Title
Bayesian Analysis of Structured and Multidimensional Functional Data with Applications to Electroencephalography Experiments

Author
Shamshoian, John

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Analysis of Structured and Multidimensional Functional Data with Applications to Electroencephalography Experiments

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

John Shamshoian

2021

ABSTRACT OF THE DISSERTATION

Bayesian Analysis of Structured and Multidimensional Functional Data with Applications to Electroencephalography Experiments

by

John Shamshoian
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2021
Professor Donatello Telesca, Chair

Many modern biomedical studies record vast amounts of data on individual subjects. The observed data may often be conceptualized as arising from an underlying smooth stochastic process after discretization and contimation with noise. Data in this form may exhibit multidimensionality and complex structural features. For example, electroencephalography (EEG) records electrical activity in the brain over continuous time. Repeated trials of cognitive tasks in EEG experiments induce longitudinal *and* and functional dimensions, complicating estimation and inference.

Regularized estimation and rigorous uncertainty quantification is highly sought after in these settings. In this dissertation I leverage techniques from factor analysis, probabilistic principal components analysis, and Gaussian processes (GPs) in the Bayesian paradigm. These techniques are crucial to achieve simultaneous flexible estimation and adaptive regularization. Model performance and calibration is assessed through a series of numerical experiments. The proposed methods are applied to analyze a wide variety of biomedical data, including cognitive EEG experiments, global age-specific fertility rates, and sleep EEG.

The dissertation of John Shamshoian is approved.

Chad Hazlett

Sudipto Banerjee

Damla Şentürk

Donatello Telesca, Committee Chair

University of California, Los Angeles

2021

To my parents, Pete and Sandi, for constantly encouraging me to explore my scientific interests and passions. To my fiance, Kristen, for her endless love, support, and encouragement throughout this adventure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not be possible without the support I received from my advisor, dissertation committee, colleagues, friends, and family. Financially, my dissertation research was supported by the grant R01 MH122428-01 from the National Institute of Mental Health.

I am exceedingly grateful to my advisor Dr. Donatello Telesca for spending countless hours supporting, mentoring, and guiding me through this dissertation and beyond. Our conversations, both in person and virtual, have formed the backbone of my scientific problem solving process. I have learned so much from Dr. Telesca throughout my graduate work and his wisdom will undoubtedly guide me on the next stage of my career. Dr. Telesca has assisted in writing all sections of this dissertation.

I would also like to thank my dissertation committee for their generous support as well. Dr. Damla Şentürk has been an engaging teacher, committee member, and mentor since my very first day at UCLA. I am very thankful for her dedication, insights, and enthusiasm throughout my entire Ph.D journey. Dr. Şentürk has provided remarkably valuable feedback on Chapters 2 & 3 in this dissertation. Dr. Banerjee Sudipto has also been a wonderful teacher and committee member from whom I received fantastic input on my papers and presentations. I also want to thank Dr. Chad Hazlett, who served on my committee with such short notice yet provided me with very valuable feedback in the little time we had together.

A special thanks goes to Dr. Shafali Spurling Jeste, former UCLA Professor of Psychiatry, Neurology, Pediatrics and current Chief of Neurology at Children's Hospital Los Angeles. I drew upon her expertise in electroencephalography in my work and I am extremely grateful for both her input and data provided. Dr. Jeste contributed important scientific interpretations in Chapter 2 and 3 of this dissertation.

I want to thank my colleagues Aaron Scheffler and Qian Li. Aaron, my former office mate, and I had many illuminating conversations surrounding electroencephalography and functional data analysis. Over time, our relationship grew more personal and we still make time to engage academically and socially to this day. Qian, a former student of Dr. Telesca's, and

I worked briefly worked together as statistical consultants for the broader UCLA researcher domain. Throughout our consulting work, Qian elucidated the major areas of statistical research, helping me hone my research interests.

I would also like to thank Roxy Naranjo, the Biostatistics Student Affairs Officer, for being there for me every step of the way. I am eternally grateful for my undergraduate advisor, Dr. Dana Paquin, for helping to set me on this path. Thank you Doug Morrison, Tommy Gibson, and Juhyun Kim for helping me get through the first few years of classes and exams during the Ph.D. Thank you to all my friends at UCLA for all the memorable and enjoyable times.

To my fiance, Kristen Keller, I cannot thank you enough for all the years of love, support, and enthusiasm pertaining to my doctoral work. Your love has meant everything, especially in isolation together during the COVID-19 pandemic. I cannot begin to thank my family enough for the support and encouragement throughout my entire life. Dad, I want to thank you for instilling the wonder of science in me from a very young age. Mom, thank you for the outpouring of love and support and for raising me with important values. In addition, I want to thank my sister Alexa Barry, my grandmother Alice Katchadourian, and late grandmother June Shamshoian. Last but not least, I want to thank my newfound family members: Jeanne Keller, David Keller, and Laura Keller. To everyone, your selfless collective efforts have made this work possible.

VITA

2015	B.S. (Mathematics), California Polytechnic State University, San Luis Obispo
2015–2017	Graduate Teaching Assistant, Department of Biostatistics, UCLA
2017-2017	Statistical Consultant, Institution of Digital Research and Education, UCLA
2017–2020	Graduate Student Researcher, Department of Biostatistics, UCLA
2019–2021	Data Scientist Visiting Research Assistant, Lucid Circuit, Santa Monica, CA

PUBLICATIONS

Watson, G. L., Xiong, D., Zoller, J. A., **Shamshoian, J.**, Sundin, P., Bufford, T., Rimoin, A. W., Suchard, M. A., Ramirez, C. M. (2021). Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLoS Computational Biology*, *17*(3), e1008837.

Xiong, D., Zhang, L., Watson, G. L., Sundin, P., Bufford, T., Zoller, J. A., **Shamshoian, J.**, Suchard, M. A., Ramirez, C. M. (2020). Psuedo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. *Epidemics*, 33, 100418.

Li, Q., **Shamshoian, J.**, Şentürk, D., Sugar, C., Jeste, S., DiStefano, C., Telesca, D. (2020). Region-referenced spectral power dynamics of EEG signals: A hierarchical modeling approach. *The Annals of Applied Statistics*, *14*(4), 2053-2068.

Shamshoian, J., Şentürk, D., Jeste, S., Telesca, D. (2020). Bayesian analysis of longitudinal and multidimensional functional data. *Biostatistics (Oxford, England)*, kxaa041.

Shamshoian, J., Şentürk, D., Jeste, S., Telesca, D. Bayesian Covariance Regression in Functional Data. (In preparation).

Shamshoian, J., Şentürk, D., Jeste, S., Telesca, D. Bayesian Analysis of Region-Referenced Functional Data. (In preparation).

CHAPTER 1

Introduction

High-dimensional data is commonly measured over continuous domains such as time or space in fields such as public health, medicine, environmental science, and biomechanics. These data may be conceptualized as realizations from an latent smooth stochastic process, after discretization and noise contamination. The simplest examples of functional are random curves defined over a one-dimensional compact interval. More generally, functional data may be (1) multidimensional, so that each latent function is a member of $\mathbb{R}^p$, and (2) highly structured, where correlated functional data arises from individual sampling units. One such example is longitudinal functional data, where units have repeated functional responses over longitudinal time. Interpretation, regularization, and uncertainty quantification become challenging in this setting. A related problem in functional data analysis (FDA) is the notion of covariance regression. The covariance function is a central object in many applications with functional data. However, most applications assume the covariance function does not depend on exogenous covariates. Covariance regression allows exogenous covariates to impact second order dynamics, quantifying heterogeneity of functional responses. A third challenge in functional data is assessing model adequacy in structured settings. The ensuing chapters address these challenges in FDA. To guide the reader through the statistical aspects of this work, I will review five key areas: FDA, Gaussian processes, multidimensional and structured functional data, covariance regression, and pivotal discrepancy measures.

1.1 Functional Data Analysis

Consider a collection of real-valued functions, $X_1(t)$, ..., $X_n(t)$ defined on a compact interval $t \in \mathcal{T}$ with $X_i(t) \in l_2$. These random functions have an overall population mean, $\mu(t) = \mathbb{E}(X_i(t))$, and within-function dependency quantified by $c(t, t') = \text{Cov}(X_i(t), X_i(t'))$. Functional data is similar to longitudinal data, where t could index longitudinal time. However, in contrast to longitudinal data analysis, FDA treats each entire random function as a single unit of information. Furthermore, estimation of quantities such as $\mu(t)$ or $c(t, t')$ often is completed in a nonparametric framework, taking advantage of the rich information embedded in each $X_i(t)$.

In the real world, one only observes discrete realizations of $X_i(t)$, potentially contaminated with noise. One such data-generating process is $Y_i(t) = X_i(t) + \epsilon_i(t)$, where $t \in \{t_1, \ldots, t_n\}$, $\epsilon_i(t)$ represents noise, and $Y_i(t)$ is the actual observed data. Various techniques have been proposed to bridge this gap between $X_i(t)$ and $Y_i(t)$ [Wang et al., 2016, Yao et al., 2005]. One particularly effective strategy is hierarchical modeling through functional principal components analysis (FPCA). Similar to principal components analysis (PCA), FPCA serves to (1) reduce the instrinsic dimension of the data and (2) extract interpretable features. Much of the work in this dissertation has strong connections with FPCA, so we review it here.

1.1.1 Functional Principal Components Analysis

FPCA is the most popular core technique in FDA [Wang et al., 2016]. FPCA expresses the random functions $X_i(t)$ as linear combinations of orthogonal basis functions. This orhogonal basis is considered optimal because it explains more variation than any other basis for a fixed number of basis functions. FPCA is owed to Mercer's theorem [Mercer] and the Karhunen-Loève theorem [Karhunen, 1946, Loève, 1946]. Mercer's theorem states that $c(t, t')$ may be written as $\sum_{j=1}^{\infty} \lambda_j \psi_j(t) \psi_j(t')$, where λ_j and $\psi_j(t)$ are the eigenvalues and eigenfunctions of

$\psi_j(t)$ respectively. With these quantities defined, the Karhunen-Loève theorem states

$$X_i(t) = \sum_{j=1}^{\infty} \psi_j(t)\eta_{ij}$$

where $\eta_{ij} = \int_{\mathcal{T}} X_i(t)\psi_j(t)], dt$ are the subject-specific coefficients.

1.2 Gaussian Processes for Functional Data

Gaussian processes are flexible approaches to nonparametric function estimation; see for example O'Hagan [1978], Williams [1998], Neal [1999], and Rasmussen and Williams [2006]. Shi et al. [2005] propose a Gaussian process (GP) mixture model to analyze standing-up manoeuvres of paraplegia patients. The authors are mainly focused on predicting center of mass at a later time or predicting the center of mass trajectory for a new patient. Each standing-up has a few hundred training data points (involving output and input variables). Let N represent the number of time points recorded in a standing-up, $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ represent the response, and $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{Qi})^\top, i = 1, \ldots, N$. Then the proposed model is

$$\mathbf{y} \sim N(\mathbf{0}, C) \tag{1.1}$$

where C is an $N \times N$ covariance matrix. The (i, j) entry is denoted $C_{ij} = C(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. An example of a covariance function is

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \nu_0 \exp\left\{ -\frac{1}{2}\sum_{q=1}^{Q} w_q(x_{iq} - x_{jq})^2 \right\} + a_0 + a_1 \sum_{q=1}^{Q} x_{iq}x_{jq} + \sigma_0\delta_{ij} \tag{1.2}$$

where δ_{ij} is the kronecker delta. Based on this setup, prediction of response y^* at test input $\boldsymbol{x}^*$ follows from fundamental multivariate normal properties. Specifically, the posterior

distribution at y^* given the training data $\mathcal{D}$ is also a Gaussian distribution, with

$$\mathbb{E}(y^*|\mathcal{D}) = \psi^\top(\boldsymbol{x}^*)C^{-1}\boldsymbol{y} \tag{1.3}$$

$$\text{Var}(y^*|\mathcal{D}) = C(\boldsymbol{x}^*, \boldsymbol{x}^*) - \psi^\top(\boldsymbol{x}^*)\boldsymbol{C}^{-1}\psi(\boldsymbol{x}^*) \tag{1.4}$$

where $\psi(\boldsymbol{x}^*) = (C(\boldsymbol{x}^*, \boldsymbol{x}_1), \ldots, C(\boldsymbol{x}^*, \boldsymbol{x}_N)^\top$. The authors embed this model within a mixture model to capture hierarchical effects. Shi et al. [2007] extends this model to incorporate a mean structure to improve predictive performance. See Shi and Choi [2011], Shi et al. [2012], and Wang and Shi [2014] for further extensions. Despite the model's flexibility incorporating functional and time-stable covariates, these models assume a pre-specified covariance structure, such as squared exponential (equation 1.2) or Matérn. Pre-specifying a covariance structure limits flexibility for covariance estimation, since the marginal covariance surface is completely determined by a set of hyper-parameters.

With the goal of mean-covariance estimation in mind, Yang et al. [2016b] considers an appropriately defined Inverse-Wishart process prior. Suppose that the functional data contain n independent trajectories, denoted by $\{Y_i(\cdot);\, i = 1. \ldots, n\}$ and the ith trajectory has p_i measurements $\boldsymbol{t}_i = \{t_{i1}, \ldots, t_{ip_i}\}$. Their model is

$$Y_i(t_{ij}) = Z_i(t_{ij}) + \epsilon_{ij} \tag{1.5}$$

$$Z_i \sim GP(\mu, \Sigma) \tag{1.6}$$

$$\mu \mid \Sigma \sim GP(\mu_0, \frac{1}{c}\Sigma) \tag{1.7}$$

$$\Sigma \sim \text{IWP}(\delta, \Psi) \tag{1.8}$$

$\text{IWP}(\delta, \Psi)$ denotes an Inverse-Wishart process, with shape parameter δ and scale parameter Ψ. An Inverse-Wishart process is defined such that for any finite grid $\boldsymbol{t} = \{t_1, \ldots, t_p\}$, Σ evaluated at this grid is Inverse-Wishart distributed, i.e., $\Sigma(\boldsymbol{t}, \boldsymbol{t}) \sim IW(\delta, \Psi(\boldsymbol{t}, \boldsymbol{t}))$. This method automatically induces mean-covariance estimation and smooths all functional observations simultaneously. However, similar to the Inverse-Wishart distribution, the IWP has issues: the uncertainty for all variances is controlled by a single degree of freedom parameter

δ Gelman et al. [2013], the marginal distribution for the variances has low density in a region near zero Gelman et al. [2006], and there is an *a priori* dependence between correlations and variances Tokuda et al. [2011].

Another approach for FDA using GPs is from Suarez and Ghosal [2017]. The goal in this paper is to estimate functional principal components and their number to retain. The following is a brief summary of their low rank model. Assume functional responses $\boldsymbol{Y}_i$ are measured on a common grid of points $\{t_1, \ldots, t_T\}$. Let H_J be a $T \times J$ basis matrix whose columns consist of basis functions evaluated at all grid points. Let $\Xi = ULU^\top$ be a prior guess for the covariance surface projected on H_J, where U is orthogonal and L is the diagonal matrix of ordered eigenvalues. Let $\Xi_K = U_K L_K U_K^\top$, where U_K is the $J \times K$ matrix formed by the first K columns of U, and L_K is the $K \times K$ matrix formed by the first K rows and K columns of L. The basic low rank model is as follows:

$$\boldsymbol{Y}_i \sim N(H_J U_K \boldsymbol{\beta}_{i,K}, \sigma^2 I) \tag{1.9}$$

$$\boldsymbol{\beta}_{i,K} \sim N(\boldsymbol{\theta}, \Sigma) \tag{1.10}$$

$$\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \tau\Sigma) \tag{1.11}$$

$$\sigma^2 \sim \text{inv-Gamma}(a, b) \tag{1.12}$$

$$\Sigma^{-1} \sim \text{Wishart}(\nu, L_K^{-1}) \tag{1.13}$$

The authors sample from the posterior for fixed (J, K). To perform model selection, the authors compute estimates of the marginal likelihood using results from Chib [1995]. The authors state asymptotic results, compare their approach with FACE Xiao et al. [2016], and apply their method to Canadian weather data, which is freely available in the `fda` package in `R`. One drawback of this procedure is the limitations induced by the Inverse-Wishart distribution, as discussed previously.

Montagna et al. [2012] developed a Bayesian latent factor model for function on scalar regression, which has connections with FPCA. Let $Y_i(t)$ denote the response at time t. An

abbreviated version of their model is as follows

$$Y_i(t) = f_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \varphi^2) \tag{1.14}$$

$$f_i(t) = \sum_{k=1}^{K} \tilde{\phi}_k(t)\eta_{ik} + r_i(t), \quad \eta_{ik} \sim N(x_i^\top \beta, 1) \tag{1.15}$$

$$\tilde{\phi}_k(t) = \sum_{p=1}^{P} \lambda_{pk} b_p(t), \quad r_i(t) = \sum_{p=1}^{P} \zeta_{ip} b_p(t) \tag{1.16}$$

When $r_i(t) = 0$, $\beta = 0$, $\tilde{\phi}_k(t)$ are eigenfunctions, and η_{ik} are scores, the model is equivalent to conventional FPCA. However, unlike FPCA, the authors induce sparsity on the loading coefficients λ_{pk} through a multiplicative gamma process shrinkage prior Bhattacharya and Dunson [2011]. The implied marginal covariance function from this model is $B(t)\Lambda\Lambda^\top B(t)^\top$, where $B(t)$ is a matrix of basis functions and Λ is a $P \times K$ loading matrix with entry (p, k) equal to λ_{pk}. This covariance function is quite flexible and hence this model serves as the building block in our proposed methods.

