delta^*) + \bar{\eta}_{a,pp'}$ which simplifies to $\mathrm{E}\{\widetilde{\gamma}^0_{pp'}\} \approx C_{pp'}(0) + \bar{\eta}_{a,pp'}$ by the assumption that $C_{pp'}(\Delta) \approx 0$ if $\Delta \geq \Delta^*$. Here $\eta_{pp'}(t, t')$ differs from the nugget effect in that it also includes the level 2 functional covariance operator. Our estimation procedure will account for the nuisance parameters $\eta_{pp'}(t, t')$, but only the parameters of the bivariate Matérn covariance are of interest. Then raw estimator for the (cross) covariogram becomes

$$\widetilde{C}_{pp'}(\delta) = \begin{cases} \widetilde{\gamma}^0_{pp'} & \text{if } \delta = 0; \\ \widetilde{\gamma}_{pp'}(\Delta^*) - \widetilde{\gamma}_{pp'}(\delta) & \text{if } 0 < \delta < \Delta^*; \end{cases} \tag{2.7}$$

we set $\widetilde{C}_{pp'}(\delta) = 0$ if $\delta \geq \Delta^*$. When $0 < \delta < \Delta^*$, the covariance estimator given in (2.7) is unbiased. For $\delta = 0$, $\widetilde{C}_{pp}(0) = \widetilde{\gamma}^0_{pp'}$ is upwardly biased due to the positive term $\bar{\eta}_{a,pp'}$.

The moments-based raw estimate of the spatial covariance from (2.7) is neither guaranteed to be smooth nor positive semi-definite. To obtain a smoothed estimate, (2.7) provides the foundation for estimating the Matérn parameters through a procedure that emulates maximum likelihood but uses the estimated covariance in place of data. In order to find a suitable function f to maximize, assume there exists some unobserved, bivariate Gaussian spatial process $\boldsymbol{Q}(s)$ measured at locations $\boldsymbol{s}_i = [s_{i1}, \ldots, s_{i\mathrm{m}}]^{\mathrm{T}}$, $s_{ij} \in \mathbb{R}$, for $j = 1, \ldots, \mathrm{m}$ within subject $i = 1, \ldots, \mathrm{n}$. The pairwise distances $\Delta_{ijj'} = |s_{ij} - s_{jj'}|$ form the distance matrix $\boldsymbol{\Delta}_i = \{\Delta_{ijj'}\}_{j,j'=1}^{\mathrm{m}}$. Let $\boldsymbol{\Sigma}(\boldsymbol{\Delta}_i; \boldsymbol{\theta}, \boldsymbol{\eta})$ be a parametric covariance matrix with elements $\Sigma_{pp'}(\Delta_{ijj'}; \boldsymbol{\theta}, \boldsymbol{\eta}) = \sigma_{pp'} M(\Delta_{ijj'} | \nu_{pp'}, a_{pp'}) + \bar{\eta}_{a,pp'} \mathrm{I}(\Delta_{ijj'} = 0)$ where $\boldsymbol{\theta}$ indicates the bivariate Matérn pa-

rameters we wish to estimate, and $\boldsymbol{\eta}$ indicates the nuisance parameters. Now, let $\boldsymbol{Q}_i \overset{indep}{\sim} \mathrm{N}_{2\mathrm{m}}\big(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\Delta}_i; \boldsymbol{\theta}, \boldsymbol{\eta})\big)$ where $\boldsymbol{Q}_i = [Q^{1,\mathrm{T}}(\mathbf{s}_i), Q^{2,\mathrm{T}}(\boldsymbol{s}_i)]^{\mathrm{T}}$. Furthermore, assume that all subjects i are observed on the same equally spaced $\mathrm{m} \times 1$ grid of points $\mathbf{s}_{\mathrm{grid}}$ with the convention $s_{\mathrm{grid},1} = 0$, and $s_{\mathrm{grid},\mathrm{m}} < \Delta^*$, so that $\boldsymbol{\Delta}_i \equiv \boldsymbol{\Delta}_{\mathrm{grid}}$ for all i, and the largest pairwise distance between locations will be less than Δ^*. Thus, the $\boldsymbol{Q}_i$ are i.i.d. multivariate normal random variables with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\Delta}_{\mathrm{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta})$.

The log-likelihood function, minus arbitrary constants, is $\ell(\boldsymbol{\theta}, \boldsymbol{\eta}; \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = -1/2 \sum_{i=1}^{n} \Big[\log|\boldsymbol{\Sigma}(\boldsymbol{\Delta}_{\mathbf{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta})| + \Big\{\boldsymbol{q}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Delta}_{\mathrm{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta}) \boldsymbol{q}_i\Big\}\Big]$, which can alternatively be written in terms of the sample covariance $\boldsymbol{C}^q = 1/n \sum_{i=1}^{n} \boldsymbol{q}_i \boldsymbol{q}_i^{\mathrm{T}}$ by using the trace: $\ell(\boldsymbol{\theta}, \boldsymbol{\eta}; \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = -n/2 \Big[\log|\boldsymbol{\Sigma}(\boldsymbol{\Delta}_{\mathbf{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta})| + \mathrm{Tr}\Big\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Delta}_{\mathrm{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta}) \boldsymbol{C}^q\Big\}\Big]$. Since we do not observe the data $\boldsymbol{q}$ and, hence, cannot directly use the sample covariance $\boldsymbol{C}^q$ in the likelihood, we must use a covariance estimate based on (2.7) as a substitute. Evaluating (2.7) for each element in $\boldsymbol{\Delta}_{\mathrm{grid}}$ gives the estimated covariance matrix $\widetilde{\boldsymbol{C}}_{\mathrm{grid}} = \{\widetilde{C}_{pp'}(\boldsymbol{\Delta}_{\mathrm{grid}})\}_{p,p' \in \{1,2\}}$, which we use in place of $\boldsymbol{C}^q$. Note that $\mathrm{E}(\widetilde{\boldsymbol{C}}_{\mathrm{grid}}) = \{\sigma_{pp'} M(\boldsymbol{\Delta}_{\mathrm{grid}} | \nu_{pp'}, a_{pp'}) + \bar{\eta}_{a,pp'} \mathbf{I}\}_{p,p' \in \{1,2\}}$, where $\mathbf{I}$ is the $\mathrm{m} \times \mathrm{m}$ identity matrix. Lastly, we maximize the function $f(\boldsymbol{\theta}, \boldsymbol{\eta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = -n/2 \Big[\log|\boldsymbol{\Sigma}(\boldsymbol{\Delta}_{\mathbf{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta})| + \mathrm{Tr}\Big\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Delta}_{\mathrm{grid}}; \boldsymbol{\theta}, \boldsymbol{\eta}) \widetilde{\boldsymbol{C}}_{\mathrm{grid}}\Big\}\Big]$ to find the bivariate Matérn parameter estimates $\widehat{\boldsymbol{\theta}}$, resulting in the smooth estimator $\widehat{\boldsymbol{C}} = \boldsymbol{C}(\widehat{\boldsymbol{\theta}})$.

To ensure the estimated parameters produce a valid bivariate Matérn structure in which the marginal covariances $\widehat{\boldsymbol{C}}_{11}$ and $\widehat{\boldsymbol{C}}_{22}$ and the entire bivariate covariance matrix $\widehat{\boldsymbol{C}}$ are positive semi-definite, we implement the three-step maximum likelihood algorithm proposed by Apanasovich et al. (2012):

1: Fit the marginal models to get the parameters σ_{pp}, ν_{pp} and a_{pp} for $p = 1, 2$.

2: Define $\nu_{12} = (\nu_{11} + \nu_{22})/2 + \Delta_A$ for $\Delta_A \geq 0$, $a_{12} = (a_{11}^2 + a_{22}^2)/2 + \Delta_B$ for $\Delta_B \geq 0$ and $\sigma_{12}^2 = \rho_{12}^2 (\sigma_{11}\sigma_{22}) \prod_{k=1}^{3} \psi_{12}^{(k)}$.

3: Holding the univariate parameters σ_{pp}, ν_{pp} and a_{pp} fixed from Step 1, fit the bivariate model and estimate Δ_A, Δ_B, and ρ_{12}.

The $\psi_{pp'}^{(k)}$ in Step 2 are defined for components p, p' as $\psi_{pp'}^{(1)} = \mathcal{B}(\nu_{pp'}, d/2)^2 / \mathcal{B}\{(\nu_{pp} + \nu_{p'p'})/2, d/2\}^2$, $\psi_{pp'}^{(2)} = (a_{pp} a_{p'p'} / a_{pp'}^2)^{2\Delta_A}$, and $\psi_{pp'}^{(3)} = \Gamma^2\{(\nu_{pp} + \nu_{p'p'})/2\} a_{pp}^{2\nu_{pp}} a_{p'p'}^{2\nu_{p'p'}} / \{a_{pp'}^{2(\nu_{pp} + \nu_{p'p'})} \Gamma(\nu_{pp}) \Gamma(\nu_{p'p'})\}$, where $\mathcal{B}(\cdot, \cdot)$ is the Beta function, and $d = 1$ since $s_{ij} \in \mathbb{R}$. Extensions to multivariate responses

are possible by slight modifications of this algorithm; see Apanasovich et al. (2012). Initial values of the Matérn parameters for the maximum likelihood procedure can be found through weighted least squares as shown in Appendix A.3.

2.3.2 Raw functional covariance estimates

In order to implement FPCA, we will need to find asymptotically consistent estimates of $\boldsymbol{K}^Z(t,t')$ and $\boldsymbol{K}^W(t,t')$. We do this in a manner similar to that in Staicu et al. (2010) and Di et al. (2009), utilizing the same notions of total and between covariance that are analogous to the variance decomposition found in mixed ANOVA models. Define the total covariance of unit-level functions measured within the same unit j as $K^{\boldsymbol{Y}}_{\text{Total},pp'}(t,t') = \text{Cov}\{Y^p_{ij}(t,s_{ij}), Y^{p'}_{ij}(t',s_{ij})\}$ and the between covariance of the unit-level functions that are distance $\Delta_{ijj'} > 0$ apart as $K^{\boldsymbol{Y}}_{\text{Between},pp'}(t,t',\Delta_{ijj'}) = \text{Cov}\{Y^p_{ij}(t,s_{ij}), Y^{p'}_{ij'}(t',s_{ij'})\}$. In terms of model (2.2), the total and between covariance quantities can be written $K^{\boldsymbol{Y}}_{\text{Total},pp'}(t,t') = K^Z_{pp'}(t,t') + K^W_{pp'}(t,t') + C_{pp'}(\Delta = 0) + \tau^2_p \text{I}(t = t', p = p')$ and $K^{\boldsymbol{Y}}_{\text{Between},pp'}(t,t',\Delta_{ijj'}) = K^Z_{pp'}(t,t') + C_{pp'}(\Delta_{ijj'})$. These covariance quantities and their decompositions in terms of model (2.2) are key in our estimation method, particularly because they provide an intuitive method of moments-based approach with desired consistency properties.

Define the moments-based nearest neighbor estimator

$$\widehat{K}^Z_{pp'}(t,t') = \frac{1}{|\mathcal{N}(\Delta^*)|} \sum_{(i,j,j')\in\mathcal{N}(\Delta^*)} \{Y^p_{ij}(t,s_{ij}) - \bar{Y}^p_{G(i)}(t)\}\{Y^{p'}_{ij'}(t',s_{ij'}) - \bar{Y}^{p'}_{G(i)}(t')\}, \quad (2.8)$$

where $p,p' = 1,2$, $\mathcal{N}(\Delta^*)$ is given in (2.4) and $\bar{Y}^p_{G(i)}(t)$ is the group mean at t over all unit locations and all subjects in the group to which subject i belongs. This estimator makes use of the property that $K^{\boldsymbol{Y}}_{\text{Between},pp'}(t,t',\Delta^*) \approx K^Z_{pp'}(t,t')$ since $C_{pp'}(\Delta) \approx 0$ if $\Delta \geq \Delta^*$.

To estimate $\boldsymbol{K}^W(t,t')$ we diverge from the methods of Staicu et al. (2010) and Di et al. (2009). Consider the residuals $R^p_{ij}(t,s_{ij}) = Y^p_{ij}(t,s_{ij}) - \widehat{L}^p_i(t)$ for some smooth estimator $\widehat{L}^p_i(t) = \widehat{\mu}^p_{G(i)}(t) + \widehat{Z}^p_i(t)$ obtained using penalized regression splines or similar methods to smooth the pooled data across units for each subject i and each bivariate response component $p = 1,2$. We can model the residuals as $\boldsymbol{R}_{ij}(t,s_{ij}) = \boldsymbol{W}_{ij}(t) + \boldsymbol{U}_i(s_{ij}) + \boldsymbol{\epsilon}_{ijt}$. Define the total covariance of the residuals as $K^{\boldsymbol{R}}_{\text{Total},pp'}(t,t') =$

$\text{Cov}\{R_{ij}^{p}(t, s_{ij}), R_{ij}^{p'}(t', s_{ij})\}$, which has the decomposition $K_{\text{Total},pp'}^{\boldsymbol{R}}(t, t') = K_{pp'}^{W}(t, t') + C_{pp'}(0) + \tau_p^2 \text{I}(t = t', p = p')$. Let $K_{pp',\text{Inflated}}^{W}(t, t') = K_{pp'}^{W}(t, t') + \tau_p^2 \text{I}(t = t', p = p')$ denote the covariance process of $\boldsymbol{W}_{ij}(t)$ that is inflated by the error. Its estimator is given by

$$\widehat{K}_{pp',\text{Inflated}}^{W}(t, t') = \frac{1}{\text{NM}} \sum_{i=1}^{\text{N}} \sum_{j=1}^{\text{M}} \{R_{ij}^{p}(t, s_{ij}) R_{ij}^{p'}(t', s_{ij})\} - \widehat{C}_{pp'}(0), \tag{2.9}$$

where $\widehat{C}_{pp'}(0)$ is defined in (2.7).

2.3.3 Multivariate multilevel FPCA

Multilevel FPCA (MFPCA) as presented in Di et al. (2009) retrieves the eigenvalues and eigenfunctions that comprise the covariance expansion for level 1 and level 2 univariate functional processes observed with error by implementing FPCA on smoothed covariance matrix estimates. We introduce a novel adaptation of MFPCA that incorporates the multivariate structure for multilevel spatial functional data.

Since we are assuming the curves are observed with error, we must smooth the raw level 1 and level 2 covariance estimates given in (2.8) and (2.9) in order to implement MFPCA. For level 1, we use (2.8) to form the raw matrices $\widehat{\boldsymbol{K}}_{pp'}^{Z}$ for $p, p' = 1, 2$ and smooth each univariate and cross covariance separately to obtain $\widetilde{\boldsymbol{K}}_{\text{sm},pp'}^{Z}$. These individually smoothed matrices combine to form the bivariate smoothed covariance $\widetilde{\boldsymbol{K}}_{\text{sm}}^{Z}$. The intuition behind smoothing the submatrices separately and then combining them into the bivariate matrix versus combining the raw estimates first and smoothing the bivariate matrix is that the delineations between the submatrices need not be smooth within the bivariate matrix.

For level 2, we use (2.9) to form $\widehat{\boldsymbol{K}}_{\text{inflated},pp'}^{W}$ for $p, p' = 1, 2$. Since the diagonals of the univariate matrices $\widehat{\boldsymbol{K}}_{\text{inflated},11}^{W}$ and $\widehat{\boldsymbol{K}}_{\text{inflated},22}^{W}$ are inflated by the error variances τ_1^2 and τ_2^2, respectively, we ignore the diagonals when smoothing. After smoothing $\widehat{\boldsymbol{K}}_{\text{inflated},pp'}^{W}$ separately for $p, p' = 1, 2$ to obtain $\widetilde{\boldsymbol{K}}_{\text{sm},pp'}^{W}$, the smoothed estimates combine to form the bivariate covariance matrix $\widetilde{\boldsymbol{K}}_{\text{sm}}^{W}$. In contrast to the raw covariance estimate $\widehat{\boldsymbol{K}}_{\text{inflated}}^{W}$, the smoothed level 2 bivariate matrix $\widetilde{\boldsymbol{K}}_{\text{sm}}^{W}$ is no longer inflated by the error variance, a property which we can use to obtain an estimator of the error

variances,

$$\widehat{\tau}_p^2 = \frac{1}{\mathrm{L}} \sum_{t=1}^{\mathrm{L}} \left\{ \widehat{K}^W_{pp,\text{Inflated}}(t,t) - \widetilde{K}^W_{pp,\text{sm}}(t,t) \right\}. \tag{2.10}$$

In implementing MFPCA, we find the eigenfunctions $\boldsymbol{e}(t)$ and the eigenvalues $\widehat{\lambda}$ of the smoothed bivariate covariance estimates $\widetilde{\boldsymbol{K}}_{\text{sm}}$ for both level 1 and level 2 processes. The truncated spectral decompositions of these covariances as presented in Section 2.2.3 lead to the estimators $\widehat{\boldsymbol{K}}^Z(t,t') = \sum_{k=1}^{\mathrm{N}_Z} \widehat{\lambda}_k^Z \boldsymbol{e}_k^Z(t)\{\boldsymbol{e}_k^Z(t')\}^{\mathrm{T}}$ and $\widehat{\boldsymbol{K}}^W(t,t') = \sum_{\ell=1}^{\mathrm{N}_W} \widehat{\lambda}_\ell^W \boldsymbol{e}_\ell^W(t)\{\boldsymbol{e}_\ell^W(t')\}^{\mathrm{T}}$ of the cross covariance matrices for the level 1 and 2 processes. The truncation values N_Z and N_W are chosen based on the proportion of variation explained by the eigenvalues as suggested in Di et al. (2009). Using level 1 as an example, specify a cumulative explained variance threshold P_1 and and individual explained variance threshold P_2. Define $\mathrm{N}_Z = \min\{k : p_{1k}^Z \geq \mathrm{P}_1, p_{2k}^Z < \mathrm{P}_2\}$ where $p_{k1}^Z = \sum_{i=1}^k \widehat{\lambda}_i^Z / \sum_{j=1}^n \widehat{\lambda}_j^Z$, $p_{k2}^Z = \widehat{\lambda}_k^Z / \sum_{j=1}^n \widehat{\lambda}_j^Z$ and the positive eigenvalues are the first $n \geq k$ eigenvalues. N_W for level 2 is found analogously.

2.3.4 Estimate group mean functions

Assume that the group means are modeled parametrically as $\boldsymbol{\mu}_{G(i)}(t) = \boldsymbol{X}_i\boldsymbol{\beta}$. Once the covariance estimates have been found, we can use generalized least squares (GLS) regression assuming known (or estimated) covariance to estimate the diet means. Let $\boldsymbol{Y}_{p,ij} = [Y_{ij}^p(t_1, s_{ij}), \ldots, Y_{ij}^p(t_{\mathrm{L}}, s_{ij})]^{\mathrm{T}}$ be the $\mathrm{L} \times 1$ vector obtained by stacking the responses over subunits t_k, $k = 1, \ldots, \mathrm{L}$, for component p in unit j within subject i. Stacking $\boldsymbol{Y}_{p,ij}$ over units $j = 1, \ldots, \mathrm{M}$ yields the $\mathrm{ML} \times 1$ vector $\boldsymbol{Y}_{p,i}$, which is then stacked over p to form the $2\mathrm{ML} \times 1$ overall response vector for subject i, $\boldsymbol{Y}_i = [\boldsymbol{Y}_{1,i}^{\mathrm{T}}, \boldsymbol{Y}_{2,i}^{\mathrm{T}}]^{\mathrm{T}}$. Define $\boldsymbol{V}_{i,pp'} = \mathrm{Cov}(\boldsymbol{Y}_{p,i}, \boldsymbol{Y}_{p',i})$. Then the response vector $\boldsymbol{Y}_i$ for subject $i = 1, \ldots, \mathrm{N}$ has the cross covariance matrix $\boldsymbol{V}_i = \{\boldsymbol{V}_{i,pp'}\}_{p,p' \in \{1,2\}}$. Combining the estimates from previous sections gives $\widehat{\boldsymbol{V}}_{i,pp'} = \mathbf{J}_{\mathrm{M}} \otimes \widehat{\boldsymbol{K}}^Z_{pp'} + \mathbf{I}_{\mathrm{M}} \otimes \widehat{\boldsymbol{K}}^W_{pp'} + \widehat{\boldsymbol{C}}^U_{pp'}(\boldsymbol{\Delta}_i) \otimes \mathbf{J}_{\mathrm{L}} + \widehat{\tau}_p^2 \delta_{pp'} \mathbf{I}_{\mathrm{ML}}$, where $\otimes$ indicates the Kronecker product, $\mathbf{J}_{\mathrm{M}}$ is a $\mathrm{M} \times \mathrm{M}$ matrix of ones, $\mathbf{J}_{\mathrm{L}}$ is a $\mathrm{L} \times \mathrm{L}$ matrix of ones, $\mathbf{I}_{\mathrm{M}}$ is the $\mathrm{M} \times \mathrm{M}$ identity matrix, $\mathbf{I}_{\mathrm{ML}}$ is the $\mathrm{ML} \times \mathrm{ML}$ identity matrix, and $\delta_{pp'} = 1$ if $p = p', 0$ else. Employing GLS estimation with the estimated cross covariance matrix, we obtain $\widehat{\boldsymbol{\mu}}_{G(i)}(t) = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ where $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}})[\sum_{i=1}^{\mathrm{N}} \boldsymbol{X}_i^{\mathrm{T}} \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{Y}_i]$ and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = [\sum_{i=1}^{\mathrm{N}} \boldsymbol{X}_i^{\mathrm{T}} \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i]^{-1}$.

2.4 Simulations

The purpose of our simulations is two-fold: 1) compare our model with simplified versions to assess performance gain (Scenarios 1-4), and 2) explore robustness to model misspecification (Scenarios 5 & 6). For Scenarios 1-4 we consider four estimating models:

1. Full (FULL): the multivariate spatial model in (2.3);
2. Non-spatial (NS): the model from (2.3) with no spatial process ($\boldsymbol{U}_i(s) \equiv 0$);
3. Univariate (UNIV): the model from (2.3) applied separately to each response;
4. True-GLS (TRUE): results of GLS estimation when using each subject's true cross covariance matrix $\boldsymbol{V}_i$ (Section 2.3.4) that comes from the correct generating model.

The next sections discuss the simulation specifications and results for Scenarios 1-4 and briefly summarize our findings for Scenarios 5 and 6. The latter scenarios are discussed in more detail in Appendix A.2.