1.3 Multidimensional and Structured Functional Data Analysis

The literature surrounding FDA has shifted to consider more complex dependency structures than independent and identically distributed curves. Functional mixed models have been investigated in Guo [2002], Morris et al. [2003a], and Morris and Carroll [2006]. Functional mixed effects models inhereit the flexiblity of the linear mixed model in handling complex designs and correlation structures, and include nested curves as a special case. The model in Guo [2002] has

$$y_{ij}(t) = X_{ij}\boldsymbol{\beta}(t) + Z_{ij}\boldsymbol{\alpha}_i(t) + \epsilon_{ij}(t), \quad \epsilon_{ij}(t) \sim N(0, \sigma_e^2) \tag{1.17}$$

where $y_{ij}(t)$ is the response from the ith unit for the jth repetition at time t, X_{ij} is a $1 \times p$ fixed effects design matrix, Z_{ij} is a $1 \times q$ random effects design matrix, $\boldsymbol{\beta}(t)$ is a $p \times 1$ vector of fixed effect functions, $\boldsymbol{\alpha}_i(t)$ is a $q \times 1$ vector of random effect functions, and $\epsilon_{ij}(t)$ captures

measurement error. Smoothing splines are used to estimate $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}_i(t)$, ensuring that the fixed effects and random effects have the same smoothness property. Morris and Carroll [2006] has a similar model to Guo [2002] with more flexible covariance assumptions and uses wavelets to model irregular functions to account for non-stationarities such as different variances and different degrees of smoothness at different locations in the curve-to-curve deviations.

Di et al. [2009] develops Multilevel Functional Principal Components Analysis (MFPCA) for repeatedly observed functional data. The MFPCA model starts from the two-way functional ANOVA model

$$X_{ij}(t) = \mu(t) + \eta_j(t) + Z_i(t) + W_{ij}(t) \tag{1.18}$$

where $X_{ij}(t)$ is the signal for the ith unit, jth repetition, μ is a fixed overall mean function, η_j is a fixed jth repetition shift from the mean μ, Z_i is a random unit-specific shift from μ, and W_{ij} is a random unit-repetition shift from $\mu + \eta_j$. The random functions Z_i and W_{ij} are expanded as a KL decomposition

$$Z_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k^{(1)}(t), \quad W_{ij}(t) = \sum_{l=1}^{\infty} \zeta_{ijl}\phi_l^{(2)}(t) \tag{1.19}$$

Estimation is carried out in several steps. First, the mean functions $\hat{\mu}$, $\hat{\eta}_j$, and the relevant covariance functions are estimated by method of moments (in principle, any consistent estimator can be used). Then estimate functional principal components and scores $\hat{\phi}_k^{(1)}(t)$, $\hat{\phi}_l^{(2)}(t)$, and ξ_{ik}, ζ_{ijl} respectively. See Di et al. [2009] for more details. Interestingly, many methodological developments over the years for more complex dependency structures seem to follow the same pattern of estimating mean and covariance functions consistently, and then estimating eigenfunctions and scores in a multi-step algorithm. Additional references for repeatedly observed functional data include Kundu et al. [2016], Crainiceanu et al. [2009], Zipunnikov et al. [2014], and Shou et al. [2015].

Longitudinal functional data analysis Greven et al. [2010], Chen and Müller [2012], Park and Staicu [2015b], Lynch and Chen [2018a], is an extension of multilevel FDA, whereby longitudinal ordering of curves is taken into account. An example of a longitudinal functional

model particularly relevant for Chapter 2 background material is the *marginal* and *product* decompositions of Lynch and Chen [2018a]. For a function observed with longitudinal argument s, functional argument t, and functional response $Y_i(s,t)$ for the ith unit, the marginal decomposition has

$$Y_i(s,t) = \mu(s,t) + \sum_{j=1}^{\infty} \xi_{ij}(t)\psi_j(s), \quad \xi_{ij}(t) = \sum_{k=1}^{\infty} \chi_{ijk}\phi_{jk}(t) \tag{1.20}$$

where $\psi_j(s)$ are eigenfunctions of a certain margin covariance function (Park and Staicu [2015b], Lynch and Chen [2018a]), $\xi_j(t)$ are random coefficient functions in $\psi_j(s)$, and $\xi_j(t) = \sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t)$ is the KL expansion of the random coefficient functions $\xi_j(t)$. The marginal decomposition is a flexible nonparametric model for longitudinal functional data. A more parsimonious model than equation 1.20 yields the product decomposition

$$Y_i(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{ijk}\phi_k(t)\psi_j(s) \tag{1.21}$$

The validity of the product decomposition depends on the assumption of weak-separability Lynch and Chen [2018b]. Both the marginal and product decompositions yield parsimonious models with interpretable functions $\psi_j(s), \phi_{jk}(t)$ and $\phi_k(t)$. These appealling properties are explored in a Bayesian context in Chapter 2.

Over the years, the literature has seen contributions to multivariate functional data analysis (Berrendero et al. [2011], Jacques and Preda [2014], Chiou et al. [2014], Chiou et al. [2016], Happ and Greven [2018]), multilevel multivariate functional data analysis Zhang et al. [2019], and region-referenced longitudinal functional data analysis Hasenstab et al. [2017], Scheffler et al. [2020a].

1.4 Covariance Regression

Regression models are often synonymous with estimating a mean function depending on covariates, i.e., a linear regression where $\mu_x = \mathbb{E}(y|x) = x^\top\beta$. Statistical models in multivariate

analysis often make homogenous assumptions on covariance-covariate relationships such as $\Sigma_x = \text{Cov}(y|x) = \text{Cov}(y) = \Sigma$, where $y \in \mathbb{R}^p$. Flury [1984] develops a likelihood ratio test for detecting equal principal components across groups. Mathematically, the common principal components assumptions is

$$B\Sigma_i B^\top = \Lambda_i, \quad i = 1, \ldots, K \tag{1.22}$$

where Λ_i is diagonal, B is an orthonormal $p \times p$ matrix, and Σ_i is the covariance matrix for the ith group. Flury [1987], Schott [1991], Schott [1999] generalized this concept to test if only q out of p eigenvectors are common to all groups. Boik [2002] extended this idea to test if q eigenvectors are common to only some of the groups. More recently, Hoff [2009] developed a hierarchical model to pool eigenvector information across groups. Group-specific eigenvectors are similar, but not necessarily equal due to the hierarchical nature of the model. Franks and Hoff [2019] developed a shared subspace model, where groups share a common subspace. Each individual group has a latent group-specific subspace. Estimating and choosing the dimension of the shared subspace is discussed in their work. Although the literature has expanded since Flury [1984], the literature surrounding pooled covariance estimation with continuous covariates has remained sparse. One method for handling continuous covariates is given in Zhao et al. [2018]. Their model embeds principal components analysis (PCA) within a generalized linear model (GLM) framework. However, interpretation of relevant covariance matrices and eigenvectors is not straightforward, which is crucial in exploratory FDA. We modify the model in Hoff and Niu [2012] to accommodate functional data. The model in Hoff and Niu [2012] is similar to factor analysis (FA), with some important differences. Let y_i be a p-dimensional outcome with d-dimensional covariate x_i and μ_{x_i} be the conditional mean. The random effects representation has the form

$$y_i = \mu_{x_i} + \gamma_i B x_i + \epsilon_i \tag{1.23}$$

$$\mathbb{E}(\epsilon_i) = 0, \quad \text{Cov}(\epsilon_i) = \Psi \tag{1.24}$$

$$\mathbb{E}(\gamma_i) = 0, \quad \text{Var}(\gamma_i) = 1, \quad \epsilon_i \perp\!\!\!\perp \gamma_i \tag{1.25}$$

where B is a $p \times d$ loading matrix, γ_i is a subject specific random effect, and ϵ_i captures residual covariance with baseline covariance Ψ. Marginalizing over the random effects, $\text{Cov}(y_i) = Bx_i x_i^\top B^\top + \Psi$, which is a rank-1 update to Ψ. Hoff and Niu [2012] discusses extensions to a rank-k update in which B will be indexed from 1 to k. Estimation is completed through either the expectation-maximization (EM) algorithm or Gibbs sampling. Our approach in Chapter 3 uses Gibbs sampling for full probabilistic inference.

1.5 Contributions and Dissertation Outline

Chapter 2 develops a Bayesian model for multidimensional and longitudinal functional data analysis. We propose a computationally efficient nonparametric Bayesian method to simultaneously smooth observed data, estimate conditional functional means and functional covariance surfaces. Statistical inference is based on Monte Carlo samples from the posterior measure through adaptive blocked Gibbs sampling. Several operative characteristics associated with the proposed modeling framework are assessed comparatively in a simulated environment. We illustrate the application of our work in two case studies. The first case study involves age-specific fertility collected over time for various countries. The second case study is an implicit learning experiment in children with Autism Spectrum Disorder (ASD).

Chapter 3 proposes a Bayesian covariance regression model for functional data, providing joint inference for both the conditional mean and covariance functions. Our work hinges on basis expansions of both the functional evaluation domain and covariate space, to define flexible non-parametric forms of dependence. To aid interpretation, we develop novel low-dimensional summaries, which indicate the degree of covariate-dependent heteroschedasticity. For illustration, our modeling framework is applied to two case studies, aiming to provide novel insight in brain imaging. The first case study evaluates a functional biomarker of neural development in children with autism spectrum disorder, and the second case study explores the relationship between sleep patterns, age, and hypertension.

Chapter 4 explores Bayesian approaches to model region-referenced functional data - multivariate functional data observed over regional subunits. The proposed methods identify

data-driven interpretable marginal functional and regional basis functions. Prior structure is imposed to encourage smoothness in the functional domain and weaken the influence of superfluous basis functions. The proposed methods incorporate scalar covariates, enabling joint mean and covariance estimation. In addition, we show how pivotal discrepancy measures can be used to assess structural covariance assumptions and aid in model selection. The proposed methods are applied to study electroencephalography data from children with autism spectrum disorder in a resting-state experiment.

In summary, the three chapters describe approaches to rigorously quantify uncertainty and perform adaptive regularization in functional data settings. Chapter 2 develops a probabilistic extension to a recently proposed methods for structured functional data, ensuring interpretable summaries for age-specific fertility rates and EEG data in an implicit learning study of children with autism spectrum disorder. Chapter 3 proposes nonparametric Bayesian methods for quantifying heterogeneity, accounting for vector-valued exogenous covariates. Chapter 4 develops two probabilistic extensions of recently proposed models for analyzing region-referenced functional data. Particular emphasis is placed on assessing implicit structural assumptions and model adequacy. Chapter 5 concludes with a discussion of the proposed methods and outlines future methodological developments that could follow this work.

CHAPTER 2

Bayesian Analysis of Longitudinal and Multidimensional Functional Data

Multi-dimensional functional data arises in numerous modern scientific experimental and observational studies. In this article, we focus on longitudinal functional data, a structured form of multidimensional functional data. Operating within a longitudinal functional framework we aim to capture low dimensional interpretable features. We propose a computationally efficient nonparametric Bayesian method to simultaneously smooth observed data, estimate conditional functional means and functional covariance surfaces. Statistical inference is based on Monte Carlo samples from the posterior measure through adaptive blocked Gibbs sampling. Several operative characteristics associated with the proposed modeling framework are assessed comparatively in a simulated environment. We illustrate the application of our work in two case studies. The first case study involves age-specific fertility collected over time for various countries. The second case study is an implicit learning experiment in children with autism spectrum disorder.

2.1 Introduction

We investigate Bayesian modeling and inference for longitudinal functional data, conceptualized as functional data observed repeatedly over a dense set of longitudinal time-points. A typical dataset would contain n patients observed over the course of multiple visit times, with each visit contributing a functional datum. Thus, for patient i we would record the outcome $y_i(s,t)$, where s is the visit time and t is the functional argument. In this setting it is reasonable to expect non-trivial correlations between functions from one visit time to

another. Therefore, appropriate modeling of this dependence pattern would be critical for the validity of statistical inference. This manuscript outlines a flexible Bayesian framework for the estimation of the functional mean structure, possibly dependent on a set of time-stable covariates, as well as an adaptive regularization framework for the estimation of the covariance operator of $y_i(s,t)$ and its eigenstructure.

The frequentist analysis of longitudinal functional data is a mature field. In particular, semiparametric modeling strategies, depending on the mixed effects modeling framework, have been proposed by Di et al. [2009] in the context of hierarchical functional data, and Greven et al. [2010] for longitudinal functional data. Important generalizations have been introduced by Chen and Müller [2012], through use structured functional principal components analysis (FPCA), with more parsimonious representations introduced by Park and Staicu [2015a], and Chen et al. [2017] in the more general context of function-valued stochastic processes. The appealing nature and flexibility of structured FPCA modeling strategies has seen the application and extension of these methods to challenging scientific problems ranging from functional brain imaging (Hasenstab et al. 2017, Scheffler et al. 2020a), to the exploration of complex data from wearable devices (Goldsmith et al. 2016).

The vast majority of approaches based on FPCA, generally focus on point estimation from a frequentist perspective, and do not provide reliable uncertainty quantification without bootstrapping. The very application of the bootstrap methodology to structured functional data has not been the subject of rigorous investigation. The literature, in fact, is ambiguous on the handling of the many tuning parameters, typical of structured FPCA models. Although there are some consistency results regarding the bootstrap for functional data (Cuevas et al. 2006, Ferraty et al. 2010), the procedure is relatively underdeveloped for hierarchical data (Ren et al. 2010).

Bayesian methods in functional data analysis define a straightforward mechanism for uncertainty quantification. This appealing inferential structure comes, however, at the cost of having to specify a full probability model, and priors with broad support on high dimensional spaces (Shi and Choi [2011], Yang et al. [2016a], Yang et al. [2017]). In hierarchical and multidimensional functional data settings, starting from the seminal work of Morris et al. [2003b],

and recent extensions in Lee et al. [2019b], the prevalent strategy has been to work within the framework of basis function transforms, defining flexible mixed effect models at the level of the basis coefficients (Morris et al. [2003b], Baladandayuthapani et al. [2008], Zhang et al. [2016]). The resulting functional mixed effects models, like their finite dimensional counterpart, require a certain degree of subject matter expertise in the definition of random effects and their covariance structure [Morris et al., 2011, Morris and Carroll, 2006].

This manuscript aims to merge the appealing characterization of longitudinal functional data through FPCA decompositions (Chen and Müller 2012, Park and Staicu 2015a, Chen et al. 2017), with flexible probabilistic representations of the classical Karhunen-Loéve expansion of square integrable random functions. Our work builds on the ideas of Suarez and Ghosal [2017] and Montagna et al. [2012], who adapted the regularized product Gamma prior for infinite factor models of Bhattacharya and Dunson [2011], to the analysis of random functions. Extensions of this framework to the longitudinal functional setting are discussed in Section 2.2. In Section 2.3 we discuss prior distributions and ensuing implications for the covariance operator. A comprehensive framework for posterior inference is discussed in Section 2.4. Section 2.5 contains a comparative simulation study. Finally, in Section 2.6 we discuss the application of our proposed methodology to two case studies. The first case study explores age-specific fertility dynamics in the global demographic study conducted by the Max Plank Institute and the Vienna Institute of Demography (HFD 2019). While purely illustrative, this data allows for a direct comparison with the original analysis of Chen et al. [2017]. The second case study, involves the analysis of electroencephalogram (EEG) data from an investigation of implicit learning in children with autism spectrum disorder (ASD) (Jeste et al. 2015). The main interest in both case studies is modeling and interpreting the longitudinal component.

2.2 A Probability Model for Longitudinal Functional Data

Let $y_i(s,t)$ denote the response for subject i, $(i = 1, \ldots, n)$, at longitudinal time $s \in \mathcal{S}$ and functional time $t \in \mathcal{T}$, where $\mathcal{S}$ and $\mathcal{T}$ are compact subspaces of $\mathbb{R}$. Furthermore,

for each subject, assume we observe a time-stable d-dimensional covariate $\boldsymbol{x}_i \in \mathbb{R}^d$. In practice, we only obtain observations $y_i(s_j, t_k)$ at discrete sampling locations $(s_j, t_k) \in \mathcal{S} \times \mathcal{T}$, $j = 1, \ldots, n_i^s$, $k = 1, \ldots, n_i^t$. However, in subsequent developments, we maintain the lighter notation $y_i(s, t)$ without loss of generality.