2.4.1 Data generation

In Scenarios 1-4 we generate data from FULL according to the differing sample sizes and spatial cross-correlations shown in Table 2.1. All scenarios use 100 Monte Carlo (MC) replications. There are D=2 groups with mean functions $\mu_d^1(t) = 3t + d$ and $\mu_d^2(t) = d - t + t^2$ for $d = 1, 2$. Each group has $\mathrm{N} = 10, 50$ subjects, $\mathrm{M} = 20$ units per subject, and $\mathrm{L} = 30$ subunits per unit that are equally spaced in $[0, 1]$. Unit locations $\{\boldsymbol{s}_{ij} : j = 1, \ldots, \mathrm{M}\}$ for subject i are assumed i.i.d. and are obtained by generating from the uniform distribution on $[0, 15]$ to emulate the colon slices in the colon carcinogenesis data which can be up to 15 millimeters in length.

Table 2.1: Specifications for Scenarios 1-4

	$\rho_{12} = 0.8$	$\rho_{12} = 0.2$
$\mathrm{N} = 50$	Scenario 1	Scenario 3
$\mathrm{N} = 10$	Scenario 2	Scenario 4

The bivariate Matérn parameters are $\rho_{12} = 0.2, 0.8$, $\sigma_{11}, \sigma_{22} = 1$, $\nu_{11}, \nu_{22}, \nu_{12} = 1$, and $a_{11}, a_{22}, a_{12} = 4$, chosen so the auto- and cross-correlations decay to zero at distances of approximately 1 unit. For the level 1 process $\boldsymbol{Z}_i(t) = \sum_{k=1}^{\mathrm{N}_Z} \xi_{i,k} \boldsymbol{\phi}_k^Z(t)$, $\mathrm{N}_Z = 2$ with orthogonal eigenfunctions $\boldsymbol{\phi}_1^Z(t) = [\sin(2\pi t), \sqrt{3/2}(2t-1)]^{\mathrm{T}}$, $\boldsymbol{\phi}_2^Z(t) = [\cos(2\pi t), \sqrt{5/2}(6t^2 - 6t + 1)]^{\mathrm{T}}$ and eigenvalues $\lambda_1^Z = 1.25$, $\lambda_2^Z = 0.25$. For the level 2 process $\boldsymbol{W}_{ij}(t) = \sum_{\ell=1}^{\mathrm{N}_W} \zeta_{ij,\ell} \boldsymbol{\phi}_\ell^W(t)$, $\mathrm{N}_W = 2$ with orthogonal eigenfunctions $\boldsymbol{\phi}_1^W(t) = [\sin(4\pi t), \cos(6\pi t)]^{\mathrm{T}}$, $\boldsymbol{\phi}_2^W(t) = [\cos(4\pi t), \sin(8\pi t)]^{\mathrm{T}}$ and eigenvalues $\lambda_1^W = 1.25$, $\lambda_2^W = 0.375$. FPC scores are generated as $\boldsymbol{\xi}_i \sim \mathrm{N}\left(\mathbf{0}, \mathrm{diag}(\lambda_1^Z, \lambda_2^Z)\right)$ and $\boldsymbol{\zeta}_{ij} \sim \mathrm{N}\left(\mathbf{0}, \mathrm{diag}(\lambda_1^W, \lambda_2^W)\right)$. Finally, the error is generated from $\boldsymbol{\epsilon}_{ij} = [\epsilon_{ij}^1, \epsilon_{ij}^2]^T \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\tau_1^2, \tau_2^2))$ where $\tau_1^2, \tau_2^2 = 0.075$.

2.4.2 Computational Details

We compare the performance of correctly specifying FULL to the performance of NS and UNIV, using TRUE as a baseline. There are several tuning parameters in the estimation method that must be specified. First, Δ^* (Section 2.3.1) is set to be 2.5, which is conservative based on the spatial correlation decay to zero around 1. Furthermore, the tolerance ϵ is set such that each spatial lag has around 100 triplets entering its nearest neighbor set, corresponding to $\epsilon = 0.1$ for $\mathrm{N} = 10$ and $\epsilon = 0.02$ for $\mathrm{N} = 50$. We found that an equally spaced grid $\boldsymbol{s}_{\text{grid}}$ of $\mathrm{m} = 50$ points (Section 2.3.1) works well for spatial estimation for this simulation. In practice, m is a tuning parameter that should be chosen to be large enough to capture the features of the covariance matrix but small enough so that the dimensionality of the covariance matrix remains reasonable and computationally feasible. Estimate N_Z and N_W using the cumulative explained variance threshold $\mathrm{P}_1 = 0.95$ and an individual explained variance threshold $\mathrm{P}_2 = 1$ (see Section 2.3.3).

Prior to implementing the bivariate estimation methods, we recommend scaling each univariate response so that the variances are on a similar scale, particularly since scalar variances for the scores from MFPCA are estimated from information from both responses. For example, one can use $Y_{ij}^p(t, s_{ij})/s_p$ where $s_p = [(\mathrm{NML})^{-1} \sum_{i,j,t} \{Y_{ij}^p(t, s_{ij}) - \bar{Y}_p\}^2]^{1/2}$ for L subunits, M units and N subjects, and $\bar{Y}_p$ is the overall mean for response p. Also, we place constraints on the smoothness parameters $\nu_{pp'} \in (0.1, 5)$ and the range parameters $1/a_{pp'} \in (\min_{i,j,j'}\{\delta_{ijj'}\}, \max_{i,j,j'}\{\delta_{ijj'}\})$ so the correlation range will be within the minimum and maximum distances between

two units.

For estimation of NS, the total and between covariance quantities are $K^{\boldsymbol{R}}_{\text{Total},pp'}(t,t') = K^{W}_{pp'}(t,t') + \tau_p^2 \text{I}(t = t', p = p')$ and $K^{\boldsymbol{Y}}_{\text{Between},pp'}(t,t') = \text{Cov}\{Y^p_{ij}(t,s_{ij}), Y^{p'}_{ij'}(t',s_{ij'})\}$ defined for $j \neq j'$ so that $K^{\boldsymbol{Y}}_{\text{Between},pp'}(t,t') = K^{Z}_{pp'}(t,t')$. This leads to straightforward modifications to the estimators in (2.8) and (2.9), while (2.10) remains the same. For the level 1 process, $\widehat{K}^{Z}_{pp'}(t,t') = \{\text{NM}(\text{M}-1)\}^{-1} \sum_{i,j} \sum_{j' \neq j} \{Y^p_{ij}(t,s_{ij}) - \bar{Y}^p_{G(i)}(t)\}\{Y^{p'}_{ij'}(t',s_{ij'}) - \bar{Y}^{p'}_{G(i)}(t')\}$. For the level 2 process, $\widehat{K}^{W}_{pp',\text{Inflated}}(t,t') = (\text{NM})^{-1} \sum_{i,j} \{R^p_{ij}(t,s_{ij}) R^{p'}_{ij}(t',s_{ij})\}$. The model assumptions and estimation procedure for UNIV are straightforward simplifications to the method previously presented for FULL.

2.4.3 Results

Methods are compared in terms of estimation accuracy of the diet mean functions $\mu^p_d(t)$. To assess model fit for response $p = 1,2$ we compute MC estimates (averaged over diet) of the mean integrated squared error: $\text{MISE} = \int_t \text{E}\{\widehat{\mu}^p_d(t) - \mu^p_d(t)\}^2 dt$; the integrated bias: $\text{Int Bias} = \int_t \text{E}\{\widehat{\mu}^p_d(t) - \mu^p_d(t)\} dt$; the integrated squared bias: $\text{Int Sq Bias} = \int_t [\text{E}\{\widehat{\mu}^p_d(t) - \mu^p_d(t)\}]^2 dt$; and the integrated variance: $\text{Int Var} = \int_t [\text{E}\{\widehat{\mu}^p_d(t) - \text{E}\{\widehat{\mu}^p_d(t)\}]^2 dt$. We construct pointwise confidence intervals $\widehat{\mu}^p_{d,r}(t) \pm l^p_{d,r}(t)$ with margin of error $l^p_{d,r}(t) = 1.645\sqrt{\text{Var}\{\widehat{\mu}^p_{d,r}(t)\}}$ and length $2l^p_{d,r}(t)$ for each MC replication $r = 1,\ldots,100$ and then average over diet d, subunit t and replication r. $\text{Var}\{\widehat{\mu}^p_{d,r}(t)\} = \boldsymbol{x}^p_{d,r}(t)^{\text{T}} \text{Cov}(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) \boldsymbol{x}^p_{d,r}(t)$ where the form of $\text{Cov}(\widehat{\boldsymbol{\beta}}_{\text{GLS}})$ is given in Section 2.3.4, and $\boldsymbol{x}^p_{d,r}(t)$ is the vector corresponding to $\boldsymbol{\beta}$ that indicates the diet d and response p.

When the data are generated from FULL, it holds for all four scenarios that fitting FULL outperforms the estimation from both UNIV and NS (Table 2.2). Only Scenarios 1 and 2 are presented since the outcomes for the other two scenarios are very similar to Scenario 1. Results from Scenarios 3 & 4 can be found in Appendix A.4, along with figures showing that the spatial correlation functions and level 1 & 2 eigenfunctions are estimated well. As Table 2.2 shows, NS underestimates the overall variability in the model by not accounting for the spatial cross covariance, leading to smaller confidence interval lengths and improper coverage. The coverage of UNIV is nominal, although the confidence interval lengths required to achieve this are longer than

those of either FULL or TRUE. Though statistically different, the MISE of FULL is close to that of TRUE, whereas the MISE of FULL is statistically smaller than that of either UNIV or NS. All methods have little or no bias, though for response 2 the variance increases wildly with UNIV, as seen in Figure 2.1. Both the bivariate and spatial features of the generating model are important enough that FULL is the only method that adequately accounts for them and is preferred over NS and UNIV in this setting.

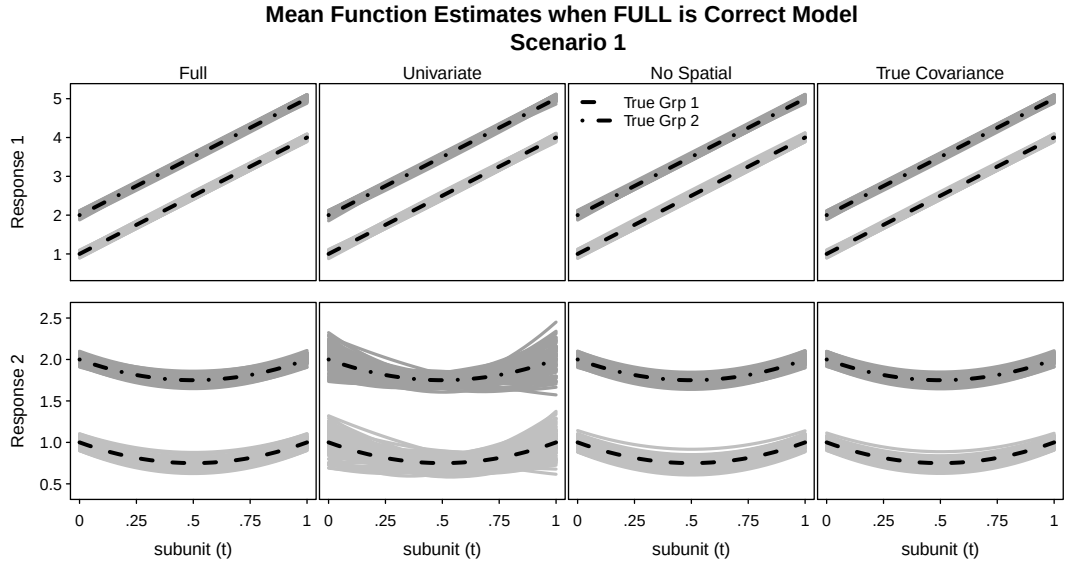

Figure 2.1: Group mean functions when FULL is the generating model with $\rho = 0.8$ and $N = 50$ (Scenario 1). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.

Table 2.2 shows that for Scenario 2, UNIV has statistically smaller MISE than FULL for response 1, which is the only major difference between this case and the other three. The difference in MISE is very small and unlikely to be practically important. We attribute this finding to the small sample size ($N = 10$). Regardless, the preferred method overall for Scenario 2 is FULL since it has smaller MISE than NS for both responses and does a much better job of estimating diet mean functions than UNIV for response 2.

Table 2.2: Mean function estimation comparisons for Scenarios 1 & 2 when FULL is generating model

	90% Coverage	C.I. Length	MISE		Int Bias	Int Sq Bias	Int Var
Scenario 1: $\rho = 0.8$ and N $= 50$							
Response 1							
FULL	87.8	13.7	0.179		0.526	0.007	0.172
UNIV	89.3	14.3	0.184	$*$	0.540	0.008	0.176
NS	82.9	11.9	0.202	$*$	0.352	0.004	0.198
TRUE	89.2	13.7	0.176	$**$	0.543	0.007	0.168
Response 2							
FULL	88.5	13.7	0.190		0.453	0.007	0.184
UNIV	89.3	29.7	0.922	$*$	0.437	0.011	0.910
NS	79.5	11.9	0.228	$*$	0.328	0.003	0.225
TRUE	90.9	13.7	0.175	$**$	0.398	0.005	0.170
Scenario 2: $\rho = 0.8$ and N $= 10$							
Response 1							
FULL	82.1	30.8	1.230		-0.160	0.011	1.219
UNIV	85.3	32.5	1.197	$**$	-0.041	0.005	1.192
NS	67.5	23.9	1.441	$*$	-0.235	0.012	1.429
TRUE	86.8	30.6	0.981	$**$	-0.275	0.007	0.974
Response 2							
FULL	85.9	30.4	1.155		0.321	0.008	1.148
UNIV	82.4	60.4	5.587	$*$	-0.472	0.127	5.461
NS	74.8	23.6	1.221	$*$	0.219	0.005	1.216
TRUE	87.5	30.6	0.865	$**$	-0.339	0.005	0.860

Results in hundredths. A '$**$' ('$*$') indicates better (worse) MISE compared to FULL by Wilcoxson rank sum test, $\alpha = 0.05$.

Scenarios 5 and 6 in Appendix A.2 address the robustness of FULL to model misspecification. Briefly, our simulation showed that overfitting FULL when the true model is less complex (NS) still results in very small MISE, small bias, and proper coverage. In Scenario 6, the generating model is more complex than the proposed model because it includes interactions between functions and spatial random effects. In this case FULL has larger MISE than TRUE, but maintains proper coverage.

2.5 Colon carcinogenesis study; joint analysis of apoptosis and p27

We now consider the colon carcinogenesis experiment that was designed to investigate how fish and corn oil diets affect colon carcinogenesis. To facilitate further understanding, Figure 2.2 provides a depiction of the data structure.

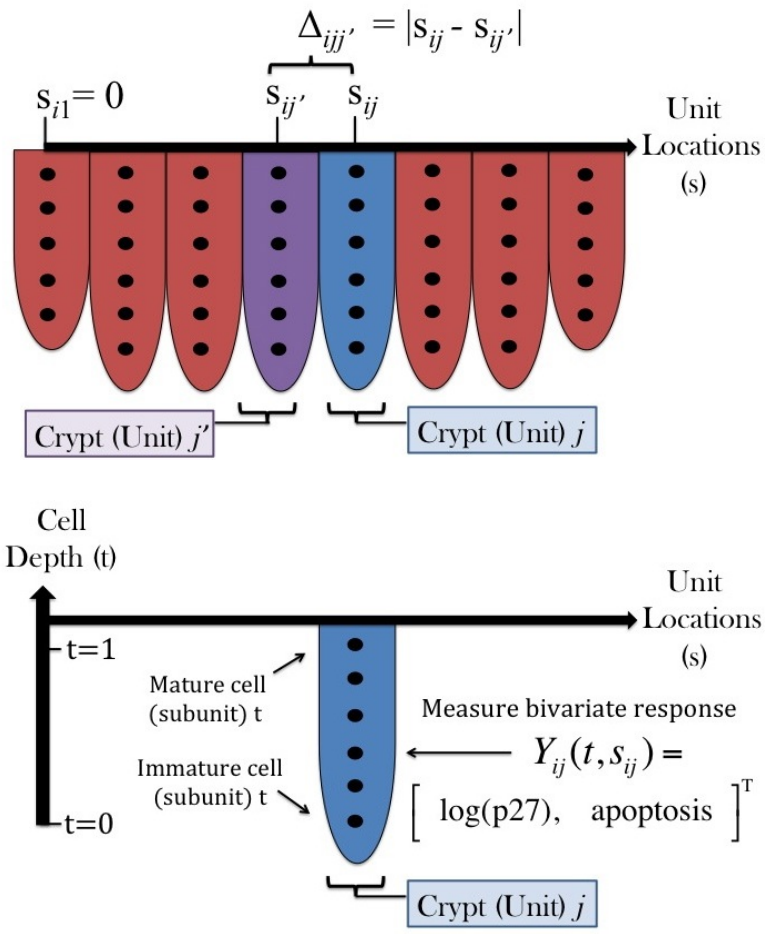

Figure 2.2: Depiction of the data structure from the motivating colon carcinogenesis application.

In this experiment, a total of N = 12 rats were divided into D = 4 diet groups defined by the combination of two types of oils (fish and corn) with two levels of

supplementation (the addition of butyrate or not). 24 hours after exposure to a colon carcinogen, a slice of each rat's colon was removed that contained approximately 20 fingerlike structures called colonic crypts. Within each crypt, multiple responses were measured at each of 18-37 cell depths. Here we focus on apoptosis (response 1), or programmed cell death (0/1 valued), and a (continuous) cell cycle inhibitor protein called p27 which is log-transformed for stability (response 2).

Since the maturity of a cell depends on its relative depth within a crypt, the data for a crypt can be modeled using functional data methods (Baladandayuthapani et al. 2008; Morris and Carroll 2006; Morris et al. 2003, 2002, 2001; Staicu et al. 2010). Stem cells which generate new (daughter) cells are located at the bottom of the crypt, and the daughter cells mature as they travel upward toward the intestinal lining where they are released. Furthermore, the spatial alignment of crypts enables us to investigate what has been called *crypt signaling* (Baladandayuthapani et al. 2008), that is, the effect that responses in one crypt have on the responses in neighboring crypts. Our goals in this analysis are to perform inference on the diet mean functions for the two responses and to investigate the complex correlation structures that exist within and between the two responses, including spatial correlation among crypts at varying distances and correlation between the responses at different depths within a crypt.

In terms of notation, let $\boldsymbol{Y}_{ij}(t, s_{ij}) = [Y^1_{ij}(t, s_{ij}), Y^2_{ij}(t, s_{ij})]^{\mathrm{T}}$ where $Y^1_{ij}(t, s_{ij})$ is log(p27)and $Y^2_{ij}(t, s_{ij})$ is apoptosis (jittered to mimic a continuous random variable). Due to the infrequency of apoptotic events in the second part of the crypt, that is, from mid-crypt to the lumen, we focus on only the first half of the crypt. The bivariate response is measured at cell depth $t \in \mathcal{T}$ within crypt $j = 1, \ldots, \mathrm{M} = 20$ at relative spatial location s_{ij} within rat (subject) $i = 1, \ldots, \mathrm{N} = 12$ of diet group $d = 1, \ldots, \mathrm{D} = 4$. For a crypt with n total cells, the relative position of cell k is assigned the value $t = (k-1)/(\mathrm{n}-1)$ so that the first cell measured (at the tip) is at position 0 and the last cell (closest to the intestinal lining) is at position 1.

In order to implement the methods presented for a balanced design, we assign each observation a new cell depth value t within the set $\{t_1 = 0, \ldots, t_{37} = 1\}$ based on its proximity to the original relative cell position $(k-1)/(\mathrm{n}-1)$. Note that a grid of 37 subunits is chosen because 37 is the maximum number of cells observed in a crypt. We bin the observations within a crypt so that all crypts have the same

number of observations, which results in $L = 10$ subunits per crypt when restricting to $\mathcal{T} = [0, 0.54]$. The value 0.54 was chosen because it is the largest value such that all bins have non-zero sample proportions of observed apoptosis when averaging over subjects and within diet group.

Prior to analysis, we log-transform p27 and jitter the apoptosis so that we can apply the methodology we have developed for jointly modeling two continuous functional responses. For simplicity, let $Y_{ijt} = Y_{ij}^2(t, s_{ij})$ be the observed apoptotic response, and let Y_{ijt}^* be the jittered response that acts as a continuous random variable. We jitter the binary data in the following way that matches the first two moments for a Bernoulli random variable: $Y_{ijt}^* = \hat{\pi}_{G(i)}(t)/2 + Y_{ijt}/2 + \epsilon_{ij}(t)$ for $\epsilon_{ij}(t) \overset{\text{indep}}{\sim} N(0, \frac{3}{4}[\hat{\pi}_{G(i)}(t)\{1 - \hat{\pi}_{G(i)}(t)\}])$ where $\hat{\pi}_{G(i)}(t)$ is the sample mean of the non-missing apoptotic responses at cell position t within diet group $G(i)$. With this, $E\{Y_{ijt}^*\} = E\{\hat{\pi}_{G(i)}(t)/2\} + E\{Y_{ijt}/2\} = \pi_{G(i)}(t)$ and $\text{Var}\{Y_{ijt}^*\} \approx \text{Var}\{Y_{ijt}/2\} + \frac{3}{4}[\hat{\pi}_{G(i)}(t)\{1 - \hat{\pi}_{G(i)}(t)\}] = \pi_{G(i)}(t)\{1 - \pi_{G(i)}(t)\}$.

The tuning parameters from the methods in Sections 2.3.1 and 2.3.3 are as follows: Δ^* is set to be 1 millimeter (1000 microns), the same correlation range used in the analysis done by Staicu et al. (2010) based on scientific information that the spatial correlation is practically zero at distances larger than Δ^*; the tolerance $\epsilon = .08$ is chosen so that the median number of triplets entering the nearest neighbor sets is 75, with a minimum of 51 and a maximum of 144; as in the simulations we use an equally spaced grid s_{grid} of $m = 50$ points for spatial estimation (see Section 2.3.1); finally, we set the cumulative explained variance threshold $P_1 = 0.95$ and an individual explained variance threshold $P_2 = 1/(2L)$ for $L = 10$ (see Section 2.3.3). Staicu et al. (2010) and Baladandayuthapani et al. (2008) have shown that it is reasonable to assume that the diet means for $\log(\text{p27})$ have a quadratic form. Through examination of sample proportions, we concluded that a quadratic form, though somewhat simplified, seems appropriate for the jittered version of this response as well, and we specify quadratic equations for each response in GLS estimation. For comparisons of predictive performance to FULL, we also analyze the data using the UNIV and NS models.

Since the Matérn smoothness and range parameters are difficult to estimate simultaneously, in practice one parameter is typically fixed while the other is estimated. Therefore, in contrast to the simulations in which we optimized the Matérn smooth-

ness and range parameters together, for the data analysis presented we estimate them separately in the marginal case by optimizing over a grid for the smoothness parameter, $\nu_{pp} \in \{0.5, 1, \ldots, 10\}$. Estimation of spatial cross parameters is still done simultaneously.

2.5.1 Results from FULL Model

A primary goal of jointly modeling log(p27) and apoptosis is to quantify the dependence between the two responses. The first indication of dependence is in the spatial cross correlation, estimated to be $\widehat{\rho}_{12} = .994$, although its inferential value is questionable because the spatial variance of apoptosis is estimated to be very close to zero ($\widehat{\sigma}_{22} = 0.0004$). The estimated spatial variance for log(p27) is $\widehat{\sigma}_{11} = 0.0047$, which is similar to the estimate obtained via univariate analysis by Staicu et al. (2010). The estimated Matérn parameters of $\hat{\nu}_{11} = 8.43$ and $\hat{a}_{11} = 10$ indicate that the auto-correlation for log(p27) decays to less than 0.05 for crypts more than 1.33 millimeters apart. According to our analysis, the spatial correlation is not as important for apoptosis or for the cross-dependence between responses as it is for log(p27). Error variance estimates are $\widehat{\tau}_1^2 = 0.002$ and $\widehat{\tau}_2^2 = 0.035$. Plots of the level 1 and level 2 eigenfunctions and their corresponding percentages of variation explained can be found in Appendix A.1.

To examine functional cross-dependence, Figure 2.3 shows the marginal (within a crypt) auto- and cross-correlation matrices found through the corresponding bivariate covariance matrix that is formed by summing the spatial variation, level 1 and level 2 covariances, and the error variances on the diagonals. The auto-correlation for log(p27) is very high and positive, only assuming values above 0.93. The auto-correlation for apoptosis seems to decrease with the crypt cells distance; the results seem to show that apoptosis around the quarter-crypt mark is negatively correlated with apoptosis at the tip of the crypt. Our analysis shows indication of negative weak cross-correlation between log(p27) and apoptosis in the bottom of the crypt. Even with this cross-dependence, we found no evidence of predictive performance gain of our model compared to the UNIV and NS models (for more details, see Appendix A.1.1). We believe the reason for the similar predictive performance of FULL to UNIV could be caused by the need to implement a jittering method to treat the binary response as continuous.

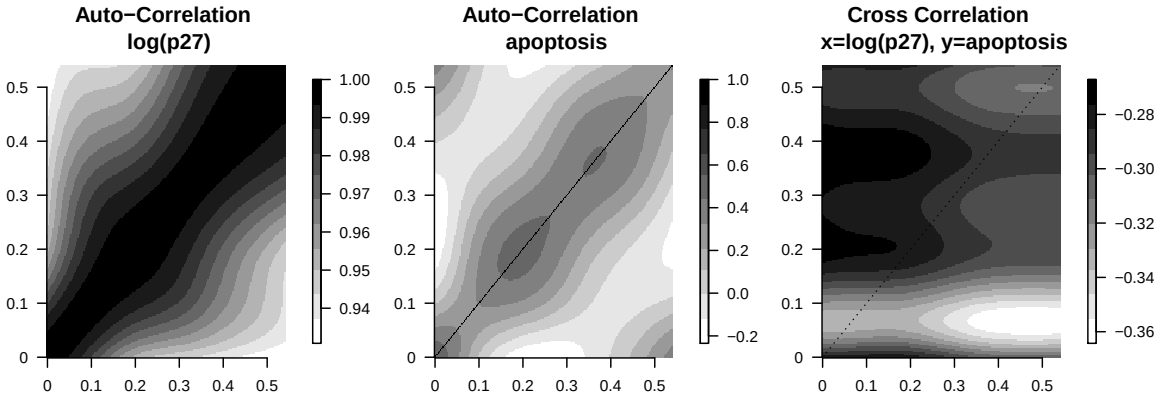

Figure 2.3: Image of marginal auto- and cross-correlation matrices for a crypt. Lines have been added to the diagonals.

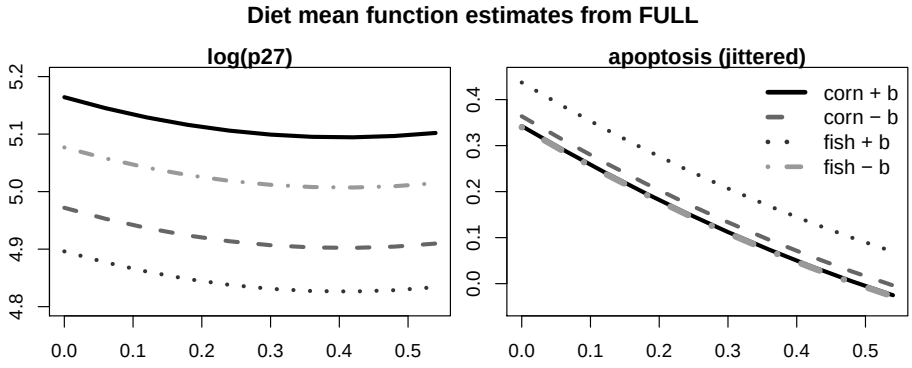

Figure 2.4: Diet mean estimates using FULL.

Another goal of the data analysis is to make inference on the diet mean functions, and estimates are given in Figure 2.4. (Pairwise diet comparisons with 95% confidence intervals based on the variances from the GLS parameter estimation are given in Appendix A.1.) For this data application, Baladandayuthapani et al. (2008) and Staicu et al. (2010) interpret higher levels of p27 to be associated with poor prognoses. Our findings align with previous literature in that rats fed the fish diet with

butyrate supplement have lower levels of $\log(\text{p27})$ than rats who were fed the corn diet with butyrate supplement. Due to the multivariate nature of our approach we also examine the diet means for apoptosis. Our analysis shows that rats fed the fish diet with butyrate supplement have higher levels of apoptosis than rats fed any other diet. This is particularly important because an increase in apoptosis is known to have a protective effect against colon cancer in all phases of tumor development (Hong et al. 2002).

CHAPTER 3

Modeling Multivariate Mixed-Response Functional Data

3.1 Background

The methodology of Chapter 2 is only applicable to a multivariate functional response in which each individual functional response is real-valued. In the analysis of the colon carcinogenesis study of Section 2.5, the binary response required pre-processing to mimic a continuous response due to the lack of available methods for handling bivariate functional responses where one response is continuous and the other is binary. We address that need in this chapter by proposing a model for multivariate functional responses of mixed type.

Until recently, the primary focus of methods employing functional principal components analysis (FPCA) has been on real-valued functional responses. Methods that can model non-Gaussian functional responses, such as repeatedly observed binary or count data, are only recently appearing for univariate functional responses (for example: Hall et al. (2008), van der Linde (2009), Serban et al. (2013)). Additionally, methods that extend functional modeling from the univariate case (i.e. one response curve) to the multivariate case (i.e. a vector of multiple response curves) are currently undergoing development (for example: Zhou et al. (2008), Berrendero et al. (2011), Jacques and Preda (2014)). These multivariate functional methods are limited

in that all curves comprising the multivariate response vector must be real-valued.

Here we propose a Bayesian multivariate functional model that utilizes a multivariate latent Gaussian process and can handle responses of different types, e.g. binary and continuous data. Our method easily incorporates covariates, a feature previously unavailable for modeling non-Gaussian functional responses. As an extension of the methods of Hall et al. (2008), we propose a way to estimate the multivariate latent covariance, in particular, the cross-covariance of latent functions corresponding to different responses. By using a reliable estimate of the multivariate latent covariance, our proposed method can implement multivariate FPCA to specify basis expansions and simplify computation.

Several approaches to modeling non-Gaussian univariate functional responses have appeared in the literature. For binary or count data observed repeatedly, Hall et al. (2008) proposed a non-parametric functional approach in which the observed responses are directly related to a latent Gaussian functional process through a link function. In order to implement FPCA, they used a Taylor series approximation to derive estimators of the latent process mean function and covariance operator and used bootstrapping methods for further inference. A similar approach by Serban et al. (2013) used logistic functional regression to model multilevel cross-dependent binary-valued functional data. In the case of non-rare events, their approach is an extension of the linear approximation methods of Hall et al. (2008) to multilevel data. For rare events, they introduced an approach centered around an exponential approximation.

In contrast to the aforementioned frequentist methods, van der Linde (2009) offered a Bayesian approach to FPCA for repeatedly observed binary or count data. They extended the variational algorithm for Gaussian responses given in van der Linde (2008), and focused on canonical links for one-parameter exponential families. The methods of Hall et al. (2008), Serban et al. (2013) and van der Linde (2009) offer ways to model univariate functional responses, whereas the approach we propose in this paper jointly models multivariate functional responses of mixed type.

To date, the literature concerning multivariate FPCA has been sparse. Ramsay and Silverman (2005) gave a brief example that uses FPCA for a bivariate functional response of hip and knee angle measurements for gait data. After assigning the two functional responses to a fine grid of points, they concatenated the two response func-

tions and proceeded with PCA in the traditional multivariate framework. Berrendero et al. (2011) proposed multivariate FPCA in which the principal components are smooth functions, a result of performing FPCA at each observed location in a domain on which curves have been smoothed. In contrast to the approach of Ramsay and Silverman (2005), Jacques and Preda (2014) presented a method that allowed for non-orthonormal bases which made it possible for each curve in the multivariate response vector to have its own basis expansion. Their approach neatly addresses how to handle responses with differing magnitudes of variation within the curves.

To our knowledge, our method that models multivariate mixed-type responses is the first of its kind within the functional data analysis literature. In the spatial literature, Reich and Bandyopadhyay (2010) developed a spatial latent factor model for multivariate mixed-response data with informative missingness. Our approach shares several similarities to that of Reich and Bandyopadhyay (2010), however our approach is able to examine complex correlation structures that their stationary spatial method is not equipped to handle.

3.2 Model

3.2.1 General Framework

We present the following methodology to jointly model P functional responses. Denote $Y_{pi}(t)$ as the observed functional response of type $p = 1, \ldots, \mathrm{P}$ for subject $i = 1, \ldots, \mathrm{N}$ at location $t \in \mathcal{T}$. In this chapter we make a change in notation for response type, from superscript ($Y_{ij}^{p}(t, s_{ij})$) in Chapter 2 to subscript here. We found that the superscript made notation a bit cleaner in Chapter 2 where the setting was that of a multilevel structure with functional and spatial arguments. Here, where the setting allows for less complicated notation, we revert to more conventional notation and use a subscript p to denote response type.

The responses $Y_{pi}(t)$ are observed only at a finite set of L_{pi} locations $t_{pi1}, t_{pi2}, \ldots, t_{pi\mathrm{L}_{pi}}$, which may be different for subject and response type. To combine responses with different supports, e.g., binary and continuous, let $Y_{pi}(t) = h_p\{W_{pi}(t)\}$ for link function $h_p(\cdot)$ and latent response $W_{pi}(t)$. Motivated by the periodontal application in Section 3.5, we restrict our attention to Gaussian and binary responses. If response p is Gaussian then we use the identity link $h_p(\eta) = \eta$; if re-

sponse p is binary, then we use the indicator link $h_p(\eta) = \mathrm{I}(\eta > 0)$.

Dependence between responses is modeled via the latent Gaussian processes

$$W_{pi}(t) = Z_{pi}(t) + \epsilon_{pi}(t) \tag{3.1}$$

where $\epsilon_{pi}(t) \overset{iid}{\sim} \mathrm{N}(0, \tau_p^2)$ is random noise and $Z_{pi}(t)$ is a random process. For identification purposes, we fix $\tau_p = 1$ for binary responses. Furthermore, let $Z_{pi}(t) = \mu_{pi}(t) + f_{pi}(t)$, the sum of a fixed mean function $\mu_{pi}(t)$ and a smooth subject-specific process $f_{pi}(t)$, assumed to be uncorrelated with $\epsilon_{pi}(t)$.

The mean can be modeled as $\mu_{pi}(t) = \sum_{j=1}^{\mathrm{m}_p} x_{pij}(t)\beta_{1pj} + s_p(t)$ so that it can incorporate m_p covariates $x_{pij}(t)$ with fixed coefficients β_{1pj} and a population-level smooth function $s_p(t)$. It is possible for a subject-specific covariate to depend on the functional location t, for example the indicator of jaw in the periodontal data of Section 3.5, and it is also possible for the same covariates to affect all responses. The smooth function $s_p(t)$ is assumed to be square integrable on $\mathcal{L}^2[0,1]$. We use a predetermined basis expansion to approximate $s_p(t)$. Let $\{\mathrm{B}_{pj}(t) : 1 \leq j \leq \mathrm{n}_p\}$ be a basis expansion in $\mathcal{L}^2[0,1]$ of dimension n_p. We approximate the smooth part by $s_p(t) = \sum_{j=1}^{\mathrm{n}_p} \mathrm{B}_{pj}(t)\beta_{2pj}$ where the type of basis expansions are allowed to differ across response p. To simplify notation, we write $\mu_{pi}(t) = \boldsymbol{u}_{pi}^{\mathrm{T}}(t)\boldsymbol{\beta}_p$ where $\boldsymbol{u}_{pi}(t) = [x_{pi1}(t), \ldots, x_{pi\mathrm{m}_p}(t), \mathrm{B}_{p1}(t), \ldots, \mathrm{B}_{p\mathrm{n}_p}(t)]^{\mathrm{T}}$ is a vector of length $\mathrm{J}_p = \mathrm{m}_p + \mathrm{n}_p$ that combines the covariates and basis functions and has corresponding coefficient vector $\boldsymbol{\beta}_p = [\beta_{1p1}, \ldots, \beta_{1p\mathrm{m}_p}, \beta_{2p1}, \ldots, \beta_{2p\mathrm{n}_p}]^{\mathrm{T}}$.

Let $\boldsymbol{f}_i(t) = [f_{1i}(t), \ldots, f_{\mathrm{P}i}(t)]^{\mathrm{T}}$ be the vector of random subject-specific deviation functions and assume $\boldsymbol{f}_i(t)$ are i.i.d. mean-zero Gaussian processes where $\mathrm{Cov}\{\boldsymbol{f}_i(t), \boldsymbol{f}_i(t')\} = \boldsymbol{K}(t,t')$ and $K_{pp'}(t,t') = \mathrm{Cov}\{f_{pi}(t), f_{p'i}(t')\}$ form the elements of $\boldsymbol{K}(t,t')$. The covariance operator $K_{pp'}(t,t')$ captures both dependence within a response over location t and the cross-dependence between two different latent responses. We assume that $f_{pi}(t)$ is a smooth process in $\mathcal{L}^2[0,1]$ and present two ways of specifying basis expansions for $f_{pi}(t)$: Section 3.2.2 details how to use predetermined bases and Section 3.2.3 gives a data-driven approach that uses multivariate FPCA.

We can write the multivariate model succinctly in matrix form. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{\mathrm{P}}^{\mathrm{T}}]$ be the fixed effect vector of length $\mathrm{J} = \sum_{p=1}^{\mathrm{P}} \mathrm{J}_p$ with corresponding $\mathrm{P} \times \mathrm{J}$

matrix $\boldsymbol{U}_i(t)$ comprised of appropriate evaluations of $\boldsymbol{u}_{pi}(t)$. Let $\boldsymbol{\epsilon}_i(t) \overset{iid}{\sim} N(0, \boldsymbol{D})$ where $\boldsymbol{D}$ is diagonal with elements $\tau_1^2, \ldots, \tau_{\mathrm{P}}^2$. Then (3.1) becomes

$$\boldsymbol{W}_i(t) = \boldsymbol{U}_i(t)\boldsymbol{\beta} + \boldsymbol{f}_i(t) + \boldsymbol{\epsilon}_i(t). \tag{3.2}$$

3.2.2 Predetermined bases

The first way in which we specify basis expansions for $f_{pi}(t)$ is by choosing predetermined bases such as B-spline, Fourier, or polynomial bases. Let

$$f_{pi}(t) = \sum_{k=1}^{\mathrm{M}_p} \psi_{pk}(t)\alpha_{pik} \tag{3.3}$$

where $\{\psi_{pk}(t) : 1 \leq k \leq \mathrm{M}_p\}$ is a basis expansion in $\mathcal{L}^2[0,1]$ of dimension M_p and $\boldsymbol{\alpha}_{pi} = [\alpha_{pi1}, \ldots, \alpha_{pi\mathrm{M}_p}]^{\mathrm{T}}$ are random coefficients with $\mathrm{E}(\alpha_{pik}) = 0$ and $\mathrm{Cov}(\alpha_{pik}, \alpha_{p'i\ell}) = \xi_{k\ell pp'}$. The multivariate covariance function induced by (3.3) is

$$K_{pp'}(t,t') = \mathrm{Cov}\{f_{pi}(t), f_{p'i}(t')\} = \sum_{k=1}^{\mathrm{M}_p}\sum_{\ell=1}^{\mathrm{M}_{p'}} \psi_{pk}(t)\psi_{p'\ell}(t')\xi_{k\ell pp'}, \tag{3.4}$$

which is a function of both the basis functions and covariance $\boldsymbol{\Sigma} = \{\xi_{k\ell pp'}\}$. Using predetermined basis expansions is extremely flexible; in Appendix B.1, we discuss how the covariance model can approximate the covariance matrix of any arbitrary finite-dimensional distribution. The choice of M_p is important in that one needs to select a number of basis functions that is sufficient to approximate the covariance well but is not unnecessarily large. We suggest choosing M_p based on a grid search, using criteria such as DIC for comparison.

3.2.3 Data-driven bases

As an alternative to using predetermined bases, we introduce a novel approach in which we use estimated basis functions that are obtained through FPCA of the multivariate latent covariance. We propose FPCA for multivariate mixed-responses, inspired by Hall et al. (2008) who introduced FPCA for binary-valued functional responses. We too require that the probability of observing a binary event is suffi-

ciently far from zero or one. For simplicity of presentation, we ignore the covariates and discuss how to account for them later in this section.

Recall from (3.1) that we model the p^{th} response as $Y_{pi}(t) = h_p\{W_{pi}(t)\}$ through the latent Gaussian process $W_{pi}(t) = Z_{pi}(t) + \epsilon_{pi}(t)$ and link function $h_p(\eta)$. Linking the latent response directly to the observed response is equivalent to assuming there is a corresponding monotone link function $g_p(\cdot)$ such that $\mathrm{E}\{Y_{ip}(t)|Z_{pi}(t)\} = g_p\{Z_{pi}(t)\}$; we focus on g_p here. Following Hall et al. (2008), assume that $g_p(\cdot)$ has bounded fourth derivative and that the latent process satisfies $Z_{pi}(t) = \mu_p(t) + \delta X_{pi}(t)$ for fixed mean $\mu_p(t)$, unknown small constant $\delta > 0$, and mean-zero Gaussian random variable $X_{pi}(t)$ that is i.i.d. across subjects i and has both finite variance and finite covariance between $X_{pi}(t)$ and $X_{p'i}(t')$. Our goal is to approximate the latent covariance matrix of $Z_{pi}(t)$ whose covariance operator is $K_{pp'}(t,t') = \mathrm{Cov}\{Z_{pi}(t), Z_{p'i}(t')\}$. Without loss of generality, we restrict our attention to one continuous Gaussian response ($p = 1$) and one binary response ($p = 2$) with link functions $g_1(\eta) = \eta$ and $g_2(\eta) = \Phi(\eta)$ where $\Phi(\cdot)$ is the standard normal cdf function. For simplicity, we use g to denote g_2 in the following exposition.

The covariance consists of variance components K_{pp} and cross-covariance components $K_{pp'}$. The variance components K_{11} and K_{22} are estimated using the common FPCA for continuous responses Ramsay and Silverman (2002, 2005) as well as binary-valued responses (Hall et al. 2008), respectively. In particular, when the responses are binary valued, the variance K_{22} is estimated using

$$\widetilde{K}_{22}(t,t') = \{\hat{S}_{22}(t,t') - \hat{\eta}_2(t)\hat{\eta}_2(t')\} / [g^{(1)}\{\hat{\mu}_2(t)\} g^{(1)}\{\hat{\mu}_2(t')\}]. \tag{3.5}$$

where $g^{(1)}$ indicates the first derivative of g. The latent mean estimator is $\hat{\mu}_p(t) = g^{-1}\{\hat{\eta}_p(t)\}$ where $\hat{\eta}_p(t)$ estimates $\mathrm{E}[g\{Z_{pi}(t)\}] = \eta_p(t)$ and is found by smoothing the data $\big(t, Y_{pi}(t)\big)$ for $i = 1, \ldots, \mathrm{N}$. $\hat{S}_{22}(t,t')$ is the estimator for $S_{22}(t,t') = \mathrm{E}\{Y_{2i}(t)Y_{2i}(t')\} = \mathrm{E}\big[g\{Z_{2i}(t)\} g\{Z_{2i}(t')\}\big]$ and is obtained through bivariate smoothing of the data $\big((t,t'), Y_{2i}(t)Y_{2i}(t')\big)$ for $i = 1, \ldots, \mathrm{N}$, removing the diagonals before smoothing.

For the cross covariance operator K_{12} we remark that

$$K_{12}(t,t') = \mathrm{Cov}\big\{Y_{1i}(t), Y_{2i}(t')\big\} / g^{(1)}\{\mu_2(t')\}, \tag{3.6}$$

which is obtained by approximating $\text{Cov}\{Y_{1i}(t), Y_{2i}(t')\} = \text{Cov}[Z_{1i}(t), g\{Z_{2i}(t')\}]$ using a Taylor expansion of $g\{Z_{2i}(t')\}$ around $\mu_2(t')$. More details are given in Appendix B.2. This leads to the estimator of the cross component given by

$$\widetilde{K}_{12}(t,t') = \{\hat{S}_{12}(t,t') - \hat{\eta}_1(t)\hat{\eta}_2(t')\}/g^{(1)}\{\hat{\mu}_2(t')\}. \tag{3.7}$$

Combining the individually smoothed estimators $\widetilde{K}_{11}(t,t')$, $\widetilde{K}_{22}(t,t')$ and $\widetilde{K}_{12}(t,t') = \widetilde{K}_{21}(t',t)$ forms the smooth 2×2 estimator $\widetilde{\boldsymbol{K}}(t,t')$ of the bivariate latent covariance operator. Note that for smoothing purposes in this paper, we implement a global smoother as opposed to the local least squares smoothing of Hall et al. (2008), though either is appropriate. In the presence of subject-specific covariates, one can find covariate estimates using least squares or logistic regression, depending on the type of response, and then use the residuals to estimate the latent covariance.