Let $f_i(s, t)$ be a Gaussian Process (GP) with mean $\mathbb{E}\{f_i(s, t)\} = \mu(\boldsymbol{x}_i, s, t)$ and covariance kernel $\text{Cov}\{f_i(s, t), f_i(s', t')\} = K\{(s, t), (s', t')\}$. A familiar sampling model for $y_i(s, t)$ assumes:

$$y_i(s, t) = f_i(s, t) + \epsilon_i(s, t), \;\; \epsilon_i(s, t) \overset{iid}{\sim} N(0, \varphi^2); \tag{2.1}$$

where $\varphi^2 > 0$ is the overall residual variance. Given a set of suitable basis functions $b_m^{(1)}(s) : \mathcal{S} \to \mathbb{R}$, $(m = 1, 2, \ldots p_1)$, and $b_\ell^{(2)}(t) : \mathcal{T} \to \mathbb{R}$, $(\ell = 1, 2, \ldots p_2)$, and a set of random coefficients θ_{iml}, the prior for the underlying signal $f_i(s, t)$ is constructed through a random tensor product expansion, so that

$$f_i(s, t) \;\; = \;\; \textstyle\sum_{m=1}^{p_1} \sum_{l=1}^{p_2} b_m^{(1)}(s) b_l^{(2)}(t) \theta_{iml}. \tag{2.2}$$

Since the truncation values p_1 and p_2 may be large to ensure small bias in the estimation of the true $f_i(s, t)$, we follow Bhattacharya and Dunson [2011] and project the basis coefficients on a lower dimensional space.

Let $\Theta_i = \{\theta_{im\ell}\} \in \mathbb{R}^{p_1 \times p_2}$ be the matrix of basis coefficients for subject i. After defining loading matrices $\Lambda \in \mathbb{R}^{p_1 \times q_1}$, $(q_1 \ll p_1)$, and $\Gamma \in \mathbb{R}^{p_2 \times q_2}$, $(q_2 \ll p_2)$, and a latent matrix of random scores $\boldsymbol{\eta}_i \in \mathbb{R}^{q_1 \times q_2}$, we assume

$$\Theta_i = \Lambda \boldsymbol{\eta}_i \Gamma^\top + \boldsymbol{\zeta}_i, \;\; \text{vec}(\boldsymbol{\zeta}_i) \sim \mathcal{N}(0, \Sigma); \tag{2.3}$$

where Σ is taken to be diagonal. The foregoing construction has connections with factor analysis. In fact, vectorizing Θ_i we obtain

$$\text{vec}(\Theta_i) = (\Gamma \otimes \Lambda)\text{vec}(\boldsymbol{\eta}_i) + \text{vec}(\boldsymbol{\zeta}_i);$$

which resembles the familiar $(q_1 \times q_2)$ latent factor model, with loading matrix $\Gamma \otimes \Lambda$ and latent factors $\text{vec}(\boldsymbol{\eta}_i)$. Differently from standard latent factor models, our use of a Kronecker product representation for the loading matrix introduces additional structural assumptions about $\text{Cov}(\Theta_i)$, and the ensuing form of the covariance kernel $K\{(s,t),(s',t')\}$.

More precisely, assuming $\text{Cov}(\boldsymbol{\eta}_i) = H$, the marginal covariance of Θ_i takes the form

$$\text{Cov}(\Theta_i) = (\Gamma \otimes \Lambda) H (\Gamma \otimes \Lambda)^\top + \Sigma = \Omega. \tag{2.4}$$

Furthermore, defining $B_1(s) = \{b_1^{(1)}(s), \ldots, b_{p_1}^{(1)}(s)\}^\top$ and $B_2(t) = \{b_1^{(2)}(t), \ldots, b_{p_2}^{(2)}(t)\}^\top$, the model in (2.3) induces the following representation for the covariance kernel $K\{(s,t),(s',t')\}$, s.t.

$$K\{(s,t),(s',t')\} = \{B_1(s) \otimes B_2(t)\}\, \Omega\, \{B_1(s') \otimes B_2(t')\}^\top. \tag{2.5}$$

The low-rank structure of Ω in (2.4), depends on the number of latent factors q_1 and q_2 in the quadratic form $(\Gamma \otimes \Lambda)^\top H (\Gamma \otimes \Lambda)$. Rather than selecting the number of factors a priori, in Section 2.3 we introduce prior distributions encoding rank restrictions through continuous stochastic regularization of the loading coefficient's magnitude. Additional structural restrictions may ensue from specific assumptions about the latent factors' covariance H. Specifically, setting $H = I_{q_1 q_2}$ leads to strong covariance separability of the longitudinal and functional dimensions. A more flexible covariance model hinges on the notion of weak-separability (Lynch and Chen 2018a). This is achieved by evaluating $H = \text{diag}(h_1, \ldots, h_{q_1 q_2}) > 0$.

Finally, let $\boldsymbol{x}_i$ be a d-dimensional time-stable covariate for subject i. Dependence of the longitudinal functional outcome $y_i(s,t)$ on this set of predictors is conveniently introduced through the prior expectation of $\boldsymbol{\eta}_i$. More precisely, let $\boldsymbol{\beta}$ be a $d \times q_1 q_2$ matrix of regression coefficients, we assume

$$\text{vec}(\boldsymbol{\eta}_i) \sim N(\boldsymbol{\beta}^\top \boldsymbol{x}_i,\, H),$$

which implies the following marginal mean structure for $y_i(s,t)$,

$$\mathbb{E}\{y_i(s,t)\} = \mu(\boldsymbol{x}_i, s, t) = \{B_1(s)\Gamma \otimes B_2(t)\Lambda\}\,\boldsymbol{\beta}^\top \boldsymbol{x}_i. \tag{2.6}$$

The model in (2.1), together with the sandwich factor construction in (2.3) defines a probabilistic representation of the product FPCA decomposition in Chen et al. [2017]. An intuitive parallel is introduced in Section 2.3, and a technical discussion is provided in the accompanying web-based supplementary document (Appendix 2A). Differently from Chen et al. [2017], we propose model-based inference through regularized estimation based on the posterior measure.

2.3 Rank Regularization and Prior Distributions

The selection of prior distributions for all parameters introduced in Section 2.2 is guided by the following considerations. Let γ_{lj} and λ_{mk} be specific entries in the loading matrices Γ and Λ in (2.3) respectively. Defining $\psi_j(s) = \sum_{l=1}^{p_1} \gamma_{lj} b_l^{(1)}(s)$ and $\phi_k(t) = \sum_{m=1}^{p_2} \lambda_{mk} b_m^{(2)}(t)$, we may expand $f_i(s,t)$ as follows:

$$\begin{aligned} f_i(s,t) &= \sum_{j=1}^{q_1}\sum_{k=1}^{q_2} \psi_j(s)\phi_k(t)\eta_{ijk} + r_i(s,t), \\ r_i(s,t) &= \sum_{j=1}^{p_1}\sum_{k=1}^{p_2} b_j^{(1)}(s) b_k^{(2)}(t)\zeta_{ijk}. \end{aligned}$$

The first component in the expression for $f_i(s,t)$ describes a mechanism of random functional variability which depends on the tensor combination of q_1 and q_2 data-adaptive basis functions $\psi_j(s)$ and $\phi_k(t)$ respectively, and $q_1 \times q_2$ basis coefficients η_{ijk}. Given q_1 and q_2, any residual variability is represented in the random function $r_i(s,t)$. When $\psi_j(s)$ and $\phi_k(t)$ are chosen to be eigenfunctions of the marginal covariance kernels in s and t, this representation is essentially equivalent to the product FPCA construction of Chen et al. [2017].

Statistical inference for FPCA constructions, commonly selects a small number of eigenfunctions on the basis of empirical considerations. Here we take an adaptive regularization

approach, choose q_1 and q_2 relatively large, and assume the variance components in the priors for Λ and Γ to follow a modified multiplicative gamma process prior (MGPP) Bhattacharya and Dunson [2011] Montagna et al. [2012].

Let λ_{mk} denote the (m,k) entry of Λ. The modified MGPP is defined by setting

$$\begin{aligned}
\lambda_{mk} &\sim N(0, \rho_{1mk}^{-1}\tau_{1k}^{-1}), \quad \rho_{1mk}^{-1} \sim \text{Gamma}(\nu_1/2, \nu_1/2), \\
\tau_{1k} &= \textstyle\prod_{v=1}^{k} \delta_{1v}, \quad \delta_{11} \sim \text{Gamma}(a_{11}, 1), \\
\delta_{1v} &\sim \text{Gamma}(a_{12}, 1)\mathbb{1}(\delta_{1v} > 1), \text{ for } v \geq 2; \;\; k = 1, 2, \ldots, q_1.
\end{aligned} \tag{2.7}$$

Using the "rate" parameterization for Gamma distributions (i.e., if $a \sim \text{Gamma}(b, c)$, then $\mathbb{E}(a) = bc$), this prior is designed to encourage small loadings in Λ as the column index increases. In the original formulation of Bhattacharya and Dunson [2011] and Montagna et al. [2012], choosing $a_{12} > 1$, insures stochastic ordering of the prior precision, in the sense that $E(\tau_{1k}) < E(\tau_{1(k+1)})$, for any $k = 1, 2, \ldots, (q_1 - 1)$. In our setting, we require the more stringent probabilistic ordering $Pr(\tau_{1k} < \tau_{1(k+1)}) = 1$, by assuming $\delta_{1v} > 1$, which results in a more stable and efficient Gibbs sampling scheme. Analogous regularization over the columns of Γ is achieved by setting:

$$\begin{aligned}
\gamma_{lj} &\sim N(0, \rho_{2lj}^{-1}\tau_{2l}^{-1}), \quad \rho_{2lj}^{-1} \sim \text{Gamma}(\nu_2/2, \nu_2/2) \\
\tau_{2l} &= \textstyle\prod_{v=1}^{l} \delta_{2v}, \quad \delta_{21} \sim \text{Gamma}(a_{21}, 1), \\
\delta_{2v} &\sim \text{Gamma}(a_{22}, 1)\mathbb{1}(\delta_{2v} > 1), \text{ for } v \geq 2; \;\; l = 1, 2, \ldots, q_2.
\end{aligned} \tag{2.8}$$

Adaptive shrinkage is induced by placing hyper-priors on a_{11}, a_{12}, a_{21}, and a_{22}, such that

$$a_{11}, a_{21} \overset{ind}{\sim} \text{Gamma}(r_1, 1), \;\; a_{12}, a_{22} \overset{ind}{\sim} \text{Gamma}(r_2, 1).$$

The model is completed with priors on residual variance components and regression coeffi-

cients. Specifically, conditionally conjugate priors are placed on the diagonal elements of Σ and H, respectively, as well as the residual variance φ, such that:

$$\sigma_j^{-1} \sim \text{Gamma}(a_\sigma, b_\sigma), \;\; h_j^{-1} \sim \text{Gamma}(a_h, b_h), \;\; \varphi^{-1} \sim \text{Gamma}(a_\varphi, b_\varphi).$$

Finally, we induce a Cauchy prior for the regression coefficients matrix $\boldsymbol{\beta}$ as in Montagna et al. [2012]. Denoting with $\beta_{j\ell}$ the (j, ℓ) entry of $\boldsymbol{\beta}$, we assume

$$\beta_{j\ell} \sim N(0, \omega_{j\ell}), \;\; \omega_{j\ell}^{-1} \sim \text{Gamma}(1/2, 1/2); \quad \ell = 1, \ldots, q_1 q_2, \quad j = 1, \ldots, d.$$

In summary, our approach starts with the projection of the observed data onto a set of known basis functions in (2.2). This initial projection is similar to the interpolation or smoothing step commonly used in functional data analysis (Chen et al. [2017], Morris and Carroll [2006]). The basis coefficients Θ_i are assumed to arise from the latent factor model in (2.3), resulting in the weakly separable covariance model in (2.4) and (2.5). Finally, the MGPP priors in (2.7) and (2.8), alllow for adaptive regularization of the covariance operator. The mean structure is made dependent on a set of time stable covariates through a varying coefficient model in (2.6).

2.4 Posterior Inference

Posterior simulation through Markov chain Monte Carlo is relatively straightforward, after selection of an appropriate basis transform and truncation of Γ and Λ to include $q_1 \ll p_1$ and $q_2 \ll p_2$ columns respectively. The use of conditionally conjugate priors allows for simple Gibbs transitions for all parameters, with the exception of the shrinkage parameters $a_{11}, a_{12}, a_{21},$ and a_{22}, which are updated via a Metropolis-Hastings step. A detailed description of the proposed algorithm is reported in Appendix 2B. We note that the decomposition of $\text{Cov}(\Theta_i)$ in (2.4) may not be unique. However, from a Bayesian perspective, one does not require identifiability of the loading elements for the purpose of covariance estimation.

Direct inference for $K\{(s,t),(s',t')\}$ and its functionals may be achieved by post-processing Monte Carlo draws from the posterior $p(\Omega \mid \boldsymbol{y})$ and evaluating the covariance function over arbitrarily dense points $\boldsymbol{t}^* \coloneqq (t_1^*, \ldots, t_{w1}^*)^\top \in \mathcal{T}$ and $\boldsymbol{s}^* \coloneqq (s_1^*, \ldots, s_{w2}^*)^\top \in \mathcal{S}$ using (2.5). Analogously, given samples from $p(\boldsymbol{\beta} \mid \boldsymbol{y})$, inference about the mean structure is achieved by evaluating $\mu(\boldsymbol{x}_i, s, t)$ over $\boldsymbol{s}^*$ and $\boldsymbol{t}^*$ using the expansion in (2.6).

Some useful posterior summaries may be obtained through marginalization. We define marginal covariance functions $K_{\mathcal{T}}(t,t')$ and $K_{\mathcal{S}}(s,s')$ as follows:

$$K_T(t,t') \;=\; \int_{\mathcal{S}} K\{(s,t)(s,t')\}ds, \qquad K_S(s,s') \;=\; \int_{\mathcal{T}} K\{(s,t)(s',t)\}dt. \tag{2.9}$$

Intuitively, $K_S(\cdot)$ and $K_T(\cdot)$ summarize patterns of functional co-variation along a specific coordinate, and their lower-dimensional posterior summaries may be obtained through functional eigenanalysis as in Chen et al. [2017]. We outline details on extracting lower dimensional summaries of the marginal covariance functions without computing $K\{(s,t),(s',t')\}$, $K_{\mathcal{T}}(t,t')$, or $K_{\mathcal{S}}(s,s')$ in Supplemental Appendix 2F. Simultaneous credible intervals for all functions of interest are easily obtained from Monte Carlo samples, by applying the methodology discussed in Crainiceanu et al. [2007] and Baladandayuthapani et al. [2005].

Specifically, M Monte Carlo draws from a posterior function of interest, say $g(\tau)$, are used to estimate the posterior mean $\widehat{g}(\tau)$, and standard deviation $\sqrt{\widehat{var}\{\widehat{g}(\tau)\}}$. Assuming approximate normality of the posterior distribution, we derive the $(1-\alpha)$ quantile c_α of the pivotal quantity

$$\max_{\tau} \left| \frac{g^{(i)}(\tau) - \widehat{g}(\tau)}{\sqrt{\widehat{var}\{\widehat{g}(\tau)\}}} \right|, \quad i = 1, \ldots, M.$$

An approximate simultaneous $(1-\alpha)$ posterior band can then be constructed as a hyper-rectangular region over τ: $\left[\widehat{g}(\tau) \pm c_\alpha \cdot \sqrt{\widehat{var}\{\widehat{g}(\tau)\}}\right]$. More general simultaneous bands have been proposed by Krivobokova et al. [2010], but are not implemented in this manuscript.

The proposed modeling framework relies on a specific basis transform strategy. While the literature has suggested the use of zero-loss transforms as a default option (Morris et al. 2003b, Lee et al. 2019a), we find that it is not uncommon to observe some sensitivity to the

number of basis functions used in the initial projection. Furthermore, the choice of more parsimonious designs, when warranted by the application, may lead to important gains in computational and estimation efficiency. Model flexibility is governed by choice of (p_1, p_2), as the number of smoothing basis functions, and (q_1, q_2) as the number of latent facors. Due to the adaptive rank regularizing prior, q_1 and q_2 should be chosen as large as possible. In practice (q_1, q_2) are chosen as fraction of (p_1, p_2).

Our simulation studies, (supplementary material, Appendix 2C), demonstrate that point estimates and uncertainty of mean and covariance functions are generally insensitive to choice of p_1 and p_2, provided q_1 and q_2 are large. Some sensitivity is, however, observed in the posterior estimate of the residual error φ^2. An alternative method is to simply rely on the minimization of information criteria. In this paper we consider simple versions of the deviance information criterion (DIC), and two versions of the Bayesian information criteria (BIC & BIC_h Delattre et al. [2014]). Our simulations studies, (supplementary material, Appendix 2C), indicate that the proposed information criteria perform well in selecting an adequate number of basis functions.

From a computational perspective, the most time consuming steps in the Gibbs sampling algorithm are the Cholesky decompositions used in updating Θ_i and η_i, requiring $O\left(p_1^3 p_2^3\right)$ and $O\left(q_1^3 q_2^3\right)$ floating point operations respectively. Therefore, scalability of naïve Gibbs sampling is a potential issue for very large samples and/or very long longitudinal or functional evaluation domains. In these cases, adapting the estimation approach of Morris and Carroll [2006], is easily implemented, by treating the estimation of Θ_i as a pre-processing step, and considering 2.3 as the sampling model. For big data applications, other analytical approximations to the posterior measure are also accessible, e.g. INLA (Rue et al. [2009]).