The final step needed to obtain the basis functions is to implement bivariate FPCA in which we find the eigenfunctions $\widehat{\boldsymbol{e}}(t) = [\hat{e}_1(t), \ldots, \hat{e}_{\text{P}}(t)]^{\text{T}}$ and the eigenvalues $\widehat{\lambda}$ of the matrix $\widetilde{\boldsymbol{K}}(t,t')$. Note that the matrix $\widetilde{\boldsymbol{K}}(t,t')$ is not guaranteed to be positive definite, but we can ensure the truncated spectral decomposition $\widehat{\boldsymbol{K}}(t,t') = \sum_{k=1}^{\text{M}} \widehat{\lambda}_k \widehat{\boldsymbol{e}}_k(t)\{\widehat{\boldsymbol{e}}_k(t')\}^{\text{T}}$ is positive definite by restricting the inclusion of only positive eigenvalues and their eigenfunction counterparts. The truncation value M is chosen based on the proportion of variation explained by the eigenvalues as suggested in Di et al. (2009). In particular, specify a cumulative explained variance threshold P_1 and and individual explained variance threshold P_2. Define $\text{M} = \min\{k : p_{1k} \geq \text{P}_1, p_{2k} < \text{P}_2\}$ where $p_{k1} = \sum_{i=1}^{k} \widehat{\lambda}_i / \sum_{j=1}^{n} \widehat{\lambda}_j$, $p_{k2} = \widehat{\lambda}_k / \sum_{j=1}^{n} \widehat{\lambda}_j$ and the positive eigenvalues are the first $n \geq k$ eigenvalues. We specify the basis functions to be the eigenfunctions scaled by their associated eigenvalues, $\widehat{\boldsymbol{\psi}}_{pk}(t) = \sqrt{\hat{\lambda}_k}\widehat{\boldsymbol{e}}_{pk}(t)$, and the subject-specific deviation function is approximated by $f_{pi}(t) = \sum_{k=1}^{\text{M}} \widehat{\boldsymbol{\psi}}_{pk}(t)\alpha_{ik}$.

Using this data-driven basis approach, the correlation across responses is largely captured by the basis functions from FPCA. Additionally, since each basis function combines information from all responses, the data-driven approach results in one set of basis functions, eliminating the need to have a set of basis functions for each response. These distinctions offer important advantages over the predetermined basis approach. First, having only one set of basis functions in turn reduces the dimensionality of the random-effect covariance matrix $\boldsymbol{\Sigma}$, making it easier to fit. Second, it allows for further simplification since one can now assume that $\boldsymbol{\Sigma}$ is diagonal. This

will offer computational advantages over the predetermined basis method where the burden of capturing the correlation across responses falls entirely on estimating a non-diagonal $\mathbf{\Sigma}$ which can potentially have very large dimension.

One important consideration to make when implementing this data-driven basis function approach is to ensure that the variance of the latent process for the continuous component is on a scale similar to that of the latent process for the binary component. We suggest scaling the continuous process by $Y_{1i}(t)/s$ where s is the overall sample standard deviation of the continuous response without regard to t. Since s is a scalar quantity, it is straightforward to scale prior to implementing the latent covariance, FPCA and MCMC estimation algorithms, rescaling only the final results back to the original scale.

3.2.4 Prior Specification

To complete the Bayesian model, we specify priors for the hyperparameters. The fixed effect parameters $\boldsymbol{\beta}$ are assigned uninformative Gaussian priors. Let the subject random effect $\boldsymbol{\alpha}_i$ have a Gaussian prior with $\text{Cov}(\boldsymbol{\alpha}_i) = \mathbf{\Sigma}$ and assign $\mathbf{\Sigma}$ an Inverse Wishart prior. For the error variances of the continuous processes, let τ_p^2 have an uninformative gamma prior; for identifiability τ_p^2 is fixed at 1 for binary processes. In summary,

$$\begin{aligned}
&\boldsymbol{\beta}|\sigma_\beta^2 \sim \text{N}_\text{J}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_\text{J}) \\
&\boldsymbol{\alpha}_i|\mathbf{\Sigma} \sim \text{N}_\text{M}(\mathbf{0}, \mathbf{\Sigma}) \\
&\mathbf{\Sigma}|q_1, q_2 \sim \text{InvWishart}_\text{M}(\boldsymbol{V} = q_2 \mathbf{I}_\text{M}, \nu = q_1) \\
&\tau_p^2|l, h \sim \text{InvGamma}(l, h)
\end{aligned} \tag{3.8}$$

for hyperparameters σ_b^2, q_1, q_2, ℓ, and h, selected to result in weak priors.

3.3 Computational Details

To facilitate MCMC sampling, we treat the continuous latent processes $W_{pi}(t)$ for binary response as unknown parameters to be updated as part of the sampling as in Albert and Chib (1993). Using this auxiliary variable approach, all parameters have conditional conjugacy due to the prior specifications given in Section 3.2.4, allowing

us to implement Gibbs sampling. The Gibbs sampling algorithm uses full-conditional posteriors derived in the supplementary material and which use notation that we now describe.

Denote the observation locations as $t_{pi\ell}$, $\ell = 1, \ldots, \mathrm{L}_{pi}$, for each subject i and response p, giving a total of $\mathrm{L}_i = \sum_{p=1}^{\mathrm{P}} \mathrm{L}_{pi}$ locations. Let $n = \sum_{i=1}^{\mathrm{N}} \mathrm{L}_i$ be the total number of locations observed across all subjects. Let $\boldsymbol{W}_{pi}$ be the vector of length L_{pi} formed by evaluating $W_{pi}(t)$ at every $t_{pi\ell}$. Furthermore, combine $\boldsymbol{W}_{pi}$ for all responses to form one vector $\boldsymbol{W}_i$ of length L_i; $\boldsymbol{U}_i$ and $\boldsymbol{\Psi}_i$ are defined analogously. Then $\boldsymbol{W}_i$ has mean $\mathrm{E}(\boldsymbol{W}_i|\boldsymbol{\alpha}_i) = \boldsymbol{U}_i\boldsymbol{\beta} + \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i$ and precision matrix $\boldsymbol{P}_i$ is comprised of the appropriate error variance parameter τ_p^{-2}.

MCMC begins by setting initial values for all parameters and then sequentially sampling each parameter conditioned on all the others (denoted by "$|\cdot$"). Sampling is performed (using the latest sample to update each parameter) according to the full conditionals in the following manner:

0. Select initial values for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\Sigma}$, $W_{pi}(t)$ for binary responses, and τ_p^2 for continuous responses;

1. For each $i = 1, \ldots, \mathrm{N}$ and $\ell = 1, \ldots, \mathrm{L}_{pi}$, update the latent response corresponding to the observed binary response by drawing from $W_{pi}(t_{i\ell})|\cdot \sim \mathrm{N}(\boldsymbol{u}_{pi}^{\mathrm{T}}(t_{i\ell})\boldsymbol{\beta} + \boldsymbol{\psi}_p^{\mathrm{T}}(t_{i\ell})\boldsymbol{\alpha}_i, 1)$ restricted to the interval $(0, \infty)$ if $Y_{pi}(t_{i\ell}) = 1$ or $(-\infty, 0)$ if $Y_{pi}(t_{i\ell}) = 0$;

2. Update the population mean parameter by drawing from $\boldsymbol{\beta}|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}})$ for $\boldsymbol{V}_{\boldsymbol{\beta}} = \left[\left(\sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i\right) + \sigma_{\boldsymbol{\beta}}^{-2}\mathbf{I}_{\mathrm{J}}\right]^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}\left[\sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i(\boldsymbol{W}_i - \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i)\right]$;

3. For each $i = 1, \ldots, \mathrm{N}$, update the random effect by sampling from $\boldsymbol{\alpha}_i|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{V}_{\boldsymbol{\alpha}})$ for $\boldsymbol{V}_{\boldsymbol{\alpha}} = \left(\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{\Psi}_i + \boldsymbol{\Sigma}^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \boldsymbol{V}_{\boldsymbol{\alpha}}\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{P}_i(\boldsymbol{W}_i - \boldsymbol{U}_i\boldsymbol{\beta})$;

4. Update the random effect covariance matrix through $\boldsymbol{\Sigma}|\cdot \sim \mathrm{InvWishart}_{\mathrm{M}}[\{\sum_{i=1}^{\mathrm{N}} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^{\mathrm{T}} + (1/q_2)\mathbf{I}_{\mathrm{M}}\}^{-1}, \mathrm{N} + q_1]$;

5. Update the error variance for the continuous responses according to $\tau_p^2|\cdot \sim \mathrm{InvGamma}(l_\omega, h_\omega)$ with $l_\omega = n/2 + l$ and $h_\omega = h + 1/2\sum_{i=1}^{\mathrm{N}}\sum_{\ell=1}^{\mathrm{L}_i}[W_{pi}(t_{i\ell}) - \boldsymbol{u}_{pi}^{\mathrm{T}}(t_{i\ell})\boldsymbol{\beta} + \boldsymbol{\psi}_p^{\mathrm{T}}(t_{i\ell})\boldsymbol{\alpha}_i]^2$.

Steps 1-5 are repeated for the desired number of samples.

3.4 Simulations

3.4.1 Data generation

We consider the case where $Y_{1i}(t)$ is continuous and $Y_{2i}(t)$ is binary. Functions are observed at a dense, balanced design with $\mathrm{L}_{pi} \equiv 30$ equally-spaced locations in $[0, 1]$ for each subject i and response p. We use the model given in (3.2) with predetermined bases as in Section 3.2.2 for data generation. We specify a separable random effect covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{A} \otimes \boldsymbol{C}$, where $\mathrm{Cov}([\alpha_{1ik}, \alpha_{2ik}]^{\mathrm{T}}) = \boldsymbol{A}$ for $\boldsymbol{A}_{11} = \boldsymbol{A}_{22} = 1$ and $\boldsymbol{A}_{12} = \boldsymbol{A}_{21} = \rho_{\boldsymbol{\alpha}}$ so that the parameter $\rho_{\boldsymbol{\alpha}}$ controls the correlation between the latent responses, and $\mathrm{Cov}([\alpha_{pi1}, \ldots, \alpha_{pi\mathrm{M}}]^{\mathrm{T}}) = \boldsymbol{C}$ for $p = 1, 2$ controls the covariance of the random effect basis function coefficients and is the same across responses. The $\boldsymbol{C}$ used for data generation has the AR(1) structure with variance 1 and correlation parameter $\rho = 1/2$.

For the fixed population mean function we assume there are no subject-level covariates so that $\boldsymbol{\mu}(t) = \mathbf{B}(t)\boldsymbol{\beta}$, and we specify a quadratic basis $\{\mathrm{B}_{pj}(t) = t^{(j-1)} : 1 \leq j \leq 3\}$ for each response p with coefficients $\boldsymbol{\beta}_1 = [-0.64, 4, -4]^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = [0.97, -6, 6]^{\mathrm{T}}$. The intercepts are chosen such that the curves are positive for approximately half of the observed locations t. The basis functions for the subject-specific deviation function $\boldsymbol{f}_i(t) = \boldsymbol{\Psi}(t)\boldsymbol{\alpha}_i$ are given by $\psi_{1k}(t) = \sin\{(2\pi k/\mathrm{M})(t + 2\pi k/\mathrm{M})\}$ and $\psi_{2k}(t) = \cos\{(2\pi k/\mathrm{M})(t + 2\pi k/\mathrm{M})\}$ for $k = 1, \ldots, \mathrm{M} = 7$. The error variance for the continuous process is $\tau_1^2 = 1$. We generate data from four scenarios given in Table 3.1 by varying the sample size ($\mathrm{N} = 50, 250$) and the cross-correlation ($\rho_{\boldsymbol{\alpha}} = 0, 0.8$). All scenarios use 100 Monte Carlo (MC) replications.

3.4.2 Models and metrics for comparison

We fit four models to each dataset.

1. Bivariate B-spline (BBSP): the multivariate model in (3.2) with B-spline bases as in Section 3.2.2;

2. Univariate B-spline (UBSP): the model from (3.1) applied separately to each response with B-spline bases as in Section 3.2.2;

3. Bivariate FPCA (BFPCA): the multivariate model in (3.2) with data-driven bases as in Section 3.2.3;

4. Univariate FPCA (UFPCA): the model in (3.1) applied separately to each response with data-driven bases as in Section 3.2.3;

For estimation using the B-spline methods, we choose B-splines of order 4 and the number of B-spline breaks for each replication is fixed at 6 based on preliminary analyses. For the FPCA methods, we specify an unstructured $\mathbf{\Sigma}$, and the number of basis functions is chosen to explain at least $\mathrm{P}_1 = 99\%$ of the cumulative variation. In practice, both the number of basis functions for the B-spline method and the percentage of variation explained for the FPCA method are tuning parameters and one should compare results over a grid parameter values. For the population mean we fit the true polynomial basis $\mathbf{B}(t)$ for estimation. We perform MCMC sampling with 20,000 draws and the first 5,000 are discarded as burn-in. The hyperparameters are specified as $\sigma_b^2 = 100$ and $q_1 = q_2 = l = h = 0.1$.

Methods are compared in terms of their predictive performance and ability to estimate the marginal mean function for each response. Let $\omega_{1i}(t) = \mathrm{E}\{Y_{1i}(t)\} = \boldsymbol{u}_{1i}^{\mathrm{T}}(t)\boldsymbol{\beta}_1$ and $\omega_{2i}(t) = \mathrm{E}\{Y_{2i}(t)\} = \Phi\{\gamma_i(t)\}$, where $\gamma_i(t) = \boldsymbol{u}_{2i}^{\mathrm{T}}(t)\boldsymbol{\beta}_2/\sqrt{v_2(t)}$ is the population effect shrunk toward zero by the square root of the marginal variance $v_2(t) = \mathrm{Var}\{Y_2(t)\} = \boldsymbol{\psi}_2(t)\mathbf{\Sigma}_{22}\{\boldsymbol{\psi}_2(t)\}^{\mathrm{T}} + 1$. Let $\widehat{\omega}_{pr}(t)$ and $\hat{v}_{pr}(t)$ be the posterior mean and variance, respectively, for MC replication $r = 1, \ldots, 100$. Metrics for comparison of estimated means found in Table 3.1 for each response are mean integrated squared error: $\mathrm{MISE} = \int_t \mathrm{E}\{\widehat{\omega}_p(t) - \omega_p(t)\}^2 dt$; coverage of 95% pointwise confidence intervals $\widehat{\omega}_{pr}(t) \pm l_{pr}(t)$ averaged over location t and MC replication r with margin of error $l_{pr}(t) = 1.96\sqrt{\hat{v}_{pr}(t)}$; and confidence interval length $2l_{pr}(t)$.

For prediction, we generate additional data $Y_{prj}(t_l)$ at equally spaced locations $t_\ell \in [0,1]$ where $\ell = 1, \ldots, 30$ for subjects $j = 1, \ldots, 20$ per response $p = 1, 2$ for each MC replication $r = 1, \ldots, 100$. To assess the value of jointly modeling the two responses, we leave out all of response 1 for 10 subjects and all of response 2 for the remaining 10 subjects per replication. Models are compared in terms of their predictive performance using mean squared prediction error (MSPE) for each response, defined as $\mathrm{MSPE} = (nm\mathrm{L})^{-1}\sum_{r=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{\mathrm{L}}\{Y_{prj}(t_\ell) - \hat{Y}_{prj}(t_\ell)\}^2$. For binary responses this is known as the Brier score and $\hat{Y}_{prj}(t_\ell)$ is the posterior probability that $Y = 1$.

3.4.3 Results

Table 3.1: Simulation Results

	Continuous Response				Binary Response			
	MISE	CI length	95 % Cvg	MSPE	MISE	CI length	95 % Cvg	MSPE
Scenario 1: $n = 50$, $\rho_{\alpha} = 0.8$								
BFPCA	3.50	65.3	92.9 ***	319	0.174	12.8	86.9 ***	23.0
BBSP	3.26	61.6 **	90.7 ***	313	0.168	12.9	87.3 ***	22.9
UFPCA	3.20	66.5 *	93.6	350 *	0.182	14.8 *	90.8 ***	24.4 *
UBSP	2.86	64.9	94.0	351 *	0.185	14.5 *	90.9 ***	24.3 *
Scenario 2: $n = 250$, $\rho_{\alpha} = 0.8$								
BFPCA	0.795	31.9	91.4 ***	284	0.039	6.66	90.7 ***	21.0
BBSP	0.798	31.2 **	91.0 ***	285	0.037	6.67	91.3 ***	21.0
UFPCA	0.790	32.8 *	91.9 ***	351 *	0.040	7.27 *	93.2	24.3 *
UBSP	0.794	32.4 *	92.0 ***	350 *	0.043	7.25 *	91.6 ***	24.3 *
Scenario 3: $n = 50$, $\rho_{\alpha} = 0$								
BFPCA	2.96	65.9	94.2	408	0.172	13.4	89.3 ***	26.3
BBSP	3.16	62.6 **	92.8 ***	421 *	0.183	13.3	88.4 ***	26.6 *
UFPCA	2.75	65.8	94.6	372 **	0.166	14.6 *	93.3	24.4 **
UBSP	2.85	63.9 **	93.5	371 **	0.162	14.5 *	92.8 ***	24.2 **
Scenario 4: $n = 250$, $\rho_{\alpha} = 0$								
BFPCA	0.802	32.5	94.6	370	0.044	6.96	91.1 ***	24.8
BBSP	0.791	31.9 **	94.3	374 *	0.042	6.97	90.7 ***	24.9
UFPCA	0.780	32.9 *	94.7	362 **	0.042	7.35 *	93.5	24.3 **
UBSP	0.765	32.5	94.5	361 **	0.040	7.35 *	94.1	24.3 **

Results in hundredths. A '**' ('*') indicates better (worse) compared to BFPCA by Wilcoxson rank sum test, $\alpha = 0.05$. For coverage, a '***' indicates that the coverage is not within the nominal 95% range.

Table 3.1 gives the simulation results. There appears to be little difference in mean function estimation between univariate and bivariate methods for all scenarios. When strong correlation is present (Scenarios 1 & 2), the bivariate methods show marked improvement in prediction for both responses over the univariate methods, a difference that becomes more pronounced with an increase in sample size. Bivariate methods perform well when the generating model is univariate (Scenarios 3 & 4). Though prediction is better when fitting the correct univariate model, the differences between the bivariate and univariate methods become very small with an increase in sample size. All methods show slight under-coverage.

For Scenarios 1 & 2 there is no clear difference between fitting predetermined bases (Section 3.2.2) or data-driven bases (Section 3.2.3); however, BFPCA has better prediction compared to BBSP in Scenarios 3 & 4 when there is no cross-correlation. The univariate models have very similar performance to one another in all scenarios.

3.5 Periodontal Data Application

We demonstrate our methods using data from a periodontal study (Fernandez et al. 2009) conducted by the Center for Oral Health Research at the Medical University of South Carolina. In addition to collecting subject-level covariates for over 200 Gullah African Americans, several measures of patients' periodontal health were observed at six sites for each of 28 teeth. The two responses we consider are (continuous) clinical attachment loss (CAL) and (binary-valued) bleeding on probing (BOP). CAL is the distance that a tooth has detached from the bone, rounded to the nearest mm. We use the average CAL over the six sites on each tooth as the tooth's CAL response. BOP is the binary indicator of whether the gums bleed when pressed with a dental probe at any of the six sites per tooth. A total of $N = 197$ patients (subjects) are included for analysis after excluding those with more than 50% missingness. Any remaining missingness is assumed to be completely at random; Reich and Bandyopadhyay (2010) and Reich et al. (2013) provide methods for accounting for non-random missingness.

For our analysis, we assign teeth the numbers 1-14 going from left to right in the upper jaw when looking at a patient and 15-28 going from right to left in the lower jaw when looking at a patient; wisdom teeth are excluded. Using this numbering system, teeth 1 & 28 are adjacent going from upper jaw to lower jaw, and it is the

same for teeth 14 & 15 on the other side of the mouth. We consider responses at each tooth to be realizations of a functional process with locations $t \in [1, 28]$. In fitting a bivariate functional model to this data, we hope to gain a better understanding of the dynamics between the responses CAL and BOP through close examination of their cross-covariance. Our extremely flexible approach to modeling the covariance will be able to capture any spatial correlation of adjacent teeth, of teeth on different sides of the mouth, and of teeth on different jaws.

The subject-specific covariates that we include in modeling the mean function are the same covariates used by Reich and Bandyopadhyay (2010) and include age (in years), gender (female=1, male=0), body mass index or BMI (in kg/m^2), smoking status (1=smoker, 0=never), and glycosylated hemoglobin or HbA1c (1 = high, 0 = controlled). All covariates have been standardized to be zero-mean with standard deviation of 1. For each tooth, we include an indicator of jaw (0=upper, 1=lower). For the smooth part of the mean, we consider a quadratic function $s_p(t) = \beta_{p0} + \beta_{p1} d(t) + \beta_{p2} d(t)^2$ of tooth distance d from the front of the mouth, where $d(t) = t - 7.5$ for teeth in the upper jaw and $d(t) = t - 21.5$ for the lower jaw.

We present analysis for 8 models given in Table 3.2 that all employ the data-driven basis method of Section 3.2.3. The 8 models differ by: 1) whether FPCA is univariate or bivariate; 2) the choice of threshold $\mathrm{P}_1 = 99\%, 95\%$ for the cumulative percentage of variation explained for FPCA; and 3) whether a random bivariate subject-level intercept $\boldsymbol{\alpha}_{0i} = [\alpha_{01i}, \alpha_{02i}]^{\mathrm{T}}$ is added to model (3.2). Models using B-splines as in Section 3.2.2 were also considered but are not presented because the best-performing models required a large number of basis functions.

For the purpose of estimating the latent covariance, we ignore the covariates. When incorporating a bivariate random subject-level intercept, we use residuals $R_{1i}(t) = Y_{1i}(t) - \mathrm{L}_{1i}^{-1} \sum_{i=1}^{\mathrm{L}_{1i}} Y_{1i}(t)$ of the continuous response CAL to estimate the latent covariance for FPCA; this is not done for the binary responses as the residuals would no longer be binary. For models that include $\boldsymbol{\alpha}_{0i}$, we estimate the covariance term $\mathrm{Cov}(\alpha_{01i}, \alpha_{02i})$ in addition to the variance terms $\mathrm{Var}(\alpha_{0pi})$. We specify a diagonal covariance matrix $\boldsymbol{\Sigma}$ for the remaining random effect parameters.