2.5 A Monte Carlo Study of Operating Characteristics

We performed a series of numerical experiments aimed at evaluating the estimation performance for both the functional mean and covariance. We study three simulation scenarios, including two weakly separable kernels (cases 1 and 2) and one non-separable covariance

function (case 3). Specifically, for $s \in [0,1]$ and $t \in [0,1]$, we take:

1. $K_{\mathcal{S}}(s,s') = \sum_{j=1}^{2} \lambda_j \psi_j(s)\psi_j(s')$, with eigenvalues $\lambda_j = \frac{1}{j^2\pi^2}$ and eigenfunctions $\psi_j(s) = \sqrt{2}\sin(j\pi s)$, $K_{\mathcal{T}}(t,t') = \sigma^2\left(1 + \frac{\sqrt{3}|t-t'|}{\rho}\right)\exp\left(-\frac{\sqrt{3}|t-t'|}{\rho}\right)$, in the Matèrn class, and mean $\mu(s,t) = \sqrt{\frac{1}{5\sqrt{s+1}}}\sin(5t)$.

2. $K_{\mathcal{S}}(s,s') = \sum_{j=1}^{2} \lambda_j \psi_j(s)\psi_j(s')$, with eigenvalues $\lambda_j = \frac{1}{(j-1/2)^2\pi^2}$ and eigenfunctions $\psi_j(s) = \sqrt{2}\sin((j-1/2)\pi s)$, $K_{\mathcal{T}}(t,t') = \sum_{k=1}^{50} \lambda_k \phi_k(t)\phi_k(t')$, with $\lambda_k = k^{-2\alpha}$ and $\phi_k(t) = \cos(k\pi t)$, and mean $\mu(s,t) = 5\sqrt{1-(s-.5)^2-(t-.5)^2}$.

3. $K((s,t),(s',t')) = \frac{1}{(t-t')^2+1}\exp\left\{-\frac{(s-s')^2}{(t-t')^2+1}\right\}$, stationary non-separable (Gneiting 2002), and mean $\mu(s,t) = \sqrt{1+\sin(\pi s)+\cos(\pi t)}$.

Scenario 1 combines a simple Matèrn class pattern on the time-domain t with a more complex oscillatory dependence pattern for the functional domain s. Scenario 2 includes an oscillatory pattern in both s and t. Finally, scenario 3, while defining simple parametric dependence in both longitudinal and functional times, is not weakly separable, allowing for comparisons on misspecified models.

We consider estimation of the mean, covariance, marginal covariance functions, and the associated two principal eigenfunctions. Each simulation includes 1,000 Monte Carlo experiments. For each experiment, posterior estimates are based on 10,000 iterations of 4 independent Markov chains, after discarding 2,500 draws for burn-in. We compare estimation of covariance, marginal covariance functions, and associated two principal eigenfunctions to the respective estimates provided by the product FPCA (Chen et al. 2017), as well as finite-dimensional empirical estimates of the mean and covariance defined as by their vectorized sample counterparts. Estimates obtained with the product FPCA have data-type set to sparse and fraction of variance explained (FVE) threshold set to .9999.

All comparisons are based on the relative mean integrated squared error. For a function f with domain D and estimator $\hat{f}$, we define $RE(\hat{f},f) = \int_D\{\hat{f}(u)-f(u)\}^2du/\int_D f(u)^2du$. Note that D can be multi-dimensional and in practice the integral is replaced with a sum. Table 2.1 compares mean $\mu(s,t)$ and covariance $K\{(s,t),(s',t')\}$ estimation under the three

Table 2.1: Mean and covariance relative errors under the three simulation cases described in section 2.5. Bayes refers to the proposed method in this paper, product refers to the product decomposition Chen et al. [2017], and empirical refers to point-wise empirical estimation. Each case is repeated 1,000 times for sample sizes of $n = 30$ and $n = 60$. We report the 50th percentile of the relative error, with the numbers in the parantheses denoting the 10th and 90th percentiles of the relative error.

Case 1		Bayes	Product	Empirical
$n = 30$	$\mu(s,t)$	.014 (.005, .038)	.019 (.010, .044)	.019 (.010, .044)
	$K\{(s,t),(s',t')\}$	.062 (.023, .224)	.085 (.047, .200)	.151 (.097, .297)
$n = 60$	$\mu(s,t)$	.007 (.003, .019)	.010 (.005, .021)	.010 (.005, .021)
	$K\{(s,t),(s',t')\}$	.030 (.010, .097)	.057 (.038, .128)	.076 (.050, .151)
Case 2				
$n = 30$	$\mu(s,t)$	.024 (.007, .101)	.031 (.013, .118)	.031 (.013, .118)
	$K\{(s,t),(s',t')\}$	.039 (.011, .184)	.050 (.012, .202)	.067 (.030, .228)
$n = 60$	$\mu(s,t)$	.014 (.004, .054)	.017 (.007, .062)	.017 (.007, .062)
	$K\{(s,t),(s',t')\}$	.019 (.005, .091)	.024 (.007, .093)	.032 (.014, .106)
Case 3				
$n = 30$	$\mu(s,t)$	.155 (.046, .389)	.160 (.051, .393)	.160 (.051, .393)
	$K\{(s,t),(s',t')\}$	.051 (.016, .187)	.051 (.014, .183)	.067 (.023, .200)
$n = 60$	$\mu(s,t)$	.073 (.019, .216)	.076 (.021, .219)	.076 (.021, .219)
	$K\{(s,t),(s',t')\}$	.028 (.008, .091)	.027 (.007, .089)	.034 (.011, .099)

settings listed above. We find that estimates from each method improve in accuracy with increasing sample size ($n = 30, 60$), with the posterior and product FPCA showing greater accuracy than empirical approach in terms of covariance estimation. Similar findings characterize the estimation performance of all marginal covariance functions ($K_{\mathcal{S}}$, $K_{\mathcal{T}}$), and the associated two principal eigenfunctions ($\psi_i(s)$, $i = 1, 2$), and ($\phi_i(t)$, $i = 1, 2$). Detailed numerical results and extended simulations are reported in the web-based supplement, Appendix 2C.

In summary, we observe that posterior estimates are associated with similar, and potentially improved accuracy in the estimation of the mean and covariance functions, when compared with product FPCA. This similarity in estimation performance, provides some empirical assurances that the chosen probabilistic representation of structured covariance functions, and estimation based on adaptive shrinkage, maintains a data-adaptive behavior with good operating characteristics.

2.6 Case Studies

We illustrate the application of the proposed modeling frameworks in two case studies. The first dataset concerns fertility rate and age of mothers by country. The second case study focuses on functional brain imaging through EEG in the context of implicit learning in children with ASD.

2.6.1 Fertility rates

The Human Fertility Database (HFD 2019) compiles vital statistics to facilitate research on fertility in the past twentieth century and in the modern era. Age-specific fertility rates are available for 32 countries over different time periods. The age-specific fertility rate $ASFR(s,t)$ is defined as

$$ASFR(s,t) = \frac{\text{births during year } s \text{ given by women aged } t}{\text{person-years lived during year } s \text{ by women aged } t}.$$

The dataset was previously analyzed and interpreted in a longitudinal functional framework using the product FPCA (Chen et al. 2017). This section focuses on a comparative analysis of product FPCA and the proposed probability model.

We follow Chen et al. 2017, and consider $n = 17$ countries, with complete data for the time period 1951 to 2006, 44 functional time points (ages 12-55), and 56 longitudinal time points (years 1951 to 2006). Since these rates are population measurements, we expect the data to contain very little noise. We use cubic b-splines as our basis functions since the data look smooth with no sharp changes in fertility rate over year or age of mother (Supplemental Fig. 1) and consider $(p_1, p_2) = (22, 28)$ splines and $(q_1, q_2) = (11, 10)$ latent factors, selected by minimizing the DIC.

Longitudinal and aging dynamics are largely determined by their associated marginal covariance functions $K_{\mathcal{S}}(s,s')$ and $K_{\mathcal{T}}(t,t')$. Figure 2.1 displays the first three marginal eigenfunctions for age and calendar year. We include the 95% simultaneous credible bands (Crainiceanu et al. 2007) as well as estimates obtained via product FPCA. We note that

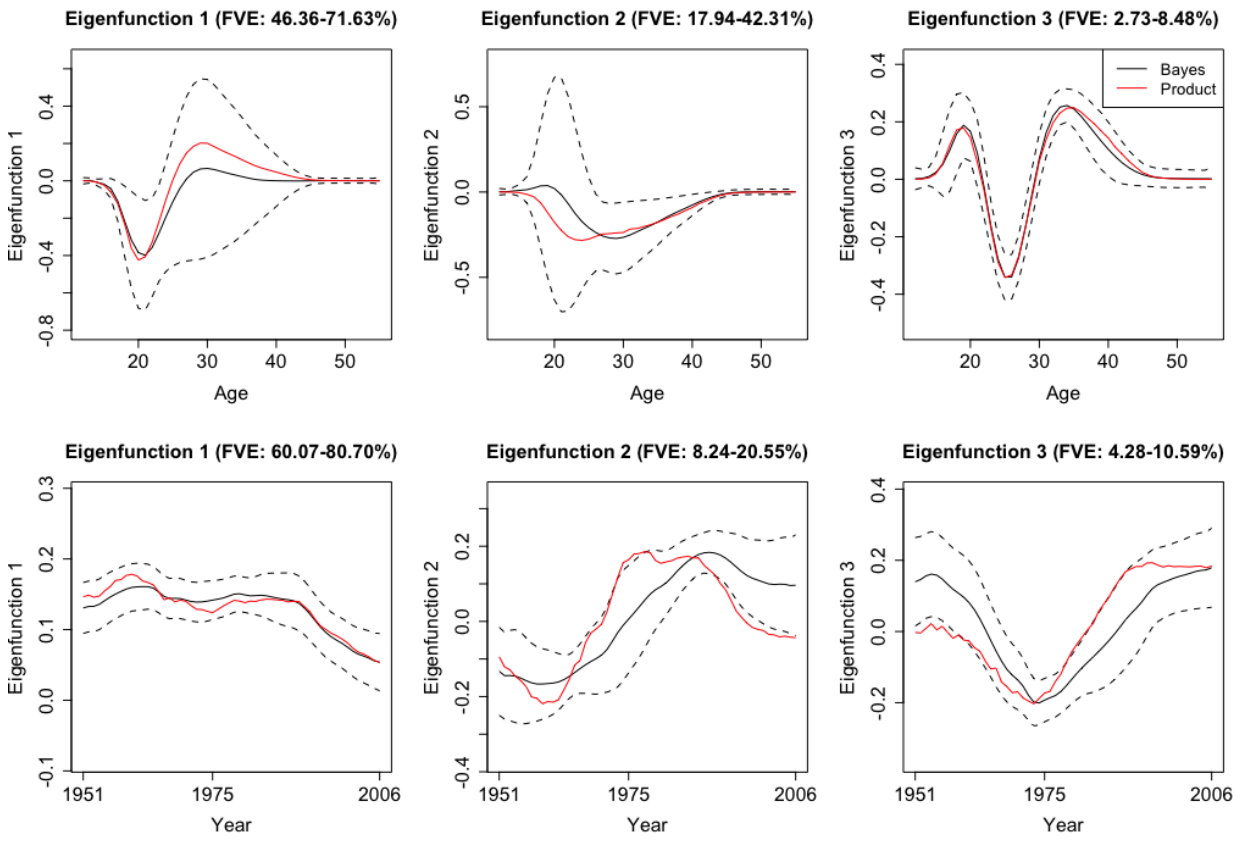


Figure 2.1: Age and calendar year marginal eigenfunctions. The above plots include the Bayesian posterior means, 95% credible bands, and the product FPCA marginal eigenfunctions.

Bayesian posterior mean eigenfunctions are qualitatively similar to the inferred product FPCA estimates, therefore warranting similar interpretations to the one originally offered by Chen et al. 2017.

In particular, the first marginal eigenfunction for age (Figure 2.1, left panel) can be interpreted as the indexing variability in young fertility before the age of 25, with the second marginal eigenfunction for age (Figure 2.1, central panel) indexing variability in fertility for mature age, between the ages of 20 and 40. As our modeling framework allows for rigorous uncertainty quantification in these posterior summaries, we note that the credible bands for the first and second eigenfunction are relatively wide, indicating that specific patterns should be interpreted with care. Examining directions of variance in fertility through the years, we note that the first marginal eigenfunction for year (Figure 2.1, left panel) is relatively

constant and can be interpreted as representing an overall "size-component" of fertility from 1951-2006. The second eigenfunction (Figure 2.1, central panel) defines a contrast of fertility in countries before and after 1975. For both the year and age coordinates the third marginal eigenfunctions capture a smaller fraction of the total variance and index higher patterns of dispersion at and around age 25 and at or around the year 1975.

We investigate sensitivity to the number of basis and latent factors considering four different models: model 1: $(p_1, p_2) = (44, 50)$, $(q_1, q_2) = (20, 20)$; model 2 $(p_1, p_2) = (44, 50)$, $(q_1, q_2) = (6, 6)$; model 3: $(p_1, p_2) = (16, 20)$, $(q_1, q_2) = (12, 12)$; and model 4: $(p_1, p_2) = (16, 20)$, $(q_1, q_2) = (6, 6)$. We also estimate the marginal covariance function with product FPCA using both the dense and sparse settings. Point estimate for $K_T(t, t')$ are reported in Figure 2.2. Comparing estimates within column (left and center panels), we assess sensitivity to a drastic reduction in the number of latent factors. Comparing estimates within row (left and center panels), we instead assess sensitivity to a drastic reduction in the number of basis functions. We note that the marginal age covariance function is relatively stable in all four settings. We contrast this relative robustness with estimates based on the product FPCA. In particular, sparse estimation using 10-fold cross-validation results in meaningfully diminished local features. A possible reason for the instability is due to the small sample size (n=17). In this example, Bayesian estimation is perhaps preferable, as adaptive penalization allows for stable estimates within a broad class of model specifications.

2.6.2 An EEG Study on Implicit Learning in Children with ASD

This analysis is motivated by a functional brain imaging study of implicit learning in young children with autism spectrum disorder (ASD), a developmental condition that affects an individual's communication and social interactions (Lord et al. 2000). Implicit learning is defined as learning without the intention to learn or without the conscious awareness of the knowledge that has been acquired. We consider functional brain imaging through EEG, an important and highly prevalent imaging paradigm aimed at studying macroscopic neural oscillations projected onto the scalp in the form of electrophysiological signals.

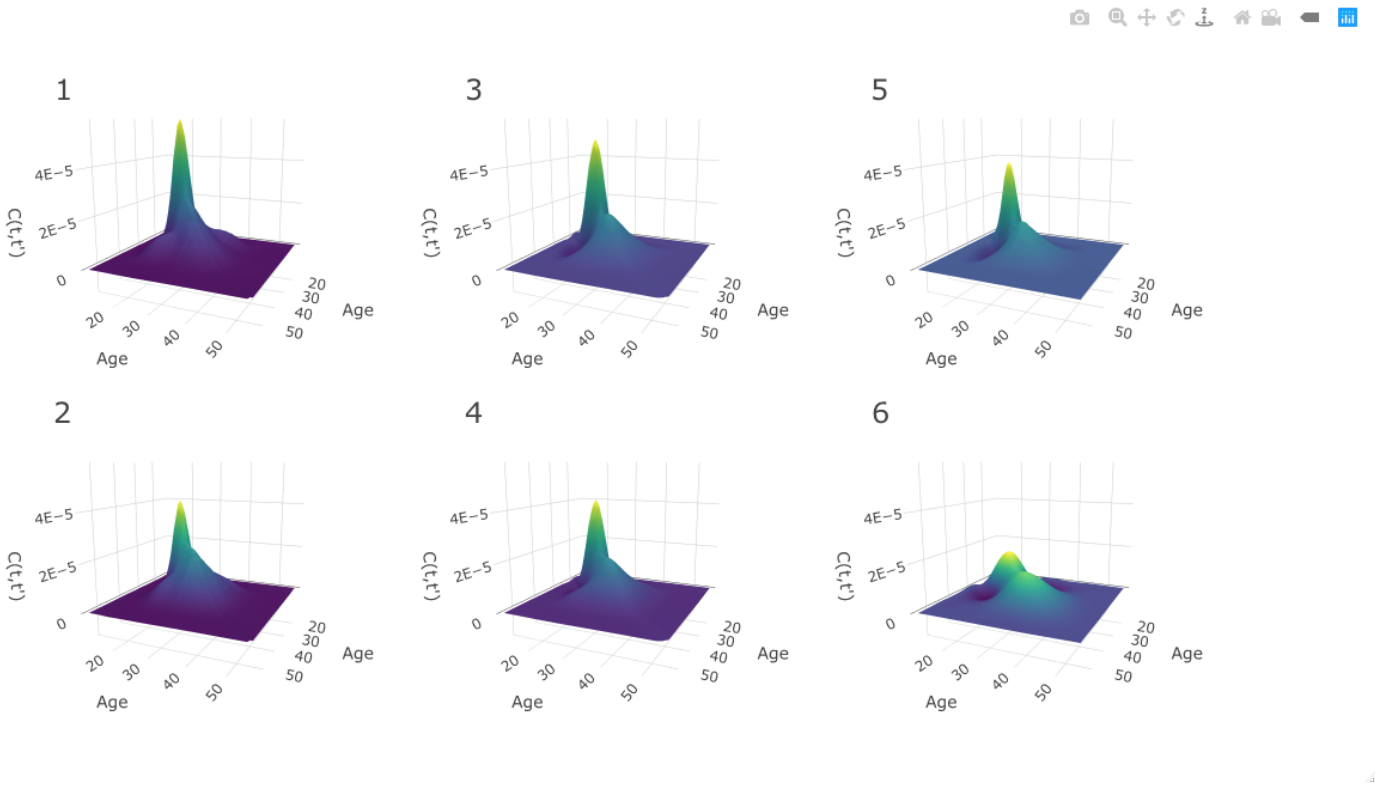


Figure 2.2: Sensitivity analysis for the marginal covariance function $K_{\mathcal{T}}(t, t')$ (HFD study). Panels (1,2,3,4) refer to posterior mean estimates obtained under different projections and numbers of latent factors (Specific details are provided in Section 2.6.1). Panels (5, 6) refer to product FPCA estimates obtained under dense (5) or sparse (6) settings.