Table 3.2 shows that Models 1-4, which include a subject random effect, outperform (based on DIC) Models 5-8 which omit the subject random effect. For this data, specifying the larger percentage of variation explained for FPCA, and hence includ-

Table 3.2: Model comparisons for the periodontal data application

Model	Subject RE	PVE	PCA	Dbar	pD	DIC
1	Y	99	B	8114	1257	9371
2	Y	99	U	8228	1231	9459
3	Y	95	B	8849	1001	9850
4	Y	95	U	8536	1114	9650
5	N	99	B	10101	942	11043
6	N	99	U	10586	832	11418
7	N	95	B	10511	788	11299
8	N	95	U	10722	760	11482

"Subject RE" indicates inclusion of a subject-specific random intercept. "PVE" is the threshold for cumulative percentage of variation explained. "PCA" indicates whether univariate ("U") or bivariate ("B") FPCA was performed.

ing more basis functions, leads to better model performance. In comparing the two leading models 1 & 2, implementing FPCA on the full bivariate covariance matrix as in Model 1, taking into account the cross-dependence between the two responses CAL and BOP, leads to superior performance. Figure 3.1 shows the subject-level coefficient estimates and 95% posterior intervals for Model 1. Models 2-4 had similar coefficient estimates. For CAL, only the coefficient interval for BMI includes zero. The other coefficient estimates show an increased level of CAL for older patients, males, smokers, patients with high HbA1c counts, and for teeth on the upper jaw. For BOP, the posterior confidence intervals are larger than those for CAL. For intervals that exclude zero, there is an increase of BOP for the upper jaw, yet a slightly lower incidence of BOP for higher BMI.

Figure 3.2 shows the fitted values (from Model 1) for two individuals in the periodontal data set. The left panels show the posterior means and 95% posterior intervals of the subject-specific mean function $\mu_{1i}(t)$ for the continuous response CAL. Most of the observed CAL values fall within the 95% interval for both subjects, indicating a reasonable model fit. The right panels show the posterior mean and 95% posterior intervals of the conditional probability of the event, $P(Y_{2i}(t) = 1|\boldsymbol{\alpha}_{2i})$. Teeth with observed BOP (= 1) are indicated by the squares on the bottom of the plot. The higher predicted probabilities tend to correspond to the incidence of BOP, again indicating a reasonable model fit.

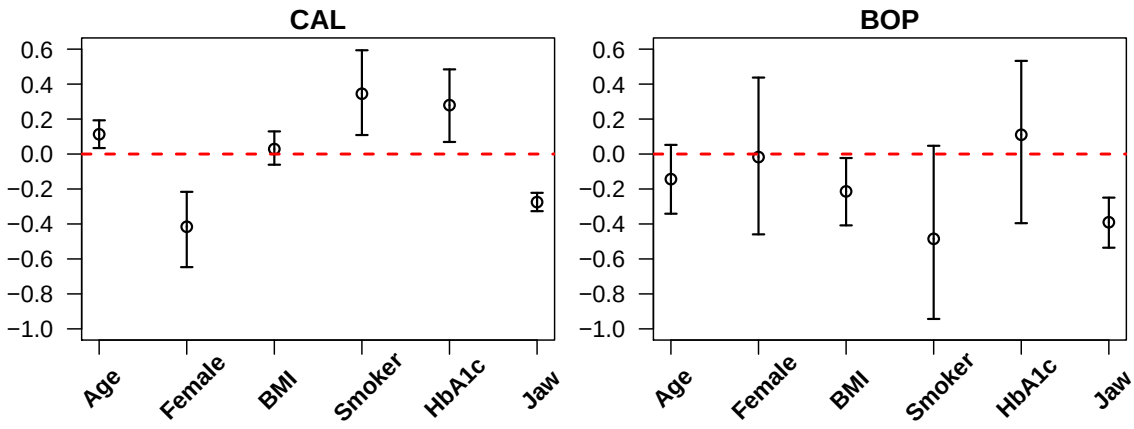

Figure 3.1: Posterior medians and 95% posterior intervals of the subject-specific covariate coefficients by response.

The posterior summaries of the auto- and cross-correlations of the subject-specific process $\boldsymbol{f}_i(t)$ from (3.2) are given in Figure 3.3; note that the correlation attributed to the subject random intercept is not included in this figure. In this periodontal application, these plots offer important and novel insights into the complex relationships that exist between and within the BOP and CAL responses in different parts of the mouth. The utility of quantifying and visualizing these complex correlation relationships is apparent for many other types of applications.

Examination of the diagonal of the auto-correlation plot for CAL in Figure 3.3 shows strong positive spatial correlation between adjacent teeth and between teeth separated by only one or two teeth on the same jaw. This plot also shows positive correlation between a tooth in the left and a tooth in the right side of the same jaw, and the relationship is particularly strong for teeth in the lower jaw. The correlation for CAL between teeth in opposite sides of the mouth and on different jaws is also positive, yet not as strong as for teeth on the same jaw; this correlation is very similar in magnitude as the correlation for teeth on the left or right side of the mouth but on different jaws. Additionally, there are mild to strong negative correlations between teeth in the center (front) of the mouth and teeth in the back of the mouth, regardless of the jaws on which the teeth are located. This is also seen in the plot of the posterior probability that the auto-correlation is positive.

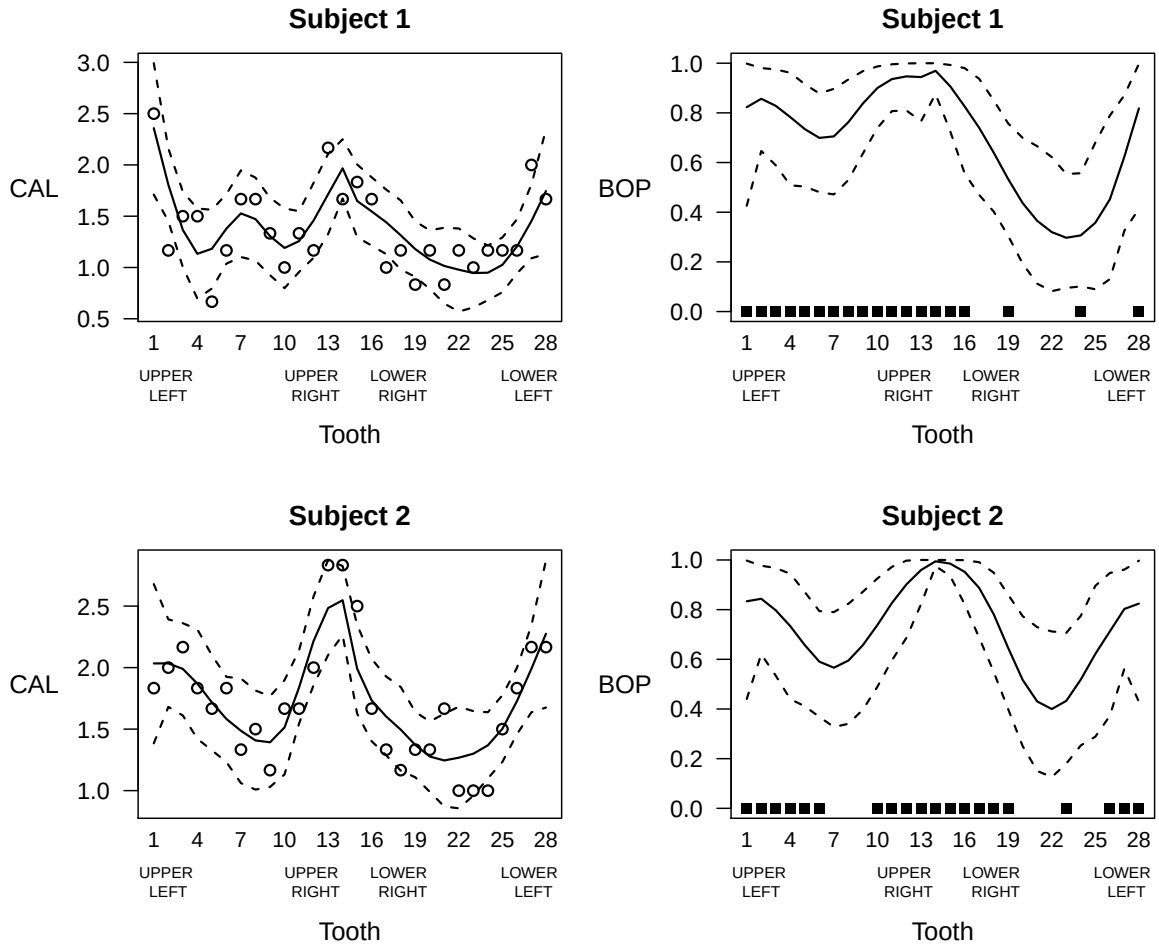

Figure 3.2: Fitted values for two individuals from the periodontal study (using Model 1). Left panels*: Observed values of CAL are shown as dots. The solid black line indicates the posterior mean of $\mu_{1i}(t)$, the subject-specific mean function, and point-wise 95% posterior intervals are given by the dotted lines.* Right panels*: The squares along the x-axis indicate the teeth for which BOP is observed. The solid black line gives the posterior mean of the conditional probability of the event, $P(Y_{2i}(t) = 1|\boldsymbol{\alpha}_{2i})$, and dotted lines show point-wise 95% posterior intervals. The label "UPPER LEFT" refers to the left side of the the upper jaw when looking at a patient, and it is analogous for the other labels.*

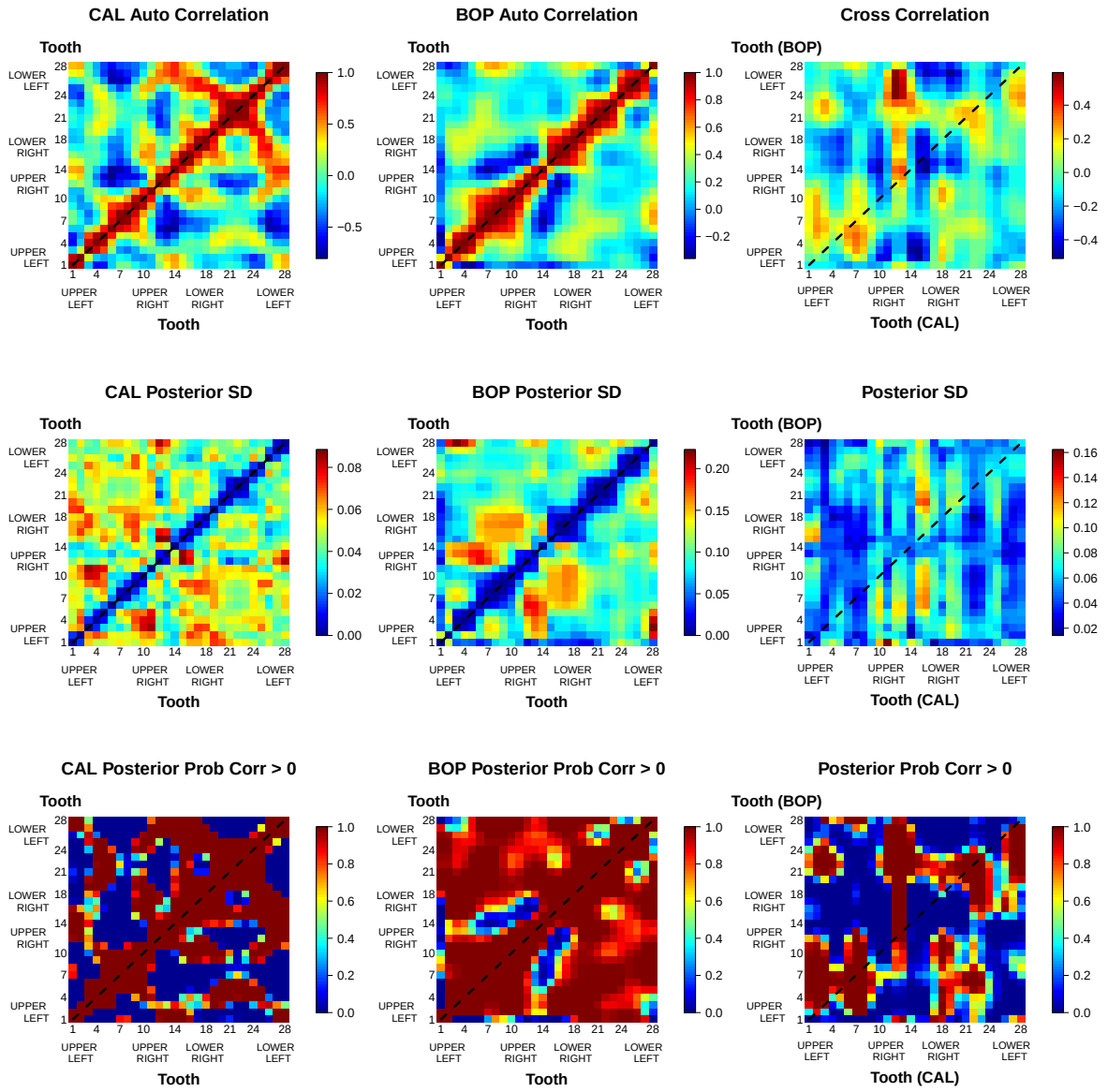

Figure 3.3: Posterior summaries of the within-response and between-response correlation structures for any two teeth when fit with Model 1 (excluding correlation from the subject random intercepts). The label "UPPER LEFT" refers to the left side of the the upper jaw when looking at a patient, and it is analogous for the other labels.

In the auto-correlation plot for BOP, again we see strong positive spatial correlation between adjacent teeth and between teeth that are close to one another on the same jaw. Additionally, the plots of the auto-correlation and of the probability of being positive show that the correlation is mostly positive with only a few areas of negative correlation. The correlation is negative between a tooth in the center and a tooth on the right side of the upper jaw, as well as between a tooth in the left and a tooth in the center of the lower jaw. There is also a strong negative correlation for teeth in the lower right and upper right, as well as for teeth in the lower left and lower right.

The cross-correlation between BOP and CAL ranges from moderately positive to moderately negative. Unlike the auto correlation plots, the cross correlation is not symmetric, which makes interpretation slightly more complex. For instance, BOP in the lower left is positively correlated with CAL in the center and upper right as indicated by the darkest patch near the top center of the cross correlation figure. Alternatively, CAL in the lower left shows slightly negative to no correlation with BOP in the center and upper right of the mouth. Another demonstration of this non-symmetric property occurs for the negative correlation of BOP in the lower left with CAL in the lower right, though BOP in the lower right shows slightly positive to no correlation with CAL in the lower left.

CHAPTER 4

Conclusion

The methodologies we present in Chapters 2 and 3 allow for joint modeling of bivariate (or multivariate) functional responses when the correlation structures within and between the response curves are very complex. A primary advantage of joint modeling over univariate methods is the ability to understand the interplay between responses, a feature which we highlight through applications of our methods to a colon carcinogenesis experiment and a periodontal disease study.

In Chapter 2 we propose a flexible Frequentist framework for jointly modeling multiple real-valued functional responses nested within a hierarchy where the functions are observed on a spatial grid and are assumed to exhibit spatial auto- and cross-correlations. We implement a novel moments-based approach to obtain a raw estimate of the spatial covariance matrix which we combine with an optimization procedure to obtain estimates of bivariate Matérn parameters. We utilize FPCA for level-1 and level-2 functional cross covariances. Accounting for these complex correlations leads to improved estimates of group mean functions.

In Chapter 3 we propose a Bayesian multivariate functional model for responses of different types, e.g. binary and continuous data, that utilizes a multivariate latent Gaussian process. We present two basis expansion options for the random subject-specific deviation, including a novel data-driven approach in which the estimated

basis functions are obtained through an extension of FPCA that we propose for mixed-type responses. Our method can account for subject-specific covariates that can be either linear or time-dependent (such as the jaw indicator used in the analysis of the periodontal study in Section 3.5).

The settings in Chapters 2 and 3 are complex, but distinct; we reiterate their differences here for added clarity. In Chapter 2, the setting involves a multilevel structure in which curves are spatially correlated across units within a subject. In Chapter 3, there is only one bivariate (or multivariate) functional response per subject. Though not explicitly shown, the method in Chapter 3 applies to the situation when subjects are nested in groups. The only modification needed is to estimate the latent covariance separately for each group and then combine the estimates before performing FPCA. Added complexity in Chapter 3 comes from allowing the response curves to be of different type, e.g. binary and continuous, whereas Chapter 2 is developed for real-valued responses.

Furthermore, the proposed method in Chapter 3 is flexible enough for functions to be observed at varying locations for different subjects and different responses. Our methodology in Chapter 2 requires responses to be observed at a dense, balanced design due to the way we estimate the spatial cross-correlation. In the imbalanced case where there is a different number of subunits measured across units, we recommend preprocessing the data by binning the observations in a way so that the data mimic the balanced case and then proceeding as previously described. One should verify that binning in this way leaves a sufficient number of subunits to warrant a functional data approach.

In Chapter 2, our proposed methodology for bivariate data is easily implemented computationally, making inferences using bootstrapping or other such techniques feasible. Extensions for functional response vectors of dimension greater than two are straightforward; estimation of the covariance of the spatial component encounters the same challenges described in Apanasovich et al. (2012). The Bayesian framework of Chapter 3 is inherently more time-consuming than Frequentist methods such at that of Chapter 2. However, important computational advantages can be gained by using the data-driven basis expansion that utilizes FPCA to capture most of the correlation across responses.

For exposition of Chapter 3 we focus on modeling a bivariate response vector

where one functional response is continuous and the other is binary, though joint modeling of more than two responses is a straightforward extension. Furthermore, the method easily models repeatedly observed categorical responses. This is achieved in a manner similar to thresholding the latent process at zero for binary data, but instead one must impose multiple thresholds on the latent process. Modeling other types of data, such as repeatedly observed count data, is not as straightforward as it would likely require using copulas (Nelsen 1999).

By estimating the multivariate covariance of the latent process, our methodology in Chapter 3 can offer novel insights into the cross-dependence of different responses, which is of interest in a wide variety of applications. Quantifying and exploring this dependence is an important contribution of our method and is a primary goal of our analysis of the periodontal data presented in Section 3.5. Reich and Bandyopadhyay (2010) and Reich et al. (2013) offer ways to incorporate informative missingness and apply their methods to the same periodontal data. We do not address the informative missingness for our analysis because it is not central to our goals, and leave it for future work.

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):pp. 669–679.

Apanasovich, T. V., Genton, M. G., and Sun, Y. (2012). A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497):180–193.

Baladandayuthapani, V., Mallick, B. K., Hong, M. Y., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64(1):64–73.

Berrendero, J. R., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634.

Besse, P. and Ramsay, J. (1986). Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311.

Boente, G. and Fraiman, R. (2000). Kernel-based functional principal components. *Statistics & Probability Letters*, 48(4):335 – 345.

Cardot, H., Mas, A., and Sarda, P. (2007). CLT in functional linear regression models. *Probability Theory Related Fields*, 138:325–561.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72.

Cressie, N. A. C. (1993). *Statistics for spatial data*. J. Wiley, New York.

Davidian M., Lin, X. and Wang, J.-L. (2004). Introduction: emerging issues in longitudinal and functional data analysis. *Statistica Sinica*, 14(3):613–614.

de Boor, C. (1978). *A practical guide to splines*. Springer, Berlin.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3(1):458–488.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

Fernandez, J. K., Wiegand, R. E., Salinas, C. F., Grossi, S. G., Sanders, J. J., Lopes-Virella, M. F., and Slate, E. H. (2009). Periodontal disease status in Gullah African Americans with Type 2 diabetes living in South Carolina. *Journal of Periodontology*, 80(7):1062–8.

Ferraty, F. and Romain, Y., editors (2011). *The Oxford handbook of functional data analysis*. Oxford University Press, New York.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and applications to spectrometric data. *Computational Statistics*, 17:545–564.

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

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Gonzalez-Manteiga, W. and Vieu, P. (2007). Statistics for functional data (editorial). *Computational Statistics & Data Analysis*, 51(10):4788–4792.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electron. J. Statist.*, 4:1022–1054.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika*, 93(4):989–995.

Hall, P. (2011). Principal component analysis for functional data: Methodology, theory, and discussion. In Ferraty, F. and Romain, Y., editors, *The Oxford handbook of functional data analysis*, pages 210–234. Oxford University Press, New York.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35(4):pp. 403–410.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2 edition.

Hong, M. Y., Chapkin, R. S., Barhoumi, R., Burghardt, R. C., Turner, N. D., Henderson, C. E., Sanders, L. M., Fan, Y.-Y., Davidson, L. A., Murphy, M. E., Spinka, C. M., Carroll, R. J., and Lupton, J. R. (2002). Fish oil increases mitochondrial phospholipid unsaturation, upregulating reactive oxygen species and apoptosis in rat colonocytes. *Carcinogenesis*, 23(11):1919–1926.

Indritz, J. (1963). *Methods in Analysis*. Macmillan, New York.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71(0):92 – 106.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, 64:411–432.

James, G. (2011). Sparseness and functional data analysis. In Ferraty, F. and Romain, Y., editors, *The Oxford handbook of functional data analysis*, pages 298–323. Oxford University Press, New York.

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

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):pp. 587–602.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining geostatistics*. Academic Press, London.

Karhunen, K. (1947). *Über lineare Methoden in der Wahrscheinlichkeitsrechnun*. Suomalainen Tiedeakatemia.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Loève, M. (1945). Fonctions aleatoires de second ordre. *C. R. Acad. Sci*, 220 469.

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

Matérn, B. (1986). *Spatial Variation*. Springer-Verlag, Berlin, 2 edition.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68:179–199.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98(463):pp. 573–583.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D., Hong, M., and Carroll, R. J. (2002). A Bayesian analysis of colonic crypt structure and coordinated response to carcinogen exposure incorporating missing crypts. *Biostatistics*, 3(4):529–546.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D., Hong, M. Y., and Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced dna adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association*, 96(455):pp. 816–826.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10:186–196.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer-Verlag, New York.

Pezzulli, S. and Silverman, B. (1993). Some properties of smoothed principal components. *Computational Statistics & Data Analysis*, 8(1):1–16.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer, New York, 1 edition.

Ramsay, J. and Silverman, B. (2002). *Applied functional data analysis: methods and case studies*. Springer, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York, 2 edition.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):pp. 539–572.