This study, carried out by our collaborators in the Jeste laboratory at UCLA, targets the neural correlates of implicit learning in the setting of an event-related shape learning paradigm (Jeste et al. 2015). Children aged 2-6 years old with ASD were recruited through the UCLA Early Childhood Partial Hospitalization Program (ECPHP). Each participant had an official diagnosis of ASD prior to enrollment. Age-matched typically developing (TD) children from the greater Los Angeles area were recruited as controls. Six colored shapes (turquoise diamond, blue cross, yellow circle, pink square, green triangle, and red octagon) were presented one at a time in a continuous "stream" in the center of a computer monitor. There were three shape pairings randomized to each child. For instance, a pink square may always be followed by a blue cross. After the blue cross would come a new shape pair. Within a shape pair would constitute an "expected" transition and between shape pairs would constitute an "unexpected" transition. Each child would wear a 128-electrode Geodesic Sensor Net and observe the stream of shapes on the computer monitor. Each stimulus, or presentation of a single shape, is referred to as a trial, and can result in frequency-specific changes to ongoing EEG oscillations, which are measured as Event Related Potentials (ERPs).

Each waveform contains a phasic component called the P300 peak which represents attention to salient information. This phasic component is typically studied in EEG experiments and is thought to be related to cognitive processes and early category recognition (Jeste et al. 2015). We use the same post-processed data as in Hasenstab et al. [2017]. Namely, we consider 37 ASD patients and 34 TD patients using data from trials 5 through 60 and averaging ERPs in a 30 trial sliding window (Hasenstab et al. 2015). The sliding window enhances the signal to noise ratio at which the P300 peak locations can be identified for each waveform. Each waveform is sampled at 250 Hz resulting in 250 within-trial time points over 1000ms. Following Hasenstab et al. [2017], we reduce each waveform to a 140ms window around each P300 peak. This 140ms window results in 37 within-trial time points. We do not apply warping techniques because each within-trial curve is centered about the P300 peak. Our analysis focuses on condition differentiation, formally defined as the difference between the expected and subsequent unexpected condition. Modeling condition differentiation for

waveforms within a narrow window about the P300 peak over trials may give insights into learning rates for the ASD and TD groups. Thus, the main interest in this study is changes in condition differentiation over trials, and a longitudinal functional framework is required for statistical inference in this setting. Our analysis is based on the condition differentiation, averaged within subject over the four electrodes in the right frontal region of the brain. In summary, for each subject we consider $n_s = 56$ observations within trial, and $n_t = 37$ total trials.

We model the ASD and TD data cohorts separately, in order to estimate ERP time and trial covariance functions within group. All inference is based on a model with $p_1 = 20$, $p_2 = 56$, $q_1 = 10$, $q_2 = 28$, selected minimizing DIC. A comprehensive analysis is reported in the web-based supplement. Statistical inference is based on 50K MCMC posterior draws, after 20K burn-in. We considered relatively diffuse priors: $a_\sigma = b_\sigma = 0.5$, $a_h = b_h = 1$, $\nu_1 = \nu_2 = 1$, $r_1 = r_2 = 1$, and $a_\varphi = b_\varphi = 0.0001$. Results are relatively insensitive to these hyperparameter settings. The estimated mean surfaces for the two groups are plotted in Figure 2.3. The ASD group tends to have positive condition differentiation between trials 30 and 55, whereas the TD group tends to have positive condition differentiation in earlier trials. Positive condition differentiation is thought to be indicative of learning, so these results suggest that the TD group is learning at a faster rate than the ASD group. However, even though qualitatively the surfaces look very different, there is a substantial amount of heterogeneity in the subject-level data, resulting in broad confidence bands around the mean, and perhaps suggesting that differential patterns of condition differentiation between ASD and TD groups are best explored considering both the mean and the covariance structure.

Next we conduct an eigen-analysis of the covariance structure for both cohorts separately. Figure 2.4 plots eigenfunctions of the marginal covariances over ERP time and trials. Credible intervals are calculated following Crainiceanu et al. [2007]. We start by analyzing summaries indexing variability in ERP time. For both the TD and ASD cohorts, the first eigenfunction explains the vast majority of the marginal covariance (84%-88% in ASD, and 86%-90% in TD). In both groups this first eigenfunction is relatively flat and can be interpreted as representing variability in the overall level of condition differentiation within

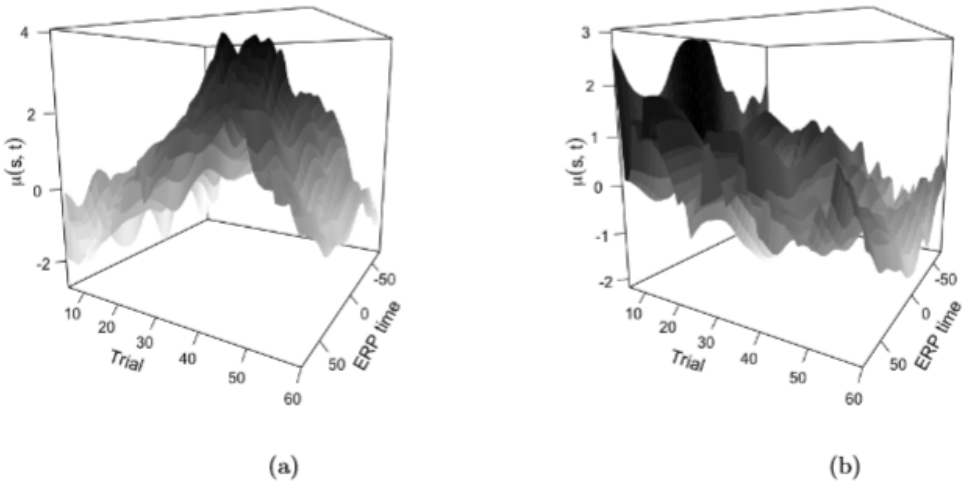
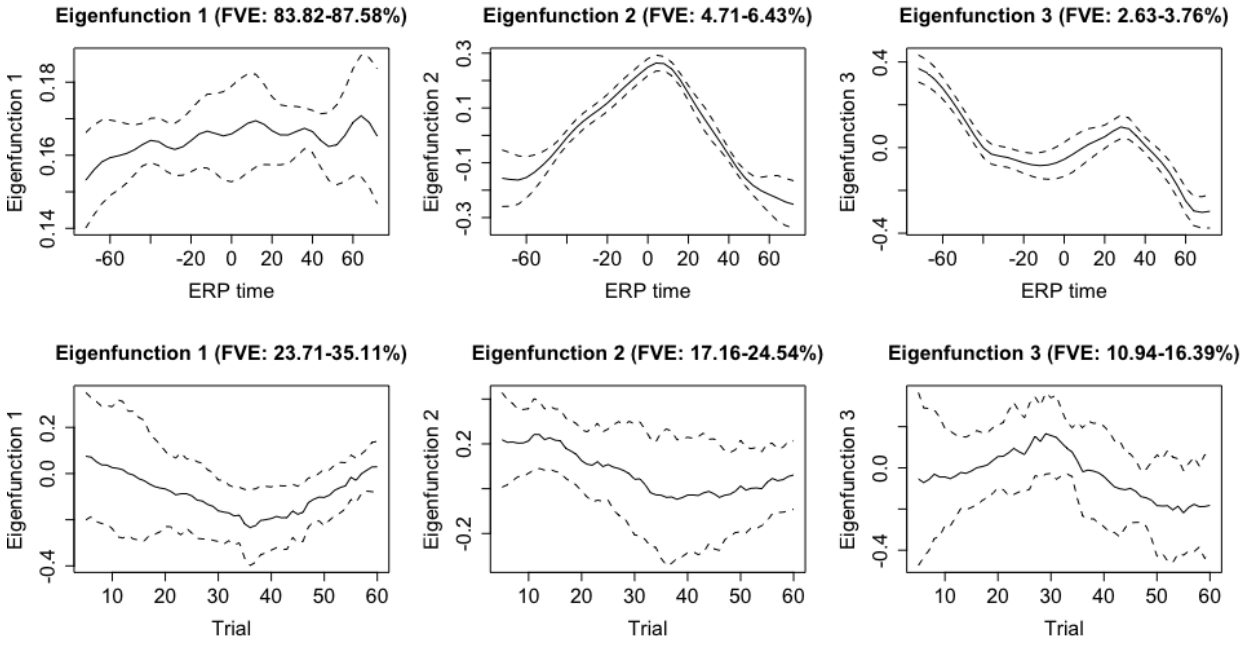


Figure 2.3: Posterior expected mean condition differentiation along trial and ERP time for the ASD (a) and the TD (b) cohorts.

a trial. The magnitude and shape of variation is comparable between TD and ASD children. Finer differences may be detected in the second and third eigenfunction, which further characterize variability in the shape of the ERP waveforms about the P300 peak. For both cohorts, however, these summaries represent only a small percentage of the variance in ERP waveform within trial.

Perhaps more interesting is an analysis of the marginal covariance across trial, as probabilistic learning patterns are likely to unfold with prolonged exposure to expected vs. unexpected shape pairings. For the ASD group, the first eigenfunction dips in an approximately quadratic fashion, suggesting enhanced variability in condition differentiation at around trial 35. Similarly, for the TD group, the first trial eigenfunction has a slight peak around trial 25. A possible interpretation of these covariance components relates to implicit learning, with higher variance in differentiation occurring earlier for TD than for ASD children. For both TD and ASD, the second eigenfunction across trials is interpreted as a contrast between high condition differentiation at early trials and low condition differentiation at later trials. Finally for the ASD cohort, the third eigenfunction exhibits a peak around trial 30. A possible interpretation would identify heterogeneity in the timing of learning, with some of the trajectories inducing variation in condition differentiation around trial 30, as opposed

ASD



TD

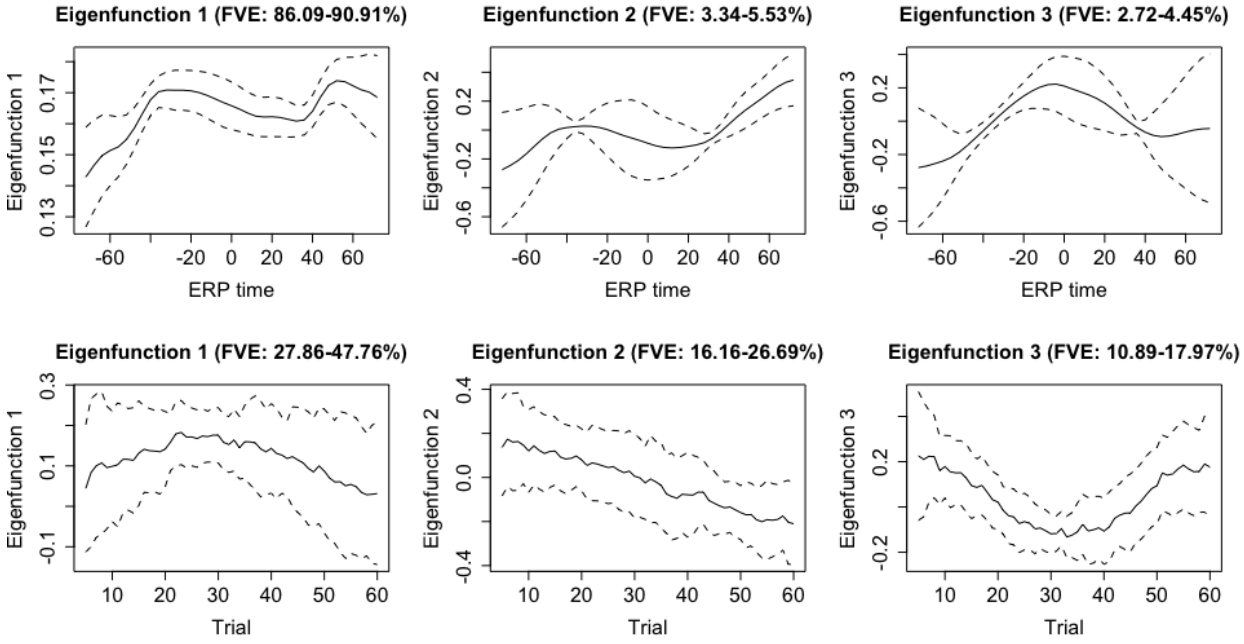


Figure 2.4: Marginal eigenfunctions with associated uncertainty for the ASD and TD groups. Solid black lines represent posterior means and dotted lines represent 95% simultaneous credible bands.

to the first eigenfunction identifying increased variance at around trial 35. Similarly for the TD group, the third trial eigenfunction has a dip around trial 35, indexing delayed increased variability in condition differentiation around trial 35.

2.7 Discussion

In this paper we provide a probabilistic characterization of longitudinal-functional data. As part of our work we propose a joint framework for the estimation of the mean or the regression function, and a flexible prior for covariance operators. Regularized estimation relies crucially on the projection of a set of basis coefficients onto a latent subspace, with adaptive shrinkage achieved via a broadly supported class of product Gamma priors. While we have not established theoretical results on posterior consistency, we have shown that the proposed framework exhibits competitive operating characteristics, when compared with alternative modeling strategies.

Importantly, uncertainty quantification, is achieved without having to rely on the asymptotic performance of bootstrap methods. From an applied perspective, analysts are charged with choosing the appropriate projection space. However, we see this as a feature rather than a problem, as different data scenarios may require and motivate the use of alternative basis systems. Because regularization is achieved jointly with estimation, inference is straightforward and does not need to account separately for the estimation of nuisance parameters or the choice of a finite number of eigenfunctions to use in a truncated version of the model, as is the case for FPCA-based methods.

Crucially, the level of flexibility afforded by the proposed method is most important when inference centers on both the mean and covariance structure. Simpler modeling strategies, e.g. the LFPCA approach of Greven et al. [2010], are likely to be more appropriate when the number of longitudinal observations is small, or if inference centers mostly on the mean structure. In a small simulation study (Appendix 2C) we found that the proposed approach performs similarly to LFPCA, even when data are generated from the latter scheme.

We have shown that posterior inference using MCMC is implemented in a relatively

straightforward fashion and need not rely on complicated posterior sampling strategies. When dealing with large data-sets, this naïve inferential strategy may not be appropriate. For example, in updating the basis coefficients Θ_i, the number of floating point operations grow at a cubic rate with respect to the dimensions of the spline bases. When naïve Gibbs is not scalable, (e.g .for very large samples or long evaluation domains), potentially promising acceleration strategies include the zero-loss projection approach of Morris and Carroll [2006], and adaptations of the INLA framework for approximate inference (Rue et al. [2009]).

From a modeling perspective, our probabilistic characterization of the longitudinal-functional covariance function is essentially equivalent to the weakly-separable model of Chen et al. [2017]. While more general than a strictly separable model, this strategy makes strong assumptions about the structure of a high-dimensional covariance operator. Testing strategies have been developed in the literature (Lynch and Chen 2018a). However, we find that a more natural approach to the problem is one of regularized estimation. In this setting, a possible extension of our modeling framework could include an embedding strategy for the regularization of a non-separable covariance operator towards a weakly separable one.

Software

Software in the form of an R package including complete documentation and a sample data set is available from https://github.com/jshamsho/LFBayes

Supplementary Material

Supplementary material is available online at [http://biostatistics.oxfordjournals.org].

Acknowledgments

Funding for the project was provided by the grant R01 MH122428-01 (DS, DT) from the National Institute of Mental Health.

CHAPTER 3

Bayesian Covariance Regression in Functional Data

Function on scalar regression models relate functional outcomes to scalar predictors through the conditional mean function. These analytical techniques find many uses in diverse applied domains ranging from medicine, to environmental science. With the exception of some recent contributions, many functional regression frameworks operate under the assumption of that covariate information does not affect patterns of covariation. In this article, we address this disparity, and develop a Bayesian functional regression model, providing joint inference for both the conditional mean and covariance functions. Our work hinges on basis expansions of both the functional evaluation domain and covariate space, to define flexible non-parametric forms of dependence. To aid interpretation, we develop novel low-dimensional summaries, which indicate the degree of covariate-dependent heteroschedasticity. For illustration, our modeling framework is applied to two case studies, aiming to provide novel insight in brain imaging. The first case study evaluates a functional biomarker of neural development in children with autism spectrum disorder, and the second case study explores the relationship between sleep patterns, age, and hypertension.