Reich, B. J. and Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *The Annals of Applied Statistics*, 4(1):439–459.

Reich, B. J., Bandyopadhyay, D., and Bondell, H. D. (2013). A nonparametric spatial model for periodontal data with nonrandom missingness. *Journal of the American Statistical Association*, 108(503):820–831.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14(3):631–647.

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 (Methodological)*, 53(1):pp. 233–243.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Sarda, P. and Vieu, P. (2012). Kernel regression. In Schimek, M. G., editor, *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons, Hoboken, NJ.

Serban, N., Staicu, A.-M., and Carroll, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics*, 69(4):903–913.

Shang, H. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2):121–142.

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

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11(2):177–194.

Ullah, S. and Finch, C. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(43).

Valderrama, M. (2007). An overview to modelling functional data (editorial). *Computational Statistics*, 22(3):331–334.

van der Linde, A. (2008). Variational Bayesian functional {PCA}. *Computational Statistics & Data Analysis*, 53(2):517 – 533.

van der Linde, A. (2009). A Bayesian latent variable approach to functional principal components analysis with binary and count data. *AStA Advances in Statistical Analysis*, 93(3):307–333.

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

Wand, M. P. and Jones, M. C. (1996). *Kernel Smoothing*. Chapman and Hall, London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Ser. A*, 26:359–372.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society Series B*, 62(2):413–428.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC, Boca Raton, FL.

Wu, C., Chiang, C., and Hoover, D. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93:1388–1402.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (1993). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59:676–685.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):pp. 577–590.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33:2873–2903.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3):601–619.

APPENDICES

APPENDIX A

Additional details for Chapter 2

A.1 Colon carcinogenesis study; additional results

Figures A.1 through A.3 show additional results from our analysis. In Figure A.3, the first eigenfunction of the level 1 process (which represents the average over subunit) explains the majority of the overall variation: if we define the total variation for response p be the sum of all eigenvalues, $\widehat{\sigma}_{pp}$, and $\widehat{\tau}_p^2$, then the first level 1 eigenvalue explains 88% and 85% of the total variation for $\log(\text{p27})$ and apoptosis, respectively. Cumulatively, the level 2 eigenfunctions explain about 10% of the variation for each response.

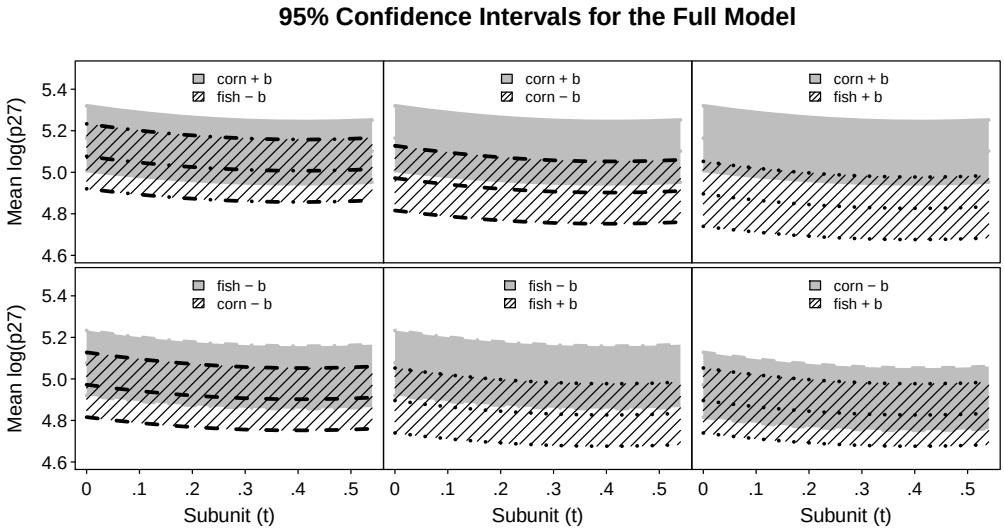

Figure A.1: Pairwise diet comparisons of mean log(p27) *estimates and* 95% *GLS confidence intervals for the FULL method. No adjustment for multiple comparisons has been made.*

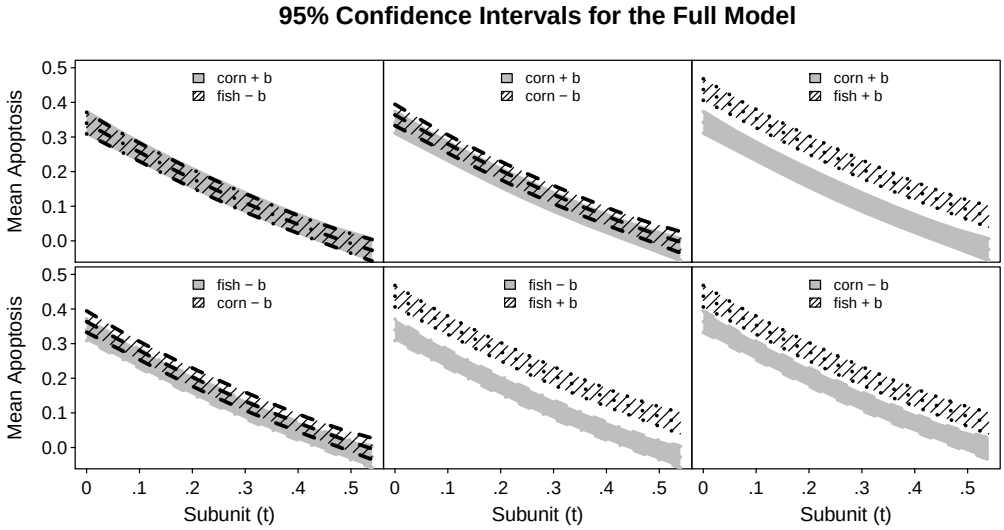

Figure A.2: Pairwise diet comparisons of mean apoptosis estimates and 95% *GLS confidence intervals for the FULL method. That the shaded regions do not overlap for the fish plus butyrate diet versus the others indicates they are significantly different. No adjustment for multiple comparisons has been made.*

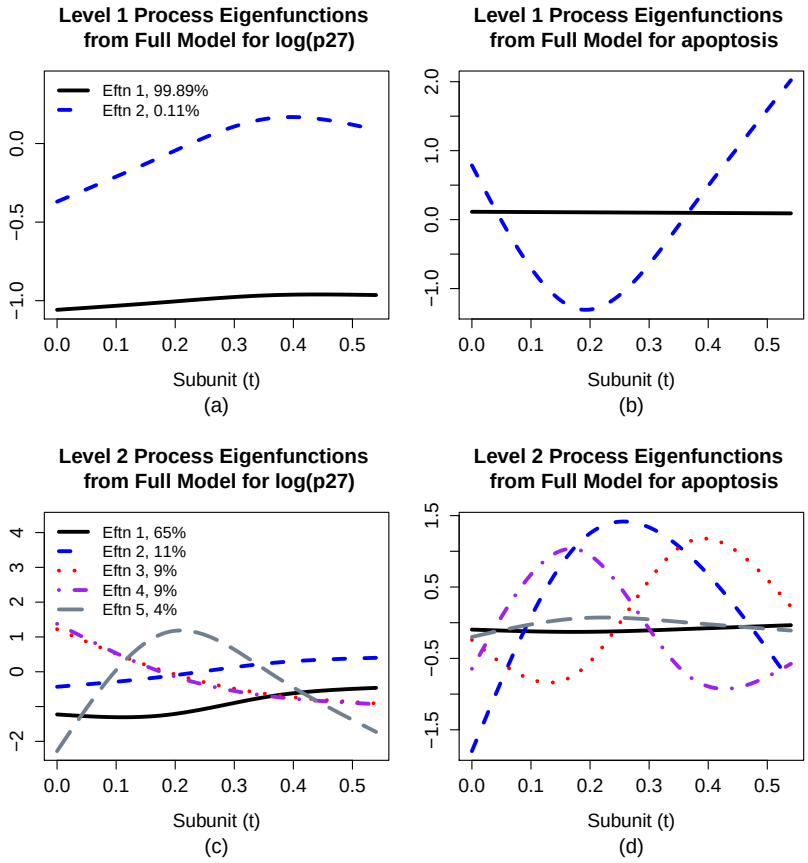

Figure A.3: Estimated eigenfunctions and corresponding percentage of variation explained for level 1 and level 2 functional processes.

A.1.1 Prediction

To assess the value of jointly modeling log(p27) and apoptosis, we perform 10-fold cross validation (CV) for each response separately, leaving the response of interest out of 2 crypts per subject in each fold. Let $\boldsymbol{Y}_{1,k}^{p}$ (hereafter $\boldsymbol{Y}_{1,k}$ for simplicity) be the vector obtained by stacking the n_k observations from the testing set from fold $k = 1, \ldots, 10$ and response p, and similarly obtain $\boldsymbol{Y}_{2,k}$ using the training set (that includes observations from both responses p and p'). Assuming $[\boldsymbol{Y}_{1,k}, \boldsymbol{Y}_{2,k}]^{\mathrm{T}} \sim \mathrm{MVN}([\boldsymbol{\mu}_{1,k}, \boldsymbol{\mu}_{2,k}]^{\mathrm{T}}, \boldsymbol{V}_k)$, the conditional distribution of $\boldsymbol{Y}_{1,k}|\boldsymbol{Y}_{2,k} = \boldsymbol{y}_{2,k}$ leads to the predicted values $\widehat{\boldsymbol{Y}}_{1,k} = \widehat{\boldsymbol{\mu}}_{1,k} + \widehat{\boldsymbol{V}}_{12,k}\widehat{\boldsymbol{V}}_{22,k}^{-1}(\boldsymbol{y}_{2,k} - \widehat{\boldsymbol{\mu}}_{2,k})$ with prediction

variance-covariance matrix $\boldsymbol{S}_k = \widehat{\boldsymbol{V}}_{11,k} - \widehat{\boldsymbol{V}}_{12,k}\widehat{\boldsymbol{V}}_{22,k}^{-1}\widehat{\boldsymbol{V}}_{21,k}$. Let $Y_{ij,k}^{p,\text{test}}(t, s_{ij})$ be the observed response and $\widehat{Y}_{ij,k}^{p,\text{test}}(t, s_{ij})$ be its corresponding predicted response from the testing set for CV fold k which has n_k observations. For apoptosis, $\widehat{Y}_{ij,k}^{p,\text{test}}(t, s_{ij})$ is truncated to be within $[0, 1]$, but the prediction coverage is calculated using non-truncated responses. All values in Table A.1 have been averaged over the 10 CV folds. For each k and each response p, we find the mean squared prediction error, MSPE $= n_k^{-1}\sum_{i,j,t}\{Y_{ij,k}^{p,\text{test}}(t, s_{ij}) - \widehat{Y}_{ij,k}^{p,\text{test}}(t, s_{ij})\}^2$; Bias $= n_k^{-1}\sum_{i,j,t}\{Y_{ij,k}^{p,\text{test}}(t, s_{ij}) - \widehat{Y}_{ij,k}^{p,\text{test}}(t, s_{ij})\}$; the CV Variance, or sample variance of the testing set; the prediction variance (Pred. Var), found by averaging the diagonals of $\boldsymbol{S}_k$; and finally the 95% prediction coverage from the prediction intervals formed using the diagonals of $\boldsymbol{S}_k$.

Table A.1 shows the results of performing prediction using the three different models. When compared to FULL, NS and UNIV have smaller MSPE for apoptosis, though practically the difference is inconsequential. All methods have very similar performance. Given the relationship between the two responses shown in Figure 2.3, we suspect that the small prediction differences between models is due to using binary data as continuous, which may not have preserved the features we expected to have been well suited for joint modeling. This could also explain why the prediction intervals have less than nominal coverage, especially for prediction of log(p27). All models produce very similar mean function estimates, seen in Figure A.4.

Table A.1: Prediction and Cross Validation

	MSPE	Bias	CV Variance	95% Pred. Coverage	Pred. Var
log(p27)					
Full	0.0098	-0.00520	0.037	0.74	0.003
No Spatial	0.0110	0.00011	0.033	0.64 $*$	0.002
Univariate	0.0098	-0.00460	0.037	0.74	0.003
apoptosis					
Full	0.13	0.0034	0.015	0.88	0.035
No Spatial	0.13 $**$	0.0030	0.014	0.88	0.035
Univariate	0.13 $**$	0.0028	0.014	0.88	0.035

Results from 10-fold Cross Validation. A '$**$' ('$*$') indicates better (worse) compared to FULL by paired t-test, $\alpha = 0.05$.

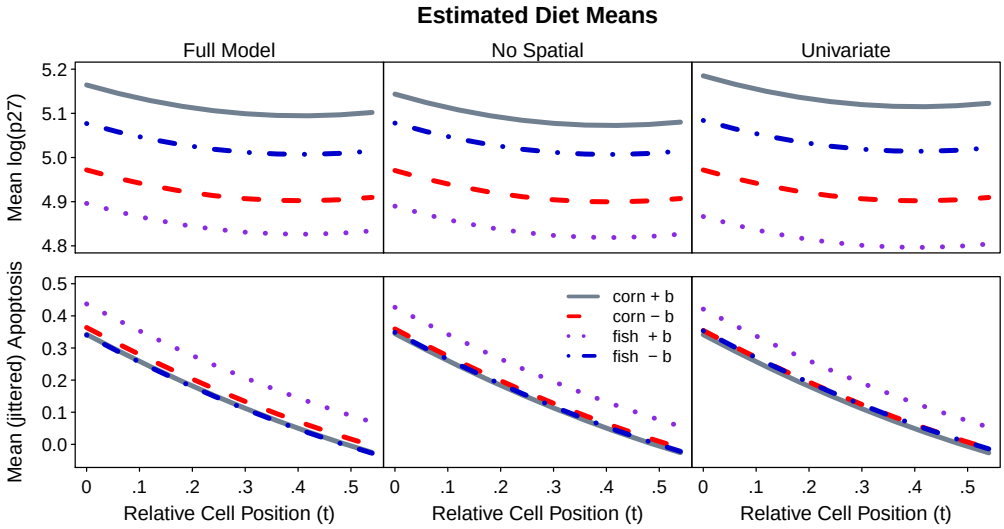

Figure A.4: Diet mean functions for each response from all models.

A.2 Simulations: Scenarios 5 & 6

In Scenarios 5 and 6 we assess the pitfalls of fitting FULL when it incorrectly incorporates spatial dependence by generating data with NS (Scenario 5) and when it incorrectly assumes additivity at the unit level (Scenario 6). For Scenario 6, we use a generating model (COMPLEX) as given in (2.1) where the quantity $\boldsymbol{\mathcal{Q}}_i(t, s_{ij})$ cannot be separated additively into one strictly functional piece and one strictly spatial piece. We compare the performance of misspecifying FULL to the performance of NS and UNIV, using TRUE as a baseline. Note that there is no estimation procedure available to fit COMPLEX in Scenario 6.

For Scenarios 5 and 6, we fix $\rho_{12} = 1$ and $\mathrm{N} = 10$, and all other parameters for Scenario 5 remain the same as in Scenarios 1-4. To generate the spatio-functional process $\boldsymbol{\mathcal{Q}}_i(t, s_{ij}) = [\mathcal{Q}_i^1(t, s_{ij}), \mathcal{Q}_i^2(t, s_{ij})]^{\mathrm{T}}$ for COMPLEX in Scenario 6, set $\mathcal{Q}_i^p(t, s_{ij}) = \sum_{\ell=1}^2 \sqrt{\lambda_\ell^W} \phi_\ell^{W,p}(t) U_{i,\ell}^p(s_{ij})$ where $\boldsymbol{U}_{i,\ell}(\mathbf{s}_i) = [U_{i,\ell}^{1,\mathrm{T}}(\mathbf{s}_i), U_{i,\ell}^{2,\mathrm{T}}(\mathbf{s}_i)]^{\mathrm{T}} \overset{i.i.d.}{\sim} \mathrm{N}\left(\mathbf{0}, \boldsymbol{C}^U(\boldsymbol{\Delta}_i)\right)$, and $\boldsymbol{C}^U(\boldsymbol{\Delta}_i) = \{\sigma_{pp'} M(\Delta | a_{pp'}, \nu_{pp'})\}_{p,p' \in \{1,2\}}$ is the bivariate Matérn correlation matrix for $p, p' = 1, 2$ with $\sigma_{11}, \sigma_{22} = 2.5$ to increase the effect of the spatial process. The covariance is given by $\mathrm{Cov}\{\mathcal{Q}_i^p(t, s_{ij}), \mathcal{Q}_i^{p'}(t', s_{ij'})\} = \sigma_{pp'} M(\Delta_{ijj'} | a_{pp'}, \nu_{pp'}) \sum_{\ell=1}^2 \lambda_\ell^W \phi_\ell^{W,p}(t) \phi_\ell^{W,p'}(t')$. All other parameter values are the same as in Scenarios 1-4.

Table A.2 shows the main results for scenarios 5 and 6. Note the very poor performance of UNIV in both scenarios. For the simpler case when NS is the generating model in Scenario 5, it is not surprising that the MISE from fitting NS is smaller than FULL. However, the coverage of FULL is closer to the nominal level and to TRUE in all cases since NS underestimates the variance in the model. In summary, overfitting FULL when the true model is less complex still results in very small MISE, small bias, and proper coverage. In Scenario 6, the generating model is more complex than the proposed model because it includes interactions between functions and spatial random effects. In this case FULL has larger MISE than TRUE, but maintains proper coverage. Estimated diet means for each scenario are shown in Figures A.5 and A.6.

Table A.2: Model comparisons for Scenarios 5 and 6

	90% Coverage	C.I. Length	MISE		Int Bias	Int Sq Bias	Int Var
Generating model: NS							
Response 1							
FULL	84.92	3.57	0.0190		0.0186	0.0002	0.02
UNIV	95.78	7.44	0.0425	*	0.0214	0.0001	0.04
NS	82.43	3.38	0.0187	**	0.0161	0.0001	0.02
TRUE	86.73	3.55	0.0159	**	0.0251	0.0002	0.02
Response 2							
FULL	84.03	3.74	0.0210		-0.0385	0.0001	0.02
UNIV	75.08	47.94	4.9622	*	0.0815	0.0260	4.94
NS	82.37	3.54	0.0208	**	-0.0360	0.0001	0.02
TRUE	89.05	3.78	0.0173	**	-0.0303	0.0001	0.02
Generating model: COMPLEX							
Response 1							
FULL	87.95	3.64	0.0193		-0.0313	0.0000	0.02
UNIV	95.75	7.36	0.0437	*	-0.0279	0.0000	0.04
NS	85.35	3.41	0.0173	**	-0.0315	0.0000	0.02
TRUE	88.55	3.56	0.0153	**	-0.0416	0.0001	0.02
Response 2							
FULL	90.27	4.14	0.0200		-0.0082	0.0002	0.02
UNIV	76.20	45.71	4.3479	*	0.0840	0.0554	4.29
NS	85.87	3.58	0.0181	**	0.0019	0.0001	0.02
TRUE	90.63	3.79	0.0160	**	0.0032	0.0002	0.02

Results in hundredths. A '$**$' ('$*$') indicates better (worse) MISE compared to FULL by Wilcoxson rank sum test, $\alpha = 0.05$.

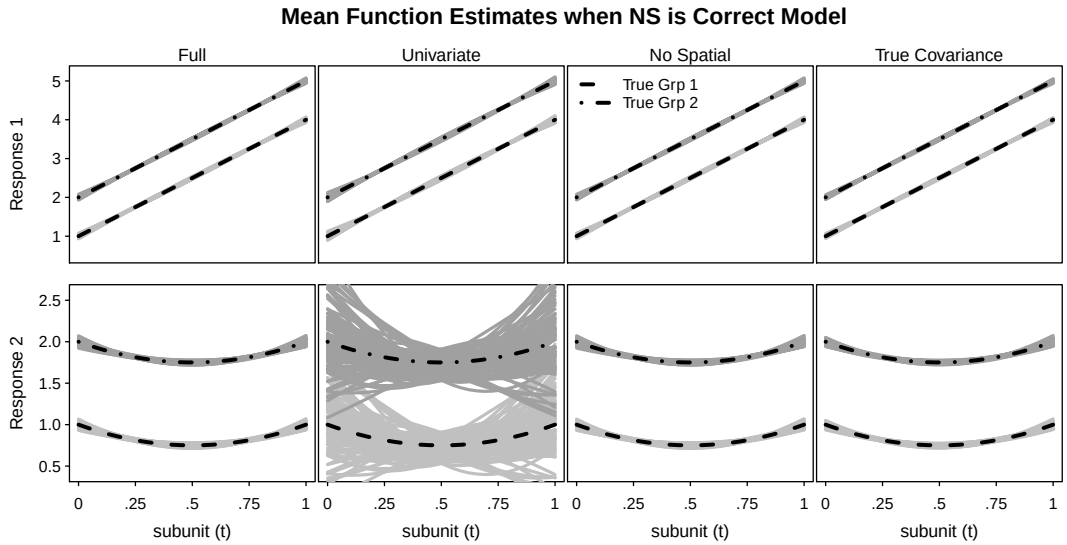

Figure A.5: Group mean functions when NS is the generating model with $\mathrm{N} = 10$ *(Scenario 5). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.*

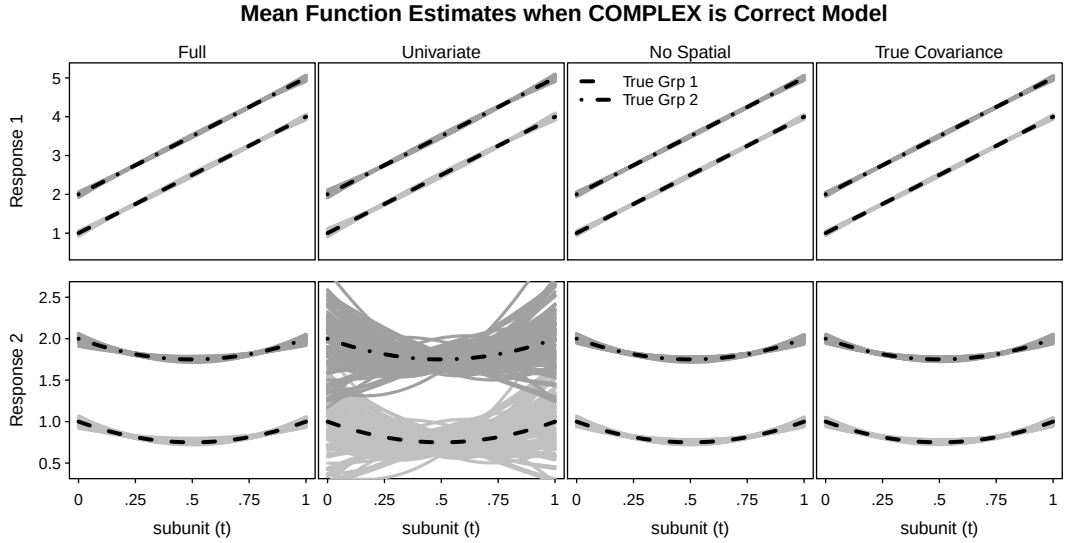

Figure A.6: Group mean functions when COMPLEX is the generating model with $\mathrm{N} = 10$ *(Scenario 6). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.*