3.1 Introduction

Functional data analysis (FDA) is a broad collection of theory and methods designed to analyze conceptually infinite-dimensional data with smoothness assumptions. Functional principal components analysis (FPCA) [Wang et al., 2016] is ubiquitous in FDA, with applications ranging from biomedicine, gene expression data, environmental science, and many other scientific disciplines. Although the FPCA literature is growing rapidly to accommo-

date data arising from complex designs [Baladandayuthapani et al., 2008, Staicu et al., 2010, Greven et al., 2010, Zipunnikov et al., 2011, Park and Staicu, 2015b, Scheffler et al., 2020a], only a handful of methods allow incorporating exogenous covariate information in the covariance estimation or resulting FPCA. Cardot [2007] developed a nonparametric technique based on kernel smoothers to smooth covariance surfaces over a covariate in dense functional data settings. Jiang et al. [2010] extended the kernel smoothing technique to accommodate sparse functional data by employing conditional expectation Yao et al. [2005]. Xiao et al. [2015] use sandwich smoothing [Xiao et al., 2013] to extract age-adjusted patterns of variation in a repeated measures circadian rhythm study. Scheffler et al. [2020b] extend the concept of weak-separability to incorporate exogenous covariate information in a multidimensional functional setting.

In this paper we develop a Bayesian method to adjust covariances based on exogenous covariate information. We focus on the independent functional response case for simplicity although extensions to more general dependency structures are possible. Our proposed method makes several novel contributions to the existing literature of covariate-adjusted covariance modelling. The Bayesian paradigm offers a straightforward mechanism for quantifying uncertainty through posterior sampling. The alternative empirical methods must rely on the bootstrap, which has some issues in the context of functional data. Parametric bootstrapping does not account for uncertainty in selecting various tuning parameters or data-adaptive basis functions [Goldsmith et al., 2013]. Theoretical guarantees surrounding nonparametric bootstrapping are underdeveloped in the context of functional covariance regression.

Unlike previous research incorporating exogenous covariate information in the covariance function, the proposed method is based on an additive structure which comes with several advantages. The additive structure allows one to incorporate discrete covariates such as diagnostic status directly into the analysis without the need for stratification. Retaining the entire original sample is important for the data (as opposed to prior assumptions) to guide the analysis. For example, Suarez and Ghosal [2017] note that prior specification for within-subject random error is a sensitive choice since it partially determines the amount of

smoothing performed on the data. Stratifying the analysis by levels of a discrete covariate would limit the model's ability to learn posterior within-subject random error magnitudes and thus obscure the amount of smoothing to perform on the data. The additive structure also allows for implicit regularization assumptions. For example, deliberately removing interactions could induce meaningful regularization in high dimensional covariate settings. Model-based estimation also allows for the calculation of low dimensional summaries which aid in describted the amount of heterogeneity to a particular covariate. It's not clear how to quantify this quantity using existing methods.

The proposed method is classified in the fifth general approach from Greven and Scheipl [2017], which directly models the observed data and expands all model terms in suitable basis expansions. Methods in this general approach account for all error sources in subsequent inferences and accommodate sparse or irregularly sampled curves. We focus on dense functional data in this article but extension to more complicated grids is straightforward with basis expansions. We are aware of several other Bayesian methods in the fifth general approach including but not limited to Van Der Linde [2009], Thompson and Rosen [2008], Montagna et al. [2012], Goldsmith et al. [2015], Suarez and Ghosal [2017], and Kowal and Bourgeois [2020]. However, we note that the proposed method is the first to go beyond varying coefficient flexibility and model covariance functions in a regression setup.

The proposed method is closely related to the notion of regularized covariance estimation. As an early reference for regularized covariance estimation, Flury [1984] developed a method to estimate a common set of principal components across k groups. This concept was generalized by Franks and Hoff [2019], who use partial pooling to estimate a set of principal components across k groups. Fox and Dunson [2015] developed a Bayesian nonparametric method for estimating a time-varying covariance matrix through factor matrix products, where the loading of the factor matrix depends on predictors. However it's unclear how to extend this method in the context of independent functional observations or include discrete covariates such as group indicators. Moreover, this particular method requires high dimensional Gibbs updates, which limits its computational feasibility. In contrast, the multivariate covariance regression model of Hoff and Niu [2012] can incorporate continuous as

well as discrete covariates and only requires low dimensional Gibbs updates. However, this flexibility is at the cost of some linear assumptions, making the method not as flexible as Fox and Dunson [2015] in some aspects. See Li et al. [2014] and Quintero and Lesaffre [2017] for extensions of this model to the multivariate multilevel case. The model presented in Section 3.2 can be viewed as a functional extension of Hoff and Niu [2012] with some added flexibility, and we will highlight the similarities and differences as we go along.

The methodological development of the proposed model is motivated by two case studies. The first case study involves electroencephalography (EEG) brain signals recorded on children with autism spectrum disorder (ASD). ASD is a complex neurodevelopmental disorder that affects about 1 in 54 children. ASD is characterized by difficulty in communication, restricted repetitive behaviors, and stereotypical behavior. Low functioning children may have limited behavioral repertoire, necessitating specialized assessment methods. Electroencephalography (EEG) provides a direct measure of postsynaptic brain activity and does not rely on behavioral output from young children with ASD, making EEG based biomarkers appealling for diagnosis, prognosis, and intervention purposes [Jeste, Kirkham, Senturk, Hasenstab, Sugar, Kupelian, Baker, Sanders, Shimizu, Norona, Paparella, Freeman, and Johnson, 2015]. In this study 59 heterogenous children with ASD and 38 age matched typically developing (TD) children had resting-state EEG signals recorded [Dickinson, DiStefano, Senturk, and Jeste, 2018]. This study focused on oscillations in the alpha rhythm, which play a role in neural coordination and communication between distributed brain regions. We describe how random patterns of variation differ between children with ASD and their TD peers, providing novel insights into their neurodevelopmental differences.

The second case study analyzes EEG data collected as part of an in-home polysomnography for the Sleep Heart Health Study (SHHS). This large study was designed to identify factors for sleep-disordered breathing, such as age, blood pressure, or sleep patterns. In this article we treat delta spectral power as a functional response and explore how simply quantify sleep as a functional response and explore how heteroschedasticity depends on age and hypertension status.

The rest of this paper is organized as follows: Section 3.2 gives the generating model for

functional data, Section 3.3 lists prior choices and discusses the reasoning behind them, Section 3.4 briefly discusses computations involved for posterior calculations, Section 3.5 gives a thorough simulation study assessing errors and coverage properties, Section 3.6 showcases the model on the two motivating case studies, and Section 3.7 concludes with a brief discussion. The sampling algorithm and additional simulation details are given in the supplement.

3.2 A Modeling Framework for Covariance Regression

In this section we present a modeling framework for relating patterns of co-variation and time-stable covariates. Let $y_i(t) \in \mathbb{R}$ denote the outcome for subject i at point $t \in \mathcal{T}$ for some real compact interval $\mathcal{T}$. Let $\boldsymbol{x} = (x_1, \ldots, x_{d_1})^\top \in \mathcal{X}$ denote a d_1-dimensional time-stable covariate for subject i, with the dependence on i removed for ease of presentation. The k-dimensional data-generating model is

$$y_i(t) = \mu(t, \boldsymbol{x}) + r_i(t, \boldsymbol{x}) + \epsilon_i(t) \tag{3.1}$$

$$r_i(t, \boldsymbol{x}) = \sum_{j=1}^{k} \psi_j(t, \boldsymbol{x}) \eta_{ij} \tag{3.2}$$

$$\eta_{ij} \sim N(0, 1), \quad \epsilon_i(t) \sim N(0, \varphi^2) \tag{3.3}$$

where $\mu(t, \boldsymbol{x})$ is the conditional mean, $\psi_j(t, \boldsymbol{x})$ form conditional latent functional bases, $\eta_{ij} \sim N(0, 1)$ are subject-specific scores, and $\epsilon_i(t) \sim N(0, \varphi^2)$ represents measurement error. Using Equations (3.1, 3.2, 3.3) the conditional covariance function $c(t, t', \boldsymbol{x})$ is

$$c(t, t', \boldsymbol{x}) = \sum_{j=1}^{k} \psi_j(t, \boldsymbol{x}) \psi_j(t', \boldsymbol{x}) + \varphi^2 \delta_t \tag{3.4}$$

Various approaches exist for specifying the form of $\mu(\cdot)$ and $\psi_j(\cdot)$, including local polynomial smoothers [Fan and Gijbels, 1996], kernel smoothers [Ferraty and Vieu, 2006], Gaussian process methods [Yang et al., 2016b, Fox and Dunson, 2015], and spline procedures [Ramsay,

2004]. Lending toward conceptually straight-forward prior assumptions, we build $\mu(\cdot)$ and $\psi_j(\cdot)$ as linear combinations of spline bases. Grouping d_1 covariates into R subsets, so that $\cup\{\boldsymbol{x}_r\}_{r=1}^R = \boldsymbol{x}$, with $|r|$ denoting the number of covariates in group r, and defining component functions $f_r(t, \boldsymbol{x}_r) : \mathcal{T} \times \mathcal{X}_r \to \mathbb{R}$, we borrow notation from Scheipl et al. [2015], and assume $\mu(t, \boldsymbol{x})$ can be written as

$$\mu(t, \boldsymbol{x}) = \sum_{r=1}^{R} f_r(t, \boldsymbol{x}_r). \tag{3.5}$$

This grouping framework leads to flexible specification of basis expansions. For example, when $\boldsymbol{x}_r = x_r \in \mathbb{R}$ is a single scalar covariate, $f_r(t, x_r)$ can be modeled as a functional linear effect $x_r f(t)$ or, more generally, as a smooth function of t and x, say $f(t, x_r)$. Similarly, if $\boldsymbol{x}_r = (x_{r_1}, x_{r_2})$ is a vector of covariates, $f_r(t, \boldsymbol{x}_r)$ could be written as $f(t, x_{r_1}, x_{r_2})$, $x_{r_1} f(t, x_{r_2})$, or $x_{r_1} x_{r_2} f(t)$, by diminishing degree of generality. These terms are approximated by a set of basis functions with corresponding priors to encourage smooth effects. For the general case, we maintain that $\boldsymbol{x}_r \in \mathbb{R}^{p_r}$ admits a basis expansion $\boldsymbol{b}^r(\boldsymbol{x}_r) = (b_1^r(\boldsymbol{x}_r), \ldots, b_{|r|}^r(\boldsymbol{x}_r)) \in \mathbb{R}^{p_r}$, and assume

$$f_r(t, \boldsymbol{x}_r) = \boldsymbol{b}(t)^\top \beta_r \, \boldsymbol{b}^r(\boldsymbol{x}_r), \tag{3.6}$$

$$\mu(t, \boldsymbol{x}) = \boldsymbol{b}(t)^\top \beta \, \boldsymbol{b}^x(\boldsymbol{x}), \tag{3.7}$$

where $\beta = \big(\beta_1 \,|\, \cdots \,|\, \beta_R\big)$ and $\boldsymbol{b}^x(\boldsymbol{x}) = \big(\boldsymbol{b}^1(\boldsymbol{x}_r) \,|\, \cdots \,|\, \boldsymbol{b}^R(\boldsymbol{x}_r)\big)^\top$. Keeping track of dimensions, β is a $p \times r(d_1)$ coefficient matrix and $\boldsymbol{b}^x(\boldsymbol{x})$ is a $r(d_1) \times 1$ vector, where $r(d_1) = \sum_{r=1}^R p_r$.

We follow a similar strategy for the representation of $\psi_j(t, \boldsymbol{x})$. Speficically, we define latent basis functions $l_{rj}(t, \boldsymbol{x}_r) : \mathcal{T} \times \mathcal{X}_r \to \mathbb{R}$, and assume

$$\psi_j(t, \boldsymbol{x}) = \sum_{r=1}^{R} l_{rj}(t, \boldsymbol{x}_r) \tag{3.8}$$

As done previously, given a basis expansion of the covariate space, we write

$$l_{rj}(t, x_r) = \boldsymbol{b}(t)^\top \Lambda_{rj}\, \boldsymbol{b}^r(\boldsymbol{x}_r),$$
$$\psi_j(t, \boldsymbol{x}) = \boldsymbol{b}(t)^\top\, \Lambda_j \boldsymbol{b}^x(\boldsymbol{x}_r),$$

where Λ_{rj} is a $p \times |r|$ loading matrix and $\Lambda_j = \big(\Lambda_{1j} \,|\, \cdots \,|\, \Lambda_{Rj}\big)$. The additivity on $\psi_j(t, \boldsymbol{x})$ implies that the covariance function in Equation (3.4) is

$$c(t, t', \boldsymbol{x}) = \sum_{j=1}^{k} \left(\sum_{r=1}^{R} \sum_{r'=1}^{R} l_{rj}(t, \boldsymbol{x}_r) l_{rj'}(t', \boldsymbol{x}_{r'}) \right) + \varphi^2 \delta_t.$$

This convolution structure makes interpretation of the latent functions $l_{rj}(t, \boldsymbol{x}_r)$, somewhat difficult as an index of covariate dependence. To aid interpretation, we define low dimensional summaries of covariate influence on the covariance function directly. More precisely, let

$$g_r(t, \boldsymbol{x}_r) = \frac{1}{|\mathcal{X}_r|} \sum_{\mathcal{X}_r} \sum_{j=1}^{k} l_{rj}(t, \boldsymbol{x}_r)^2 \tag{3.9}$$

summarize the effect of $\boldsymbol{x}_r$ across $\psi_j,\ j = 1, \ldots, k$. Here $\mathcal{X}_r$ represents a set of covariate values for the rth group of covariates and $|\cdot|$ denotes cardinatly. If the impact of $\boldsymbol{x}_r$ on $\psi_j,\ j = 1, \ldots, k$ is negligible, then $g_r(t, \boldsymbol{x}_r)$ will be near zero. Consequently if $g_r(t, \boldsymbol{x}_r)$ is near zero, $c(t, t', \boldsymbol{x})$ will not be sensitive to changes in $\boldsymbol{x}_r$. Crucially, the definition of these covariate influence functions, quantifying the effect of exogenous predictors on covariance operators, is independent of assumptions of strict additivity, resulting in a truly non-parametric measure of influence on patterns of covariation. Unlike previous work on functional covariance regression [Cardot, 2007, Jiang et al., 2010], Equations (3.1, 3.2, 3.3) specify a generative model for functional covariance regression. Complete with priors detailed in Section 3.3, posterior inference is completed through Markov-Chain Monte Carlo (MCMC).

Given a finite observation grid, the structure of the likelihood is similar to that of Hoff and

Niu [2012] who considered covariance regression with multivariate response data. However, as Fox and Dunson [2015] note, their mapping from predictors to covariance assumes a parametric form, thus limiting the model's expressivity. In particular, Hoff and Niu [2012] assumes increasing heteroschedasticity as covariate magnitude becomes large. To overcome this parametric limitation, Fox and Dunson [2015] develop a factor matrix process which involves many nonparametric Gaussian processes. This approach is flexible but (1) each gibbs sample iteration requires many cholesky decompositions of $n \times n$ matrices where n is the number of subjects and (2) cannot incorporate discrete covariates. The basis expansion approach taken in this article would only require a cholesky decomposition of a $p \cdot r(d_1)$ by $p \cdot r(d_1)$ matrix for each iteration and accommodate discrete covariates. The basis expansion approach is likely to scale better for large data sets such as the SHHS.

3.3 Prior Distributions

In this section we place priors on all unknown quantities of interest. We begin by placing prior on $\mu(t, \boldsymbol{x})$. As we have seen in Equations 3.6 and 3.7, this amounts to placing a prior on each β_r submatrix. The rows of β_r are associated with a $p \times p$ penalty matrix K, and the columns of β_r are associated with a $|r| \times |r|$ penalty matrix K_r. These penalties are designed to encourage smoothness and can target magnitude penalization, squared derivative shrinkage, or local changes in β_r through a differencing penalty. In this paper we penalize the second order difference of β_r coefficients in both directions, but other penalties could be used as well. A prior for β_r respecting the tensor structure is constructed as follows [Wood, 2017]. Let $\tilde{K} = I_{|r|\times|r|} \otimes K$ and $\tilde{K_r} = K_r \otimes I_{p\times p}$. The prior for the vectorized form of β_r is

$$\text{vec}(\beta_r) \,|\, \tau_{1xr}, \tau_{1tr} \sim$$
$$\exp\{-0.5\text{vec}(\beta_r)^\top(\tau_{1xr}\tilde{K_r} + \tau_{1tr}\tilde{K})\text{vec}(\beta_r)\}$$

where τ_{1xr}, τ_{1tr} are smoothing parameters. If $|r| = 1$ then β_r is a $p \times 1$ vector and $\tilde{K}_r = 0$. In this case the prior simplifies to

$$\beta_r \,|\tau_{1tr} \sim \exp\{-0.5\tau_{1tr}\beta_r^\top K \beta_r\}$$

This prior is improperm, but provided that proper priors are set for τ_{1tr}, τ_{1xr}, the posterior of β_r will be proper [Lang and Brezger, 2004].

As we define priors for $\psi_j(t, \boldsymbol{x})$, we take into consideration their interpretation as pseudo eigen-components of conditional covariance functions, therefore we require that for for larger values of the index j, $\psi_j(t, , \boldsymbol{x})$ contributes an increasingly smaller amount to the overall magniture of covariance. Therefore, the prior for $\psi_j(t, \boldsymbol{x})$ should encourage both smoothing and shrinkage aspects. Let Λ_{rj} be the analogous component to β_r. Re-using the same penalty matrices as above, the prior for Λ_{rj} is

$$\begin{gathered}\mathrm{vec}(\Lambda_{rj}) \,|\, \tau_{2jxr}, \tau_{2jtr}, \tau^*_{rj}, \phi_{rj} \sim \\ \exp\{-0.5\mathrm{vec}(\Lambda_{rj})^\top(\tau_{2jxr}\tilde{K}_r + \tau_{2jtr}\tilde{K} + \tau^*_{rj}\phi_{rj})\mathrm{vec}(\Lambda_{rj})\},\end{gathered}$$

where τ_{2jxr}, τ_{2jtr} are smoothing parameters and ϕ_{rj}, τ^*_{rj} are shrinkage parameters. Here ϕ_{rj} is a diagonal matrix with dimension $p \cdot |r|$ by $p \cdot |r|$. If $|r| = 1$ (so that Λ_{rj} is a column vector), then the prior becomes

$$\Lambda_{rj} \,|\tau_{2jtr}, \tau^*_{rj}, \phi_{rj} \sim \exp\{-0.5\Lambda_{rj}^\top(\tau_{2jtr}\tilde{K} + \phi_{rj}\tau^*_{rj})\Lambda_{rj}\},$$

similar to the case when β_r is a column vector. In summary, these priors for $\mu(t, \boldsymbol{x})$ and $\psi_j(t, \boldsymbol{x})$ are design to reflect our assumptions of smoothness in the functional data outcomes. All smoothing parameters are given independent gamma prior distributions,

$$\begin{aligned}\tau_{1xr}, \tau_{1tr} &\sim \mathrm{Gamma}(a_\tau, b_\tau) \\ \tau_{2jxr}, \tau_{2jtr} &\sim \mathrm{Gamma}(a_\tau, b_\tau)\end{aligned}$$

where we use the 'rate' parameterization of the Gamma distribution (i.e., if $x \sim \text{Gamma}(a, b)$, then $\mathbb{E}[x] = a/b$). In our implementation we follow Kowal and Bourgeois [2020], Gelman et al. [2006] and place uniform priors on all smoothing parameters, so that $a_\tau = -0.5$ and $b_\tau = 0$. In our experience, model fitting can be poor for some choices of a_τ and b_τ but uniform priors tend to have favorable results. We borrow the Gamma Multiplicative Process Prior (GMPP) from Bhattacharya and Dunson [2011], Montagna et al. [2012] to assign priors to τ^*_{rj} and ϕ_{rj}. Let ϕ_{irj} denote the ith diagonal element of ϕ_{rj}.

$$\phi_{irj} \sim \text{Gamma}(a_\phi, b_\phi) \tag{3.10}$$

$$\tau^*_{rj} = \prod_{l=1}^{j} \delta_{rl} \tag{3.11}$$

$$\delta_{r1} \sim \text{Gamma}(a_{r0}, 1) \tag{3.12}$$

$$\delta_{rl} \sim \text{Gamma}(a_{r1}, 1), \ l > 1 \tag{3.13}$$

$$a_{r0}, a_{r1} \sim \text{Gamma}(2, 1) \tag{3.14}$$

so that the τ_{rj} are stochastically increasing in j. This shrinkage is data-adaptive so that later entries of Λ_{rj} may or may not be shrunk toward zero depending on the model fit. The ϕ_{irj} parameters are local shrinkage parameters, and removing these will tend to result in over-shrinkage [Bhattacharya and Dunson, 2011]. Critically, specifying a conservative choice of k (number of latent functional factors) should not change results too much compared to setting k to some "optimal" choice. Finally, we place an Inverse-Gamma prior on random noise variability φ^2. Although Suarez and Ghosal [2017] notes that results may be sensitive to this prior, this has not been the case in our experience. This may be because in our prior specification smoothing comes from the the prior on β and Λ_j, whereas smoothing comes from the prior on φ^2 in the model from Suarez and Ghosal [2017]. Regardless, empirical choices of hyperparameters are available [Wang et al., 2016, Suarez and Ghosal, 2017].

3.4 Posterior inference

We use Gibbs sampling to draw parameter realizations from the joint posterior distribution. In particular, we sample β, Λ_j, ϕ_{rj}, τ_{1tr}, τ_{1xr}, τ_{2jtr}, τ_{2jxr}, δ_{rl}, ϕ_{irj}, φ^2, and η_{ij} through their respective conditional conjugate distributions and sample a_{r0}, a_{r1} through a Metropolis-Hastings step. The most computationally intensive part of the sampling algorithm requires approximately $1/3 \cdot p \cdot r(d_1)$ number of floating point operations used to calculate a cholesky decomposition involved in updating β and Λ_j. Detailed steps of the sampling algorithm are provided in Appendix 3A.

We discuss posterior inference for subject-specific latent trajectories, mean functions, and principal directions of variation. In this paper we use the posterior mean as a point estimate and symmetric pointwise credible intervals for uncertainty quantification. Let $\hat{\beta}_m$, $\hat{\Lambda}_{jm}$, and $\hat{\eta}_{ijm}$ be the mth posterior draw of β, Λ_j, and η_{ij}. The point estimate for the ith subject-specific latent trajectory, $\tilde{y}_i(t)$, is $\frac{1}{M}\sum_{m=1}^{M}\{\boldsymbol{b}(t)^\top\hat{\beta}_m\boldsymbol{b}^x(\boldsymbol{x}_i) + \sum_{j=1}^{k}\boldsymbol{b}(t)^\top\hat{\Lambda}_{jm}\boldsymbol{b}^x(\boldsymbol{x}_i)\}$. Similarly, the point estimate for the covariate-adjusted mean, $\mu(t,\boldsymbol{x})$, is $\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{b}(t)\hat{\beta}_m\boldsymbol{b}^x(\boldsymbol{x}_i)$. Let $C_m(\boldsymbol{x})$ be the matrix given by evaluating the mth posterior draw of $c(t,t',\boldsymbol{x})$ on a subset of $\mathcal{T}\times\mathcal{T}$. Performing a spectral decomposition on $C_m(\boldsymbol{x})$ yields the mth draw of ordered orthonormal posterior principal directions of variation (eigenfunctions) $\tilde{\psi}_{jm}(t,\boldsymbol{x})$, $j = 1,\ldots,k$.

However, this method can be computational expensive due to the spectral decomposition of arbitrarily high-dimensional matrix $C_m(\boldsymbol{x})$ due to the nature of functional data. To alleviate this intensive procedure, we follow Aguilera and Aguilera-Morillo [2013] and only perform a spectral decomposition on a low-dimensional matrix, $\sum_{j=1}^{k}\hat{\Lambda}_{jm}\boldsymbol{b}^x(\boldsymbol{x}_i)\boldsymbol{b}^x(\boldsymbol{x}_i)^\top\hat{\Lambda}_{jm}^\top$, so that we never actually form $C_m(\boldsymbol{x})$ in our implementation. Since $\tilde{\psi}_{jm}(t,\boldsymbol{x})$ is only identifiable up to sign change, each $\psi_{jm}(t,\boldsymbol{x})$ may be potentially multiplied by -1 to orient all eigenfunctions correctly. More details on extracting and orienting posterior eigenfunctions are provided in Appendix 3B. Once all posterior eigenfunctions are oriented correctly, a point-estimate is obtained by simply taking the posterior pointwise mean.

Credible intervals are computed in the following manner. Let $\mu_f(t)$ and $\sigma_f(t)$ be the

posterior mean and standard deviation of some functional $f(\cdot)$ evaluated at t. A $(1-\alpha)\cdot 100$ credible interval for $f(t)$ is $[\hat{f}(t) - z^*_{1-\alpha/2}\sigma_f(t), \hat{f}(t) + z^*_{1-\alpha/2}\sigma_f(t)]$ where $z^*_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Pointwise intervals have nominal average coverage when the model is correctly calibrated. However, pointwise intervals are anti-conservative when performing inference for an entire function $f(\cdot)$. In this case we recommend computing simultaneous credbile bands. Our implementation includes this option and we give more details in Appendix 3B.

3.5 Simulations

We assess the performance of the proposed method in terms of mean, covariance, and latent response point estimation and coverage. We generate data under two scenarios. The first scenario has heteroscedasticity depending on a covariate x scaled within the unit interval, and the second scenario imposes no such relationship between heteroscedasticity and covariates. Both scenarios include a covariate-dependent mean surface. We fit data generated in both scenarios by (1) the proposed method, adjusting for covariate-dependent heteroscedasticity and (2) foregoing this adjustment and assuming the covariate only impacts the mean structure. We generate data in both scenarios by simulating from the model. We generate 300 datasets per scenario and true model parameters are kept fixed over each dataset. We include two sample sizes $N = 100$ and $N = 400$ to numerically verify mean and covariance point estimation convergence.

We begin by using a p-spline of dimension 10 to expand the functional arguement t and a p-spline of dimension 5 (not including an intercept) to expand the covariate argument x. According to notation from section 3.2, $\boldsymbol{b}(t) = (b_1(t), \ldots, b_{10}(t))^\top$, $\boldsymbol{b}^2(x) = (b^2_1(t), \ldots, b^2_5(t))^\top$, and $\mu(t,x) = f_1(t) + f_2(t,x) = \boldsymbol{b}(t)\,\beta_1 + \boldsymbol{b}(t)\,\beta_2\,\boldsymbol{b}^2(x)$. The model parameters β_1 and $\text{vec}(\beta_2)$ have prior mean zero and precision $\Omega_1 + \epsilon \cdot I_{10\times 10}$ and $\Omega_2 + \epsilon \cdot I_{50\times 50}$ respectively. The nugget term $\epsilon \cdot I$ is used to simulate from a full-rank multivariate normal and in our experiments we set $\epsilon = .1$. To ensure smoothness of the resulting response functions $y_i(t)$, we set $\Omega_1 = \tau_{1t1}K$ and $\Omega_2 = \tau_{1x2}(K_2 \otimes I_{10\times 10}) + \tau_{1t2}(I_{6\times 6} \otimes K)$, where K and K_2 are second order discrete

penalty matrices associated with p-splines. We use $k = 4$ latent functional factors for the random component $r_i(t, x)$, so $r_i(t, x) = \sum_{j=1}^{4} \psi_j(t, x)$. Under data-generating scenario 1, we set $\psi_j(t, x) = l_{1j}(t) + l_{2j}(t, x)$. Under data-generating scenario 2, we set $\psi_j(t, x) = l_{1j}(t)$. Expanding out the terms, $l_{1j}(t) = \boldsymbol{b}(t)^\top \Lambda_{1j}$ and $l_{2j}(t, x) = \boldsymbol{b}(t) \Lambda_{2j} \boldsymbol{b}^2(x)$. The model parameters Λ_{1j} and Λ_{2j} prior mean zero and prior precision Γ_{j1} and Γ_{j2} respectively. We set $\Gamma_{j1} = \tau_{2t1} K + \tau^*_{1j}$ and $\Gamma_{j2} = \tau_{2x2}(K_2 \otimes I_{10\times 10}) + \tau_{2t2}(I_{6\times 6} \otimes K) + \tau^*_{2j}$. Smoothing and shrinkage parameters are $\tau_{1t1} = \tau_{1t2} = \tau_{2t1} = \tau_{2t2} = 10$, $\tau_{1x2} = \tau_{2x2} = 100$, $\tau^*_{1j} = 2^j$, and $\tau^*_{2j} = 4^j$. Measurement error is $\varphi^2 = .20^2$ and functional data $y_i(t)$ is generated according to Equation 3.1. The functional argument t is evaluated at 100 equally spaced locations within the unit interval and x_i is equal to $(i - 1)/(N - 1)$.

It will be convenient to divide data-generating scenarios and fitting procedures into four separate cases. Case 1 has data generated from scenario 1, fit with a model that adjusts for heteroscedasticity due to the covariate. Case 2 has data generated from scenario 2, fit with a model that does not adjust for heteroscedasticity due to the covariate. Case 3 has data generated from scenario 2, fit with a model that adjusts for heteroscedasticity due to the covariate. Case 4 has data generated from scenario 2, fit with a model that does not adjust for heteroscedasticity due to the covariate. We fit each dataset with with $k = 6$ latent functional factors and expand the functional argument into a p-spline basis of dimension 10. We expand the covariate argument into a p-spline basis of dimension 7. In doing so, we simply enlarge the model space and test its ability to regularize unnecessary latent dimensions and basis expansions. We place uniform priors on all smoothing parameters and a Gamma$(0.0001, 0.0001)$ prior on the inverse measurement error, φ^{-2}. We fit each model with 50,000 samples, discarding 25,000 as burn-in and keeping every 5th observation to save computer memory. We keep track of coverage, average credible interval width, and relative integrated squared error (RISE). The coverage performance assess a model's calibration and goodness of fit, whereas RISE assesses a model's performance in terms of point estimation.

Let $t_1, \ldots, t_{100}$ be 100 uniformly placed points along the domain of the functional argument, $x^*_1, \ldots, x^*_{10}$ be uniformly placed points along the interval $[0.1, 0.9]$, and $\tilde{y}_i(t)$ denote the latent response of $y_i(t)$ uncontaminated with random error. Let $\mathbb{1}(\cdot)$ be the indica-

tor function that returns 1 when the argument is true and 0 otherwise. The coverage of $\mu(t,x)$ is $\frac{1}{1000}\sum_{i=1}^{10}\sum_{m=1}^{100}\mathbb{1}\{\mu(t_m,x_i^*)\in\widehat{I}_{mi}\}$, the coverage of $c(t,t',x)$ is $\frac{1}{50500}\sum_{i=1}^{10}\sum_{m=1}^{100}\sum_{m'=1}^{m}\mathbb{1}\{c(t_m,t_{m'},x_i^*)\in\widehat{I}_{mm'i}\}$, and the coverage of $\tilde{y}_i(t)$ is $\frac{1}{N\times 100}\sum_{i=1}^{N}\sum_{m=1}^{100}\mathbb{1}\{\tilde{y}_i(t_m)\in\widehat{I}_{y,mi}\}$. Here $\widehat{I}_{mi}$ is the nominal 95% posterior interval about $\mu(t_m,x_i)$ and $\widehat{I}_{mm'i}$ is the nominal 95% posterior interval about $c(t_m,t_{m'},x_i)$, and $\widehat{I}_{y,mi}$ is the nominal 95% posterior interval about $\tilde{y}_i(t)$. The RISE of $\mu(t,x)$ is $100\cdot\frac{\sum_{i=1}^{10}\int_{\mathcal{T}}\{\widehat{\mu}(t,x_i^*)-\mu(t,x_i^*)\}^2\,dt}{\sum_{i=1}^{10}\int_{\mathcal{T}}\mu(t,x_i^*)^2\,dt}$, the RISE of $c(t,t',x)$ is $100\cdot\frac{\sum_{i=1}^{10}\int_{\mathcal{T}}\int_{\mathcal{T}}\{\widehat{c}(t,t',x_i^*)-c(t,t',x_i^*)\}^2\,dt\,dt'}{\int_{\mathcal{T}}\int_{\mathcal{T}}\sum_{i=1}^{10}c(t,t',x_i^*)^2\,dt\,dt'}$, and the RISE of $\tilde{y}_i(t)$ is $100\cdot\frac{\sum_{i=1}^{N}\int_{\mathcal{T}}\{\widehat{y}_i(t)-\tilde{y}_i(t)\}^2\,dt}{\sum_{i=1}^{N}\int_{\mathcal{T}}\tilde{y}_i(t)^2\,dt}$. In practice we use the trapezoidal rule to evaluate all integrals. We also compute average credible band width for $\mu(t,x)$, $c(t,t',x)$ and $\tilde{y}_i(t)$. All coverage, RISE, and interval width results are reported in Table 3.1.

3.6 Case Studies

3.6.1 Application to ASD study

Peak alpha frequency (PAF), the frequency at which oscillations in the alpha range demonstrate maximal power, shows well-characterised increases with chronological age during childhood in typically developing children [Somsen et al., 1997, Dustman et al., 1999, Stroganova et al., 1999, Chiang et al., 2011, Cragg et al., 2011, Miskovic et al., 2015]. PAF has been shown to index neural development in TD children [Valdés-Hernández et al., 2010, Segalowitz et al., 2010, Rodríguez-Martínez et al., 2017]. However, a recent study by Dickinson et al. [2018] found that children with ASD did not show the typical increase in PAF with age. In the present study, we investigate the relationship between diagnostic status, age, and alpha spectral density. Treating alpha spectral density as a functional response avoids complicated PAF identification procedures [Dickinson et al., 2018] and retains more information as opposed to collapsing the entire alpha spectral band to a single point [Scheffler et al., 2019, 2020b]. In this study we wish to characterize functional mean and covariance dependence of alpha spectral density in terms of age and diagnostic status.

In the Dickinson et al. [2018] study, resting-state EEG was collected on 39 TD children

Table 3.1: Simulation results over 300 data sets for each sample size and case combination. Coverage, error (RISE), and interval width are averaged over all data sets. Refer to Section 3.5 for details on how coverage, error, and interval width are calculated.