A.3 Weighted least squares for initial values

Initial values of the Matérn parameters for the maximum likelihood procedure are found through the method of weighted least squares. Define an equally spaced grid of m increasing spatial lags $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_{\mathrm{m}}]$ where $\delta_1 = 0$ and $\delta_k \in (0, \Delta^*)$ for $k > 1$. As in (2.7), estimation of the cross semi-variogram is split into two cases, the first for $\delta_1 = 0$ and the second for $\delta_k \in (0, \Delta^*)$. For non-zero grid values δ_k, the estimator $\tilde{\gamma}_{pp'}(\delta_k, \epsilon)$ of the cross semi-variogram is given in (2.5) with expectation $\mathrm{E}\{\tilde{\gamma}_{pp'}(\delta_k, \epsilon)\} = C_{pp'}(0) - C_{pp'}(\delta_k) + \bar{\eta}_{b,pp'}$. Assuming a bivariate Matérn structure, this becomes $\gamma_{pp'}(\delta_k|\boldsymbol{\theta}_{pp'}, \bar{\eta}_{b,pp'}) = \mathrm{E}\{\tilde{\gamma}_{pp'}(\delta_k, \epsilon)\} = \bar{\eta}_{b,pp'} + \sigma_{pp'}\{1 - M(\delta_k|a_{pp'}, \nu_{pp'})\}$, where $\boldsymbol{\theta}_{pp'} = [\sigma_{pp'}, a_{pp'}, \nu_{pp'}]$ are bivariate Matérn parameters. For $\delta_1 = 0$, the cross semi-variogram estimator $\tilde{\gamma}^0_{pp'}$ is given in (2.6) and has expectation $\gamma^0_{pp',\epsilon} = \mathrm{E}\{\tilde{\gamma}^0_{pp',\epsilon}\} = \sigma_{pp'} + \bar{\eta}_{a,pp'}$. Let $\boldsymbol{\eta}_{pp'} = [\bar{\eta}_{a,pp'}, \bar{\eta}_{b,pp'}]$ be nuisance parameters. Define the loss function $\mathrm{L}_\epsilon(\boldsymbol{\theta}_{pp'}, \boldsymbol{\eta}_{pp'}) = |\mathcal{N}(\Delta^*, \epsilon)|\{1 - \gamma^0_{pp'}/\tilde{\gamma}^0_{pp',\epsilon}\}^2 + \sum_{k=2}^{\mathrm{m}} \mathcal{N}(\delta_k, \epsilon)\{1 - \gamma_{pp'}(\delta_k|\boldsymbol{\theta}_{pp'}, \bar{\eta}_{b,pp'})/\hat{\gamma}_{pp'}(\delta_k, \epsilon)\}^2$. We proceed in the same way as the three-step maximum likelihood procedure outlined in Section 2.3.1 in order to find $[\hat{\boldsymbol{\theta}}_{pp'}, \hat{\boldsymbol{\eta}}_{pp'}] = \underset{\boldsymbol{\theta}_{pp'}, \boldsymbol{\eta}_{pp'}}{\mathrm{argmin}}\{\mathrm{L}_\epsilon(\boldsymbol{\theta}_{pp'}, \boldsymbol{\eta}_{pp'})\}$ for $p, p' = 1, 2$, beginning with fitting the marginal parameters and treating them as fixed when estimating the cross parameters in terms of the quantities Δ_A, Δ_B, and σ^2_{12}.

A.4 Additional simulation results for Scenarios 1-4

The remaining figures and tables give additional simulation results from Scenarios 1-4. Diet mean estimates from each model are given for Scenarios 2 through 4 in Figures A.7 through A.9, respectively. Tables A.3 and A.4 show numeric results for Scenarios 3 & 4. Figure A.10 shows that the Matérn parameters are estimated well, and Figures A.11 through A.14 show that the level 1 & 2 eigenfunctions are estimated well.

Table A.3: Mean function estimation comparisons for Scenario 3 when FULL is generating model

	90% Coverage	C.I. Length	MISE		Int Bias	Int Sq Bias	Int Var
Response 1							
FULL	91.2	13.7	0.168		0.119	0.001	0.167
UNIV	91.6	14.2	0.172	$*$	0.122	0.001	0.171
NS	79.2	11.8	0.214	$*$	0.098	0.001	0.214
TRUE	90.8	13.7	0.163	$**$	0.091	< 0.001	0.162
Response 2							
FULL	87.8	13.7	0.190		0.076	< 0.001	0.190
UNIV	90.5	29.8	0.934	$*$	0.152	0.008	0.927
NS	75.3	11.7	0.224	$*$	0.209	0.001	0.224
TRUE	89.3	13.7	0.186	$**$	0.085	< 0.001	0.186

Results in hundredths. A '$**$' ('$*$') indicates better (worse) MISE compared to FULL by Wilcoxson rank sum test, $\alpha = 0.05$. Scenario 3: $\rho = 0.2$ and N $= 50$.

Table A.4: Mean function estimation comparisons for Scenario 4 when FULL is generating model

	90% Coverage	C.I. Length	MISE		Int Bias	Int Sq Bias	Int Var
Response 1							
FULL	84.9	30.9	1.149		-1.145	0.014	1.135
UNIV	85.5	32.1	1.182	$*$	-0.989	0.010	1.171
NS	69.1	23.6	1.293	$*$	-0.838	0.008	1.286
TRUE	88.3	30.6	0.956	$**$	-0.903	0.009	0.947
Response 2							
FULL	90.0	30.8	0.856		0.166	< 0.001	0.856
UNIV	86.1	60.2	4.771	$*$	0.178	0.160	4.610
NS	72.4	23.4	1.085	$*$	0.669	0.005	1.080
TRUE	92.6	30.6	0.768	$**$	0.217	0.003	0.766

Results in hundredths. A '$**$' ('$*$') indicates better (worse) MISE compared to FULL by Wilcoxson rank sum test, $\alpha = 0.05$. Scenario 4: $\rho = 0.2$ and N $= 10$.

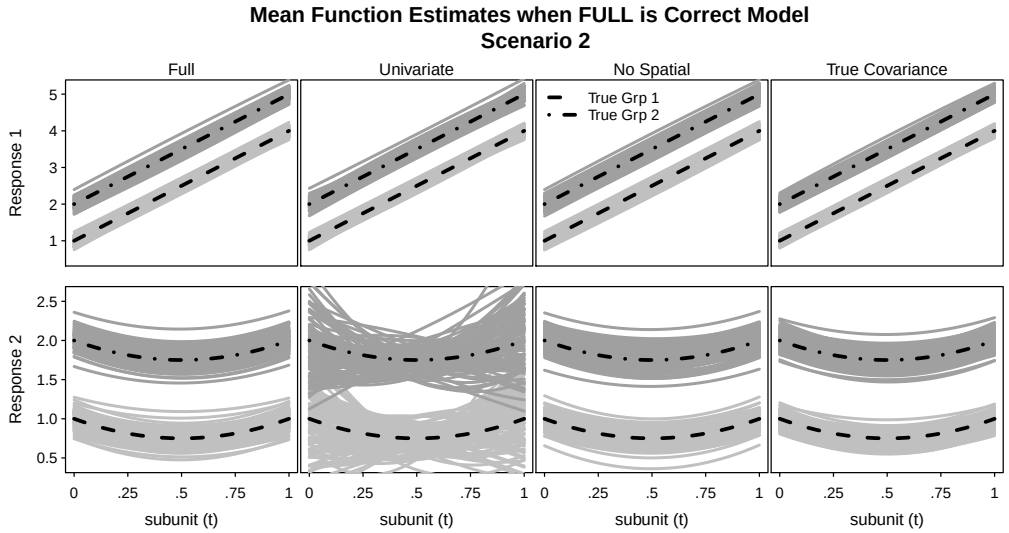

Figure A.7: Group mean functions when FULL is the generating model with $\rho = 0.8$ and $\mathrm{N} = 10$ (Scenario 2). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.

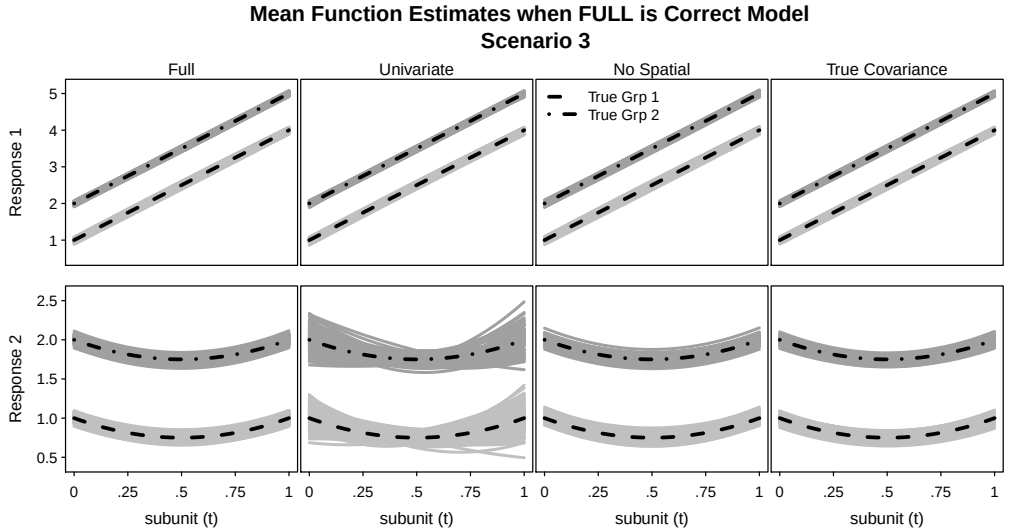

Figure A.8: Group mean functions when FULL is the generating model with $\rho = 0.2$ and $\mathrm{N} = 50$ (Scenario 3). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.

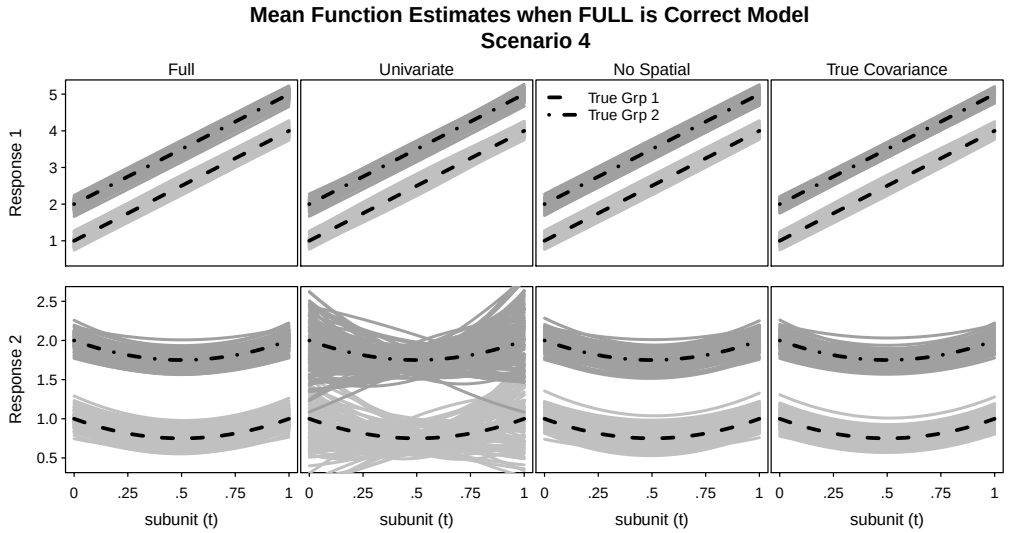

Figure A.9: Group mean functions when FULL is the generating model with $\rho = 0.2$ and $\mathrm{N} = 10$ (Scenario 4). Gray lines indicate estimated mean functions from each of the 100 Monte Carlo replications.

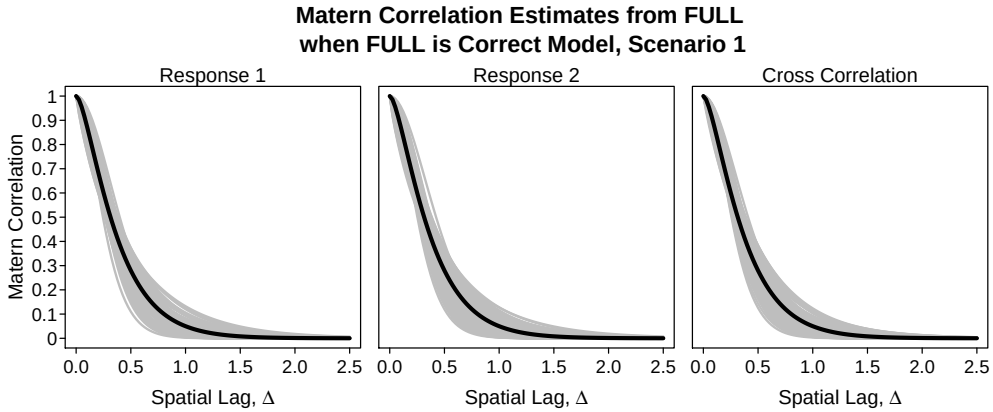

Figure A.10: Matérn correlation function estimates. Gray lines indicate estimated correlations from each of the 100 Monte Carlo replications.

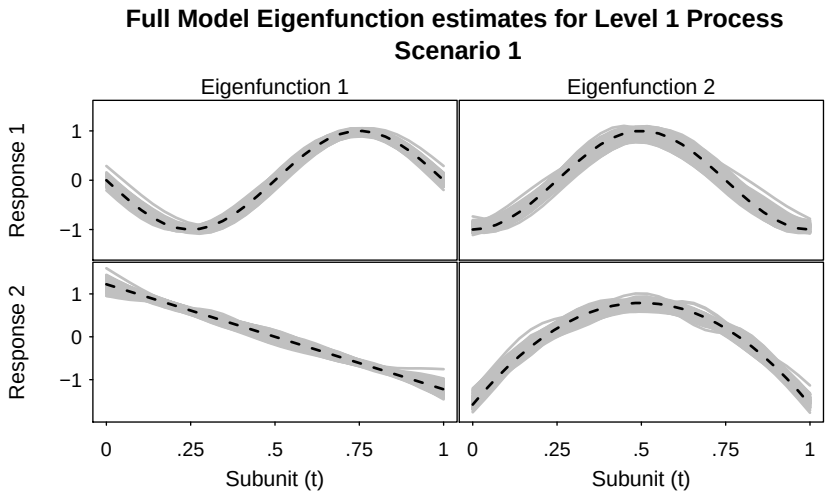

Figure A.11: Eigenfunctions obtained by fitting FULL for level 1 functional process when FULL is the generating model with $\rho = 0.8$ and $N = 50$ (Scenario 1). Solid gray lines indicate estimated functions from each of the 100 Monte Carlo replications while dashed black lines indicate true eigenfunctions.

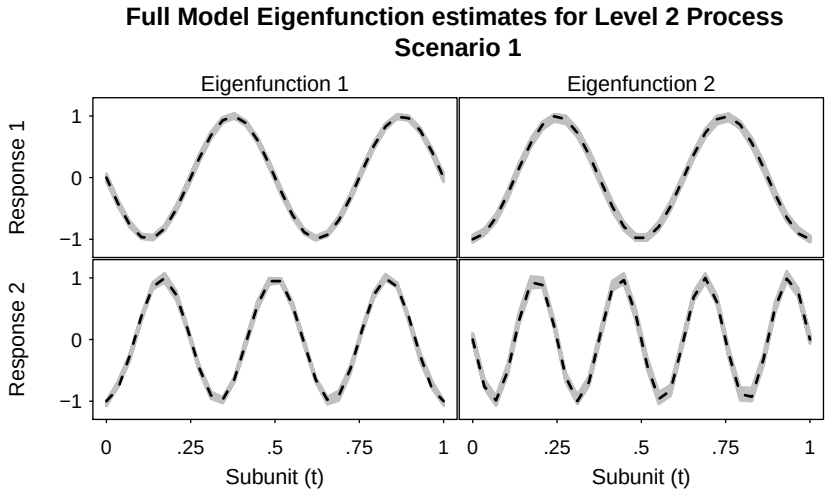

Figure A.12: Eigenfunctions obtained by fitting FULL for level 2 functional process when FULL is the generating model with $\rho = 0.8$ and $N = 50$ (Scenario 1). Solid gray lines indicate estimated functions from each of the 100 Monte Carlo replications while dashed black lines indicate true eigenfunctions.

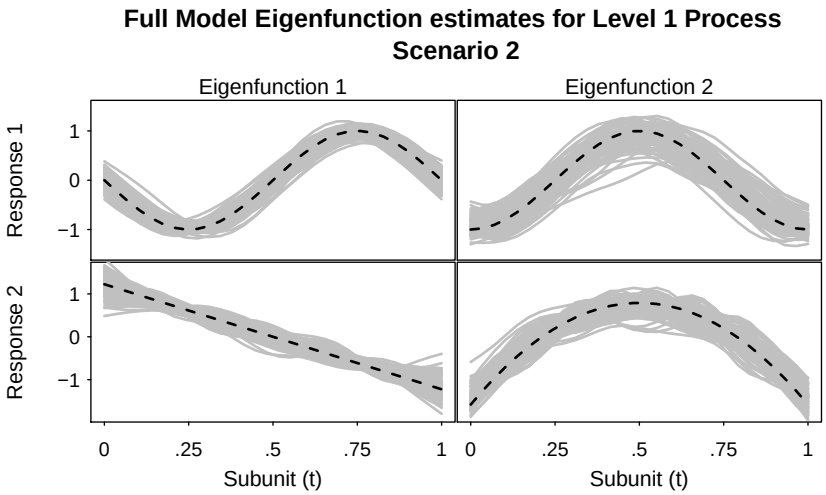

Figure A.13: Eigenfunctions obtained by fitting FULL for level 1 functional process when FULL is the generating model with $\rho = 0.8$ and $\mathrm{N} = 10$ (Scenario 2). Solid gray lines indicate estimated functions from each of the 100 Monte Carlo replications while dashed black lines indicate true eigenfunctions.

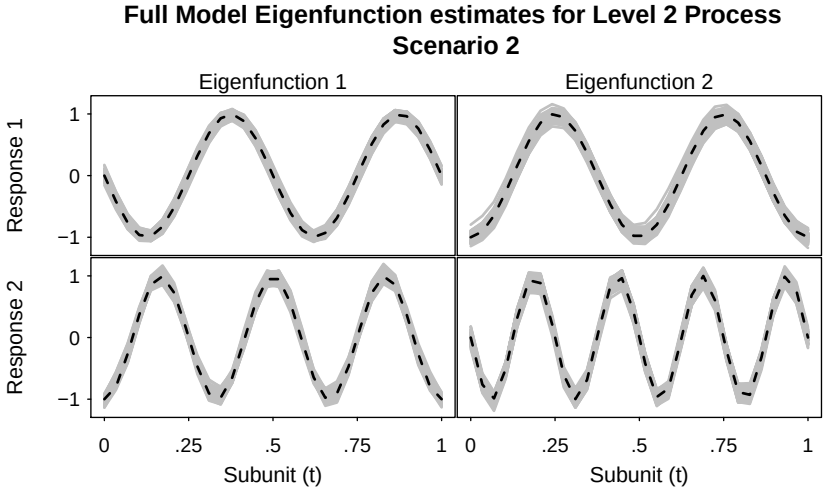

Figure A.14: Eigenfunctions obtained by fitting FULL for level 2 functional process when FULL is the generating model with $\rho = 0.8$ and $\mathrm{N} = 10$ (Scenario 2). Solid gray lines indicate estimated functions from each of the 100 Monte Carlo replications while dashed black lines indicate true eigenfunctions.

APPENDIX B

Additional details for Chapter 3

B.1 Approximating Smooth Covariance

In this section, we show that we can approximate any smooth covariance using the predetermined basis method. For simplicity, assume that the functional responses are observed at the same locations $t_{p\ell} \equiv t_\ell$ for $\ell = 1, \ldots, \mathrm{L}$ for each response p. We specify this model for an arbitrary subject, and thus drop the subscript i. Let $\boldsymbol{\psi}_{pk}$ be the vector of length L formed by evaluating at every t_ℓ the basis functions $\psi_{pk}(t)$, $k = 1, \ldots, \mathrm{M}_p$, and define the vector $\boldsymbol{f}_p$ analogously. Then we form the $\mathrm{L} \times \mathrm{M}_p$ matrix $\boldsymbol{\Psi}_p = [\boldsymbol{\psi}_{pk}, \ldots, \boldsymbol{\psi}_{p\mathrm{M}_p}]$ and the coefficient vector $\boldsymbol{\alpha}_{pi} = [\alpha_{p1}, \ldots, \alpha_{p\mathrm{M}_p}]^{\mathrm{T}}$ so that we can write $\boldsymbol{f}_p = \boldsymbol{\Psi}_p \boldsymbol{\alpha}_p$. We combine $\boldsymbol{f}_p$ for all responses to form one vector $\boldsymbol{f}$ of length $\mathrm{n} = \mathrm{PL}$, and define the coefficient vector $\boldsymbol{\alpha}^{\mathrm{T}} = [\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_{\mathrm{P}}^{\mathrm{T}}]$ of length $\mathrm{m} = \sum_{p=1}^{\mathrm{P}} \mathrm{M}_p$ and corresponding block-diagonal matrix $\boldsymbol{\Psi}$ of dimension $\mathrm{n} \times \mathrm{m}$ with blocks $\boldsymbol{\Psi}_p$. The resulting vector $\boldsymbol{f} = \boldsymbol{\Psi}\boldsymbol{\alpha}$ has length n, and we assume $\boldsymbol{\alpha} \overset{iid}{\sim} N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a covariance matrix of dimension $\mathrm{m} \times \mathrm{m}$ with elements $\mathrm{Cov}(\alpha_{pk}, \alpha_{p'\ell}) = \xi_{k\ell pp'}$.

To illustrate the flexibility of the model, assume $\boldsymbol{\Omega}_0$ is the true $\mathrm{n} \times \mathrm{n}$ covariance matrix of $\boldsymbol{f}$ evaluated at locations t_l. $\boldsymbol{\Omega}_0$ is now approximated by the variance-covariance matrix $\boldsymbol{\Omega} = \mathrm{Cov}(\boldsymbol{\Psi}\boldsymbol{\alpha}) = \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{\mathrm{T}}$. Since the basis comprising $\boldsymbol{\Psi}$ is pre-specified, the quality of the approximation $\boldsymbol{\Omega} \approx \boldsymbol{\Omega}_0$ is reliant on $\boldsymbol{\Sigma}$. By fitting a large

number of basis functions, i.e. setting m = n, it is possible to fit any smooth covariance function. When m = n then $\mathbf{\Psi}_i$ is a square matrix. Assume $\mathbf{\Psi}_i$ is full rank and thus $\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Psi}_i$ is invertible. Pre- and post- multiplication gives $\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Omega}\mathbf{\Psi}_i = \mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Psi}_i\mathbf{\Sigma}\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Psi}_i$. Since $\mathbf{\Theta} = \{\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Psi}_i\}^{-1}$ exists we can recover $\mathbf{\Sigma} = \mathbf{\Theta}\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Omega}\mathbf{\Psi}_i\mathbf{\Theta}$. Though this approach is quite flexible, it is hard to estimate $\mathbf{\Sigma}$ if it is high-dimension; therefore it is unlikely to perform well if the processes cannot be represented by a small number of basis functions.

B.2 Latent Cross Covariance Estimator

Here we describe in more detail the derivation of the latent cross covariance estimator. As our approach is inspired by Hall et al. (2008), we start with a brief summary of the method they proposed for finding the auto-covariance of the latent process corresponding to the binary response, that is, response $p = 2$. First, estimate the mean function for $p = 2$, $\hat{\mu}_2(t) = g^{-1}\{\hat{\eta}_2(t)\}$ where $\hat{\eta}_2(t)$ estimates $\mathrm{E}[g\{Z_{2i}(t)\}] = \eta_2(t)$ and is found by smoothing the data $\big(t, Y_{2i}(t)\big)$ for $i = 1, \ldots, \mathrm{N}$. Next, find the estimator $\hat{S}_{22}(t,t')$ of $S_{22}(t,t') = \mathrm{E}\{Y_{2i}(t)Y_{2i}(t)\} = \mathrm{E}\big[g\{Z_{2i}(t)\}g\{Z_{2i}(t)\}\big]$ by performing bivariate smoothing of the data $\big((t,t'), Y_{2i}(t)Y_{2i}(t')\big)$ for $i = 1, \ldots, \mathrm{N}$, once again removing the diagonals before smoothing. The estimator of the latent process covariance operator for the second response is given by

$$\widetilde{K}_{22}(t,t') = \{\hat{S}_{22}(t,t') - \hat{\eta}_2(t)\hat{\eta}_2(t')\}/[g^{(1)}\{\hat{\mu}_2(t)\}g^{(1)}\{\hat{\mu}_2(t')\}]. \tag{B.1}$$

Equation (B.1) was developed for a univariate response, so in order to estimate the latent cross covariance operator $K_{12}(t,t') = K_{21}(t',t) = \mathrm{Cov}\{Z_{1i}(t), Z_{2i}(t')\}$, we must derive an analogous estimator. This requires the following Taylor expansion, also given by equation (5) in Hall et al. (2008),

$$\begin{aligned} g\{Z_i(t)\} =& g\{\mu(t)\} + \delta X_i(t) g^{(1)}\{\mu(t)\} + \frac{1}{2}\delta^2\{X_i(t)\}^2 g^{(2)}\{\mu(t)\} \\ &+ \frac{1}{6}\delta^3\{X_i(t)\}^3 g^{(3)}\{\mu(t)\} + O_p(\delta^4). \end{aligned} \tag{B.2}$$

We can expand the covariance of the observed processes $\mathrm{Cov}\big\{Y_{1i}(t), Y_{2i}(t')\big\} = \mathrm{Cov}\big[Z_{1i}(t), g\{Z_{2i}(t')\}\big]$ by substituting (B.2) for $g\{Z_{2i}(t')\}$ and $\mu_1(t) + \delta X_{1i}(t)$ for

$Z_{1i}(t)$, which gives

$$\mathrm{Cov}\{Y_{1i}(t), Y_{2i}(t')\} = g^{(1)}\{\mu_2(t')\}\mathrm{Cov}\{\delta X_{1i}(t), \delta X_{2i}(t')\} + O(\delta^4). \tag{B.3}$$

Note that the term (suppressed from equation (B.3)) $\delta^3 \frac{1}{2} g^{(2)}\{\mu_2(t')\}\mathrm{Cov}\{X_{1i}(t), X_{2i}^2(t')\} = 0$ due to $\mathrm{Cov}\{X_{1i}(t), X_{2i}^2(t')\} = \mathrm{E}\{X_{1i}(t)X_{2i}^2(t')\} = \mathrm{E}[X_{2i}^2(t')\mathrm{E}\{X_{1i}(t)|X_{2i}(t')\}] = \sigma_1/\sigma_2\rho\mathrm{E}[X_{2i}^3(t')] = 0$ since $X_{1i}(t)|X_{2i}(t') \sim \mathrm{N}(\sigma_1/\sigma_2\rho X_{2i}(t'), (1-\rho^2)\sigma_1^2)$. Now, because $\mathrm{Cov}\{Z_{1i}(t), Z_{2i}(t')\} = \mathrm{Cov}\{\delta X_{1i}(t), \delta X_{2i}(t')\}$, we have from (B.3) that $K_{12}(t,t') = \mathrm{Cov}\{Z_{1i}(t), Z_{2i}(t')\} = \mathrm{Cov}\{Y_{1i}(t), Y_{2i}(t')\}/g^{(1)}\{\mu_2(t')\} + O(\delta^4)$, which, assuming the effect of $O(\delta^4)$ is negligible, leads to

$$K_{12}(t,t') = \mathrm{Cov}\{Y_{1i}(t), Y_{2i}(t')\}/g^{(1)}\{\mu_2(t')\}. \tag{B.4}$$

Estimation of (B.4) requires a smooth estimate $\hat{\eta}_1(t)$ of $\mathrm{E}[Y_{1i}(t)] = \eta_1(t)$ which is found by smoothing the data $\big(t, Y_{1i}(t)\big)$ for $i = 1, \ldots, \mathrm{N}$. We obtain the estimator $\hat{S}_{12}(t,t')$ of $S_{12}(t,t') = \mathrm{E}\{Y_{1i}(t)Y_{2i}(t')\} = \mathrm{E}\big[Y_{1i}(t)g\{Z_{2i}(t')\}\big]$ by performing bivariate smoothing of the data $\big((t,t'), Y_{1i}(t)Y_{2i}(t')\big)$ for $i = 1, \ldots, \mathrm{N}$, removing the diagonals before smoothing. The resulting smooth estimator of the latent cross covariance is

$$\widetilde{K}_{12}(t,t') = \{\hat{S}_{12}(t,t') - \hat{\eta}_1(t)\hat{\eta}_2(t')\}/g^{(1)}\{\hat{\mu}_2(t')\}, \tag{B.5}$$

which is the direct analogue to (B.1).

B.3 Derivations of Posteriors

In this section we present the derivations of the conditional posterior distributions.

B.3.1 Random effects

Let L_i be the number of subunits t observed for subject i and define the latent response vector for subject i as $\boldsymbol{W}_i = [W_1(t_1), \ldots, W_1(t_{\mathrm{L}_i}), W^2(t_1), \ldots, W^2(t_{\mathrm{L}_i})]^{\mathrm{T}}$, with corresponding mean vector $\mathrm{E}(\boldsymbol{W}_i) = \boldsymbol{U}_i\boldsymbol{\beta} + \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i$. Assume $\boldsymbol{W}_i|\boldsymbol{\alpha}_i \sim \mathrm{N}_{2\mathrm{L}_i}\big(\boldsymbol{U}_i\boldsymbol{\beta} + \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i, \boldsymbol{D}_i\big)$ where $\boldsymbol{D}_i = \mathrm{diag}(\tau_1^2, 1) \otimes \mathbf{I}_{\mathrm{L}_i}$, or in terms of the precision, $\boldsymbol{P}_i = \boldsymbol{D}_i^{-1} =$

$\text{diag}(\omega_1, 1) \otimes \mathbf{I}_{\text{L}_i}$. Also assume the m $\times$ 1 vector $\boldsymbol{\alpha}_i|\boldsymbol{Q} \sim \text{N}_\text{m}(0, \boldsymbol{Q}^{-1})$ for $i = 1, \ldots, \text{N}$ and for the m $\times$ m covariance matrix $\boldsymbol{Q}^{-1}$, or equivalently, the precision matrix $\boldsymbol{Q}$. Define $\boldsymbol{R}_i = \boldsymbol{W}_i - \boldsymbol{U}_i\boldsymbol{\beta}$. To find the posterior for $\boldsymbol{\alpha}_i|\cdot$ we know

$$\begin{aligned} p(\boldsymbol{\alpha}_i|\cdot) &\propto p(\boldsymbol{W}_i|\cdot) \times p(\boldsymbol{\alpha}_i|\boldsymbol{Q}) \\ &\propto \exp\left[-\frac{1}{2}\left\{(\boldsymbol{\Psi}_i\boldsymbol{\alpha}_i - \boldsymbol{R}_i)^\text{T}\boldsymbol{P}_i(\boldsymbol{\Psi}_i\boldsymbol{\alpha}_i - \boldsymbol{R}_i) + \boldsymbol{\alpha}_i^\text{T}\boldsymbol{Q}\boldsymbol{\alpha}_i\right\}\right], \\ &\propto \exp\left[-\frac{1}{2}\left\{\boldsymbol{\alpha}_i^\text{T}\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i\boldsymbol{\alpha}_i - 2\boldsymbol{R}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i\boldsymbol{\alpha}_i + \boldsymbol{R}_i^\text{T}\boldsymbol{P}_i\boldsymbol{R}_i + \boldsymbol{\alpha}_i^\text{T}\boldsymbol{Q}\boldsymbol{\alpha}_i\right\}\right], \end{aligned}$$

and ignoring terms not involving $\boldsymbol{\alpha}_i$ or $\boldsymbol{Q}$ results in

$$p(\boldsymbol{\alpha}_i|\cdot) \propto \exp\left(-\frac{1}{2}\left[-2\boldsymbol{R}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i\boldsymbol{\alpha}_i + \boldsymbol{\alpha}_i^\text{T}\left\{\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i + \boldsymbol{Q}\right\}\boldsymbol{\alpha}_i\right]\right).$$

We want to form this term into the kernel of a Gaussian distribution where the exponent is $-1/2(\boldsymbol{\alpha}_i - \boldsymbol{M})^\text{T}\boldsymbol{V}^{-1}(\boldsymbol{\alpha}_i - \boldsymbol{M}) = -1/2(\boldsymbol{\alpha}_i^\text{T}\boldsymbol{V}^{-1}\boldsymbol{\alpha}_i - 2\boldsymbol{M}^\text{T}\boldsymbol{V}^{-1}\boldsymbol{\alpha}_i + \boldsymbol{M}^\text{T}\boldsymbol{V}^{-1}\boldsymbol{M})$ for some matrices $\boldsymbol{M}$ and $\boldsymbol{V}$. To complete the square, set $\boldsymbol{V} = \{\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i + \boldsymbol{Q}\}^{-1}$ and match the coefficients of $\boldsymbol{\alpha}_i$, giving $\boldsymbol{R}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i = \boldsymbol{M}^\text{T}\boldsymbol{V}^{-1} \implies \boldsymbol{M} = \boldsymbol{V}\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{R}_i$. Thus, the full conditional posterior for $\boldsymbol{\alpha}_i$ is given by

$$\boldsymbol{\alpha}_i|\cdot \sim \text{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{V}_{\boldsymbol{\alpha}})$$

for $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \{\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i + \boldsymbol{Q}\}^{-1}\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i(\boldsymbol{W}_i - \boldsymbol{U}_i\boldsymbol{\beta})$ and $\boldsymbol{V}_{\boldsymbol{\alpha}} = \{\boldsymbol{\Psi}_i^\text{T}\boldsymbol{P}_i\boldsymbol{\Psi}_i + \boldsymbol{Q}\}^{-1}$.

B.3.2 Random effects precision matrix

Assume the m $\times$ 1 vector $\boldsymbol{\alpha}_i|\boldsymbol{Q} \sim \text{N}_\text{m}(0, \boldsymbol{Q}^{-1})$ for $i = 1, \ldots, \text{N}$ and the m $\times$ m precision matrix $\boldsymbol{Q} \sim \text{Wishart}_\text{m}(\boldsymbol{V}, \nu)$ for which the kernel of the density is given by $|\boldsymbol{Q}|^{(\nu-\text{m}-1)/2}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{Q})\right\}$. Define $\boldsymbol{S} = \sum_{i=1}^\text{N}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\text{T}$ as the sum of squares matrix of $\boldsymbol{\alpha}_i$. We use $\boldsymbol{S}$ to write $\sum_{i=1}^\text{N}\boldsymbol{\alpha}_i^\text{T}\boldsymbol{Q}\boldsymbol{\alpha}_i = \text{tr}(\sum_{i=1}^\text{N}\boldsymbol{\alpha}_i^\text{T}\boldsymbol{Q}\boldsymbol{\alpha}_i) = \text{tr}(\sum_{i=1}^\text{N}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\text{T}\boldsymbol{Q}) = \text{tr}(\boldsymbol{S}\boldsymbol{Q})$ in the kernel of the multivariate normal density, using the properties $\text{tr}(a) = a$ for scalar a and $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})$. We also use the following properties of the trace to combine like-terms: 1) $|\boldsymbol{A}|^{-1} = |\boldsymbol{A}^{-1}|$ for $\boldsymbol{A}$ invertible and 2) for two square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension, $\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$. Using this, we can show the conditional posterior $p(\boldsymbol{Q}|\cdot)$ for $\boldsymbol{Q}$ is proportional to the kernel of a

$\text{Wishart}_\text{m}\{(\boldsymbol{S}+\boldsymbol{V}^{-1})^{-1}, \text{N}+\nu\}$:

$$\begin{aligned} p(\boldsymbol{Q}|\cdot) &\propto \prod_{i=1}^{\text{N}} p(\boldsymbol{\alpha}_i|\boldsymbol{Q}) \times p(\boldsymbol{Q}) \\ &\propto |\boldsymbol{Q}^{-1}|^{-\text{N}/2} \text{exp}\left\{-\frac{1}{2}\text{tr}(\boldsymbol{S}\boldsymbol{Q})\right\} \times |\boldsymbol{Q}|^{(\nu-\text{m}-1)/2} \text{exp}\left\{-\frac{1}{2}\text{tr}(\boldsymbol{V}^{-1}\boldsymbol{Q})\right\} \\ &\propto |\boldsymbol{Q}|^{(\text{N}+\nu-\text{m}-1)/2} \text{exp}\left[-\frac{1}{2}\text{tr}\{(\boldsymbol{S}+\boldsymbol{V}^{-1})\boldsymbol{Q}^{-1}\}\right]. \end{aligned} \tag{B.6}$$

Thus, $\boldsymbol{Q}|\cdot \sim \text{Wishart}_\text{m}\{(\boldsymbol{S}+\boldsymbol{V}^{-1})^{-1}, \text{N}+\nu\}$.

B.3.3 Fixed effects

Assume as before that the $2\text{L}_i \times 1$ response vector for subject i is $\boldsymbol{W}_i|\boldsymbol{\alpha}_i \overset{indep}{\sim} \text{N}_{2\text{L}_i}(\boldsymbol{U}_i\boldsymbol{\beta} + \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i, \boldsymbol{D}_i)$ for $\boldsymbol{D}_i = \text{diag}(\tau_1^2, 1) \otimes \mathbf{I}_{\text{L}_i}$, or in terms of the precision, $\boldsymbol{P}_i = \boldsymbol{D}_i^{-1} = \text{diag}(\omega_1, 1) \otimes \mathbf{I}_{\text{L}_i}$. Define the $2\text{L}_i \times 1$ vector $\boldsymbol{U}_i = \boldsymbol{W}_i - \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i$. The $\text{r} \times 1$ vector $\boldsymbol{\beta} \sim \text{N}_\text{r}(\mathbf{0}, \boldsymbol{C}^{-1})$ for $\boldsymbol{C}^{-1} = \sigma_{\boldsymbol{\beta}}^2\mathbf{I}_\text{r}$. To find the full conditional distribution of $\boldsymbol{\beta}|\cdot$, we have

$$\begin{aligned} p(\boldsymbol{\beta}|\cdot) &\propto \prod_{i=1}^{\text{N}} p(\boldsymbol{W}_i|\cdot) \times p(\boldsymbol{\beta}) \\ &\propto \text{exp}\left(-\frac{1}{2}\left[\sum_{i=1}^{\text{N}}\left\{(\boldsymbol{U}_i\boldsymbol{\beta} - \boldsymbol{U}_i)^\text{T}\boldsymbol{P}_i(\boldsymbol{U}_i\boldsymbol{\beta} - \boldsymbol{U}_i)\right\} + \boldsymbol{\beta}^\text{T}\boldsymbol{C}\boldsymbol{\beta}\right]\right) \\ &\propto \text{exp}\left[-\frac{1}{2}\sum_{i=1}^{\text{N}}\left\{(\boldsymbol{U}_i\boldsymbol{\beta} - \boldsymbol{U}_i)^\text{T}\boldsymbol{P}_i(\boldsymbol{U}_i\boldsymbol{\beta} - \boldsymbol{U}_i)\right\} - \frac{1}{2}\boldsymbol{\beta}^\text{T}\boldsymbol{C}\boldsymbol{\beta}\right] \\ &\propto \text{exp}\left[-\frac{1}{2}\sum_{i=1}^{\text{N}}\left\{\boldsymbol{\beta}^\text{T}\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\boldsymbol{\beta} - 2\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\boldsymbol{\beta} + \boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\right\} - \frac{1}{2}\boldsymbol{\beta}^\text{T}\boldsymbol{C}\boldsymbol{\beta}\right] \end{aligned}$$

and ignoring constant terms results in

$$\begin{aligned} &\propto \text{exp}\left[-\frac{1}{2}\left\{\boldsymbol{\beta}^\text{T}\boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{\beta}^\text{T}\left(\sum_{i=1}^{\text{N}}\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\right)\boldsymbol{\beta} - 2\left(\sum_{i=1}^{\text{N}}\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\right)\boldsymbol{\beta}\right\}\right] \\ &\propto \text{exp}\left[-\frac{1}{2}\left\{\boldsymbol{\beta}^\text{T}\left(\boldsymbol{C} + \sum_{i=1}^{\text{N}}\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\right)\boldsymbol{\beta} - 2\left(\sum_{i=1}^{\text{N}}\boldsymbol{U}_i^\text{T}\boldsymbol{P}_i\boldsymbol{U}_i\right)\boldsymbol{\beta}\right\}\right] \end{aligned}$$

As we did before, we want to form this term into the kernel of a Gaussian distribution and where the exponent is $-1/2(\boldsymbol{\beta}-\boldsymbol{M})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{M}) = -1/2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\beta} - 2\boldsymbol{M}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\beta} + \boldsymbol{M}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{M})$ for some matrices $\boldsymbol{M}$ and $\boldsymbol{V}$. To complete the square, set $\boldsymbol{V} = \left(\boldsymbol{C} + \sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i\right)^{-1}$ and match the coefficients of $\boldsymbol{\beta}$, giving $\sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{V}^{-1} \implies \boldsymbol{M} = \boldsymbol{V}\left(\sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i\right)^{\mathrm{T}}$. Thus, the full conditional posterior for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}})$$

$$\text{for } \boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{\mathrm{N}} (\boldsymbol{W}_i - \boldsymbol{\Psi}_i\boldsymbol{\alpha}_i)^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i\right\}^{\mathrm{T}} \text{ and } \boldsymbol{V}_{\boldsymbol{\beta}} = \left(\sigma_{\boldsymbol{\beta}}^{-2}\mathbf{I}_{\mathrm{r}} + \sum_{i=1}^{\mathrm{N}} \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{U}_i\right)^{-1}.$$

B.3.4 Error Variance (Precision)

Assume that the error precision $\omega_1 \sim \text{Gamma}(g, h)$ where we parameterize the density such that g is the shape parameter and $h = 1/s$ is the inverse of the scale parameter, called the rate parameter. Specifically, if $X \sim \text{Gamma}(g, h)$ then $p(x|g, h) = x^{g-1}e^{-xh}\{h^g/\Gamma(g)\}$. For simplicity of notation, denote the continuous response at t_ℓ for subject i as $Y_{i\ell} = Y_{1i}(t_\ell)$ for $i = 1, \ldots, \mathrm{N}$ and $\ell = 1, \ldots, \mathrm{L}_i$, and define the total number of responses observed as $n = \sum_{i=1}^{\mathrm{N}} \mathrm{L}_i$. Let $Y_{i\ell} \overset{indep}{\sim} \mathrm{N}(\mu_{i\ell}, \text{precision} = \omega_1)$. Then

$$\begin{aligned}
p(\omega_1|\cdot) &\propto \prod_{i=1}^{\mathrm{N}}\prod_{\ell=1}^{\mathrm{L}_i} p(Y_{i\ell}|\cdot) \times p(\omega_1) \\
&\propto \omega_1^{n/2} \exp\left\{-\frac{\omega_1}{2}\sum_{i=1}^{\mathrm{N}}\sum_{\ell=1}^{\mathrm{L}_i}(Y_{i\ell} - \mu_{i\ell})^2\right\} \times \omega_1^{g-1}\exp(-\omega_1 h) \\
&\propto \omega_1^{(n/2+g)-1} \exp\left[-\omega_1\left\{\frac{1}{2}\sum_{i=1}^{\mathrm{N}}\sum_{\ell=1}^{\mathrm{L}_i}(Y_{i\ell} - \mu_{i\ell})^2 + h\right\}\right].
\end{aligned}$$

This is the kernel of a Gamma density, so the posterior for $\omega_1|\cdot \sim \text{Gamma}(g_\omega, h_\omega)$ with shape and rate parameters $g_\omega = n/2 + g$ and $h_\omega = 1/2\sum_{i=1}^{\mathrm{N}}\sum_{\ell=1}^{\mathrm{L}_i}(Y_{i\ell} - \mu_{i\ell})^2 + h$.