Sample size	Quantity	Case	Coverage	Error	Interval width
$N = 100$	$\mu(t, x)$	1	.98	.84	.30
		2	.98	.91	.32
		3	.98	.78	.30
		4	.98	.79	.30
	$c(t, t', x)$	1	.89	7.00	.14
		2	.72	9.76	.11
		3	.98	4.31	.13
		4	.97	3.29	.11
	$y_i(t)$	1	.97	.25	.17
		2	.97	.29	.19
		3	.98	.21	.17
		4	.98	.20	.16
$N = 400$	$\mu(t, x)$	1	.99	.19	.18
		2	.98	.21	.19
		3	.99	.18	.18
		4	.98	.18	.18
	$c(t, t', x)$	1	.88	3.10	.09
		2	.46	7.23	.06
		3	.98	1.01	.06
		4	.97	.84	.05
	$y_i(t)$	1	.97	.13	.16
		2	.97	.18	.18
		3	.98	.12	.16
		4	.98	.12	.16

and 58 children with ASD aged 2 to 12 years old. EEG signals were sampled at 500 Hz for two minutes and interpolated to a 10-20 25 channel montage [Jasper, 1958, Perrin et al., 1989]. Alpha spectral density $\Omega \in [6 \text{ Hz}, 14 \text{ Hz}]$ were obtained for each electrode. To facilitate comparisons across regions and subjects, alpha spectral density was normalized to unit area. See Scheffler et al. [2019] for more details on data acquisition and pre-processing. The resulting data structure is region-referenced functional data, which induces correlated functional responses within subjects. Unfortunately, the proposed method can not accommodate correlated functional data. Instead, for demonstration purposes, we only examine one region at a time. We also acknowledge that Scheffler et al. [2020b] already successfully applied a covariate-adjusted hybrid principal components analysis, which can accommodate correlated functional responses within subject. While the case study of Scheffler et al. [2020b] is similar to the one in this article, we note that the two methods are very different in their methodological work. In addition, the low dimensional summary in Equation 3.9 provides some insight into how the variability of normalized alpha spectral density as a function of age and diagnostic status.

We consider EEG data arising from the T8 region while adjusting mean and and covariance functions by diagnostic status, age, and a diagnostic status by age interaction. This interaction is scientifically motivated because as previously mentioned, PAF tends to increase with age for TD children but not so for children with ASD. For the frequency dimension, we use a p-spline basis with 12 degrees of freedom using a second order differencing penalty. We expand age via p-splines with 6 degrees of freedom, again with a second order difference penalty. We run a markov chain monte carlo algorithm for 200,000 iterations. To keep memory management light, we only save every 20 iterations. Of the leftover 10,000 iterations, we discard the first 5,000 as burn-in and keep the subsequent 5,000 for inference.

Figure 3.1 plots normalized alpha spectral density over several ages of child and diagnostic group with associated 95% pointwise credible bands. The means curves drift apart as the cross-sectional age increases, which supports the notion that differences in alpha spectral band dynamics between ASD and TD children at similar ages can successfully predict diagnostic status Scheffler et al. [2019]. Figure 3.2 displays the leading age and group adjusted

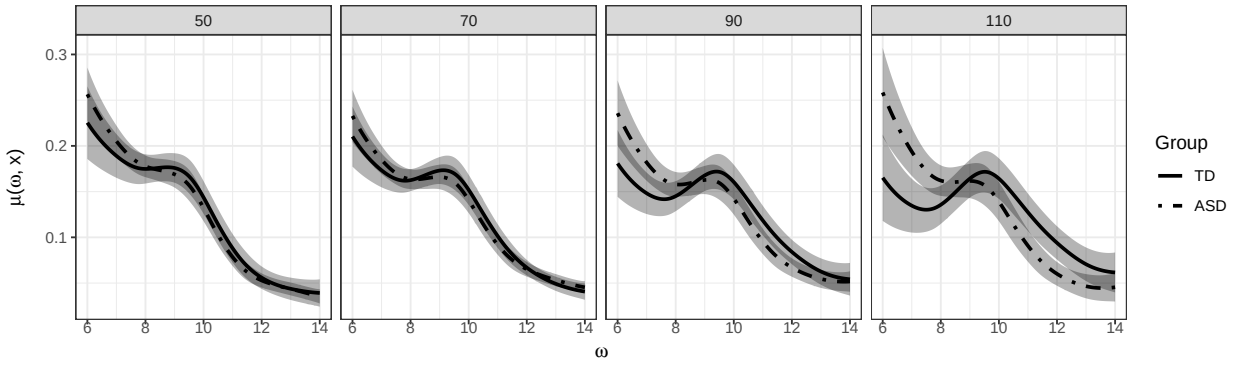


Figure 3.1: Posterior mean alpha spectral power for ASD and TD groups at age 50, 70, 90, and 110 months. The shaded area represents 95% pointwise credible intervals.

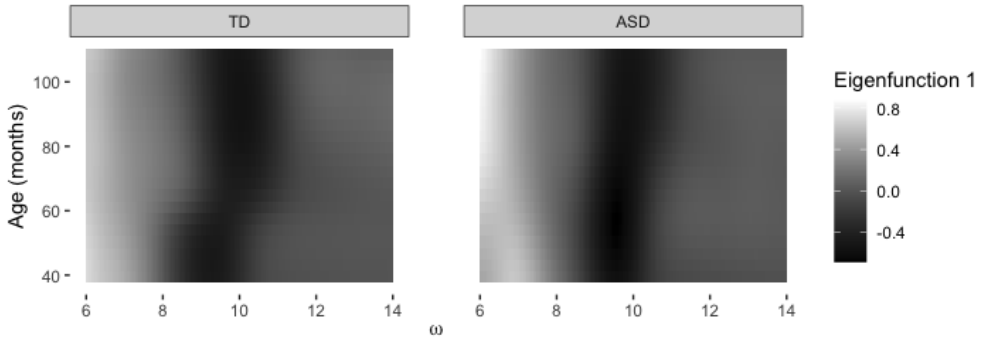


Figure 3.2: Posterior mean of the leading eigenfunction adjusted by age and diagnostic status for the resting state ASD experiment. The TD group has a clear shift in shape over age, but this shift is obscured in the ASD group.

eigenfunction. The general shape for a particular age is a unimodal curve peaked around 8-10 Hz. Notice that the peak of this curve tends to shift from lower to higher frequencies as age increases for the TD group, whereas the peak frequency is relatively constant over age for the ASD group, which is supported by the analysis in Scheffler et al. [2020b]. Posterior summaries $g(\omega, \boldsymbol{x})$ indicate heterogeneity is similar between ASD and TD groups. Moreover, heterogeneity does not depend on either group. See Appendix 3C for further discussion.

3.6.2 Application to Sleep Heart Health Study

The Sleep Heart Health Study (SHHS) was a prospective cohort study designed to investigate obstructive sleep apnea (OSA) and other sleep-disordered breathing (SDB) as risk factors for the development of cardiovascular disease [Quan et al., 1997]. Parent cohort

studies and recruitment targets for these cohorts are the following: Atherosclerosis Risk in Communities Study (1,750 participants), Cardiovascular Health Study (1,350 participants), Framingham Heart Study (1,000 participants), Strong Heart Study (600 participants), New York Hy- pertension Cohorts (1,000 participants), and Tucson Epidemiologic Study of Airways Obstructive Diseases and the Health and Environment Study (900 participants). An unattended in-home polysomnography was completed by participants between November 1, 1995 and January 31, 1998. Between January 2001 and June 2003, a second polysomnogram was obtained in 3295 of the participants.

The American Academy of Sleep recognizes four sleep stages: stage N1 (light sleep), stage N2 (relaxation), stage N3 (slow-wave sleep), and stage R (rapid-eye movement). Importantly, N3 sleep, or slow-wave sleep (SWS) consists of high amplitude (≥ 75 μV) and low-frequency ($0.5 - 4.0$) delta waves. SWS is considered to be most restorative sleep stage and to be associated with sleep quality, sleep maintenance, and functions toward memory consolidation [Bonnet, 1987, Akerstedt et al., 1997, Walker, 2009]. Mander et al. [2017] reports that advancing into the fifth decade of age comes with (1) reduced SWS time and (2) increased time spent in lighter sleep stages (N1, N2). In addition, Javaheri et al. [2018] reports that lower levels of percent SWS sleep are associated with increased odds of incident hypertension in both men and women, independent of confounders such as sleep apnea, age, and sex. Most clinical studies (including the papers referenced above) quantify sleep by classifying time-varying electrical phenoma into discrete sleep stages. Subsequently, amount of SWS is represented by a single number, which is commonly percentage of sleep time in stage N3. However, this approach comes with several limitations [Crainiceanu et al., 2009] including low intraclass correlation coefficient, no biological basis, and loss of temporal information. In this paper follow Crainiceanu et al. [2009], Di et al. [2009] and use power spectral density analysis to quantify sleep EEG. Our present goal is to characterize age-related changes in sleep between hypertension and non-hypertension groups.

We use the discrete fast Fourier transform to analyze the sleep EEG data in the frequency domain. The entire night of sleep is broken into 30-second adjacent sleep epochs to account for temporal effects. These 30-second windows are processed using Welch's method of 50%

overlapping windows (4 second windows with 2 seconds overlap), where the intervals are windowed using a tapered Tukey window. The epoch-level power spectral density estimate is the average of the windowed power spectra. We use a band-pass filter to attenuate signals less than .3 Hz and greater than 35 Hz with a .5 Hz transition width. We mask epochs where artifacts detected using the method described in Buckelmüller et al. [2006]. We also mask epochs that are statistical outliers using (2, 2) Hjorth parameters Hjorth [1970]. We compute delta spectral power by summing power spectra from 0.5 Hz to 4.0 Hz in .25 Hz increments on an epoch-by-epoch basis. To facilitate comparisons across participants, we compute epoch-level relative delta power spectral density (RDPSD) by dividing this power density by a summation of spectral density from .5 Hz to 35 Hz in .25 Hz increments. All filtering and power spectral density computations were performed in Luna [Purcell, 2020], which is an open-source software package for manipulating and analyzing polysomnographic readings.

We restrict our attention to EEG data collected on the first visit. We also filter participants to only include those who scored at least four hours of artifact-free signal. We analyze the first two hours of sleep, resulting in a dense grid of length 240 epochs for each participant. The 10th and 90th percentiles of age are 47 and 77 years-old, respectively. There are 2177 participants with hypertension and 3081 participants without hypertension. We use a p-spline of dimension 24 for the marginal basis for epoch and a p-spline of dimension 8 for the marginal basis for age. Each basis is associated with a second order differencing penalty matrix. We also include a hypertension dummy variable and a hypertension-age interaction. Exploratory analysis from Castro et al. [1986], Carroll et al. [2020] shows 9 principal components are needed to explain 99% of variability (unadjusted for age and hypertension group). We fit our model with $k = 12$ latent factors since unneeded latent factors will be set close to zero due to the prior shrinkage process in Equations 3.11-3.14. We ran the MCMC algorithm described in the supplement for 200,000 iterations with a burn-in of 100,000, saving every 20th iteration for memory requirements.

Figure 3.3 displays mean RDPSD with associated 95% point-wise credible intervals for ages 50, 60, 70, and years old for both hyertension and non-hypertension groups. Rela-

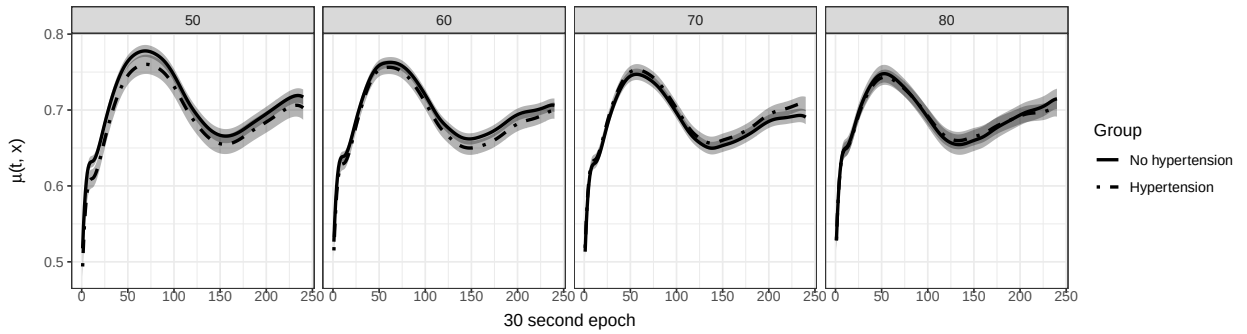


Figure 3.3: Relative delta power spectral density for the first two hours of sleep adjusted by age at 50, 60, 70, and 80 years and hypertension diagnostic group.

tive power spectral density decreases in N3 (around epoch 60) with age for both groups. The seemingly little decrease of SWS from midlife to late life is supported by Van Cauter et al. [2000], who found that percent SWS predominantly decreases from early adulthood to midlife, with no further decrease into late life. Their study included 149 healthy men aged 16 to 83 years. Our methods support their finding after considering a much richer metric than percent SWS in the context of a much larger observational study. There is considerable overlap in the credible intervals between the hypertension and non-hypertension RDPSD, which does not support the hypothesis that hypertension is associated with increased levels of RDPSD.

We conducted an eigen-analysis on the model-based covariance surfaces. The estimated leading eigenfunction for a 50 year old with no hypertension accounts for 31% of total variability and is positive over the entire epoch grid, suggesting that this component represents an overall RDPSD size construct. This eigenfunction, nor subsequent eigenfunctions vary much over age or group. Heterogeneity of RDPSD is similar between hypertensive and non-hypertensive groups. In addition, heterogeneity does not depend on age of the subject for either group. See Appendix 3C for further discussion.

3.7 Discussion

We developed a probabilistic model suitable for functional data with covariate-dependent heteroscedasticity. We carefully designed adaptive smoothing and shrinkage priors for co-

variance regularization purposes. In particular, we note that the smoothing prior reflects smoothness assumptions and the shrinkage prior makes the latent structure robust to misspecified number of latent functional factors. We reiterate that the proposed approach enables joint mean/covariance estimation, straightforward inference, incorporating group indicators, and the calculation of key low dimensional summaries. The simulation experiment illustrated attractive qualities such as calibrated coverage and decreasing error/interval width as the sample size increases. The proposed apprach is computationally feasible for moderately large datasets such as the sleep study, which has $N = 5258$ subjects, 240 timepoints, and 16 basis functions for the covariate dimension. This feasibility is owed to tractible lower dimensional gibbs sampling updates. We applied our methodology to better understand the neural mechanisms surrounding ASD and distinguishing slow-wave sleep patterns over age by hyptertension status.

Several promising extensions are possible to the existing methodology. The framework can be extended to accommodate multilevel functional data for nested, longitudinal, and repeated measures applications by incorporating a latent structure in addition to $r_i(t, \boldsymbol{x})$. Variable selection in the latent structure is largely unexplored. Employing variable selection techniques in the flavor of Kowal and Bourgeois [2020] in the context of functional covariance regression requires would require intricate design considerations for computational feasbility. Along those lines, it would be interesting to adjust only a small number of adaptively chosen latent functional factors instead of every latent functional factor as we have done in this article. It would also be interesting to adaptively choose the number of latent functional factors as in Montagna et al. [2012].

Data Availability Statement

Data used in the resting state EEG experiment is not publicly available. Please contact Shafali Jeste at sjeste@mednet.ucla.edu. Access to Sleep Heart Health Study annotated sleep waveform data requires registering an account at sleepdata.org and requesting dataset access.

Supporting Information

Software in the form of an R package including complete documentation and sample data set is publicly available at github.com/jshamsho/bfcr.

CHAPTER 4

Bayesian Analysis of Region-Referenced Functional Data

We develop Bayesian approaches to model region-referenced functional data - multivariate functional data observed over regional subunits. The proposed methods identify data-driven interpretable marginal functional and regional basis functions. Prior structure is imposed to encourage smoothness in the functional domain and weaken the influence of superfluous basis functions. The proposed methods incorporate scalar covariates, enabling joint mean and covariance estimation. In addition, we show how pivotal discrepancy measures can be used to assess covariance structural assumptions and aid in model selection. The proposed methods are applied to study electroencephalography data from children with autism spectrum disorder in a resting-state experiment. Supplementary materials, including a documented Rcpp package with a tutorial, are available online.

4.1 Introduction

Functional principal components analysis (FPCA) and the associated Karhunen-Loéve (KL) decomposition [Karhunen, 1946, Loève, 1946] is the most prevalent tool in univariate functional data analysis, appearing in regression, classification, dimension reduction, and exploratory applications. Owing to the versatility of the KL decomposition, this idea has been extended to accommodate multivariate functional data [Jacques and Preda, 2014, Chiou et al., 2014, Happ and Greven, 2018]. *Region-referenced* functional data - multivariate functional data observed over regional subunits - usually possess a structured dependency pattern that allows one to analyze marginal patterns of covariation along the functional domain and

the spatial domain separately. For example, Scheffler et al. [2020b] considered repeated oscillatory measurements along a number of electrodes on the scalp from electroencephalography (EEG) recordings. The authors proposed hybrid principal components analysis (HPCA), which involves a tensor basis of one-dimensional marginal eigenvectors and eigenfunctions consisting of regional, longitudinal, and functional bases. The use of marginal eigenvectors and eigenfunctions has advantages for interpretability. In a related *longitudinal*-functional data setting, functional data repeated over longitudinal time, Park and Staicu [2015b], Lynch and Chen [2018a] propose decomposed random longitudinal-functional effects in terms of a tensor of marginal eigenfunctions along the longitudinal and functional dimension. As noted by Li et al. [2020], these marginal eigenvector/eigenfunction decompositions are special cases of the general multivariate KL factorization [Happ and Greven, 2018].

In this article we direct our attention to region-referenced functional data. This data structure is distinct from spatially indexed functional data [Morris and Carroll, 2006, Baladandayuthapani et al., 2008, Staicu et al., 2010]. Contrary to spatial functional data analyses, region-referenced functional data analyses do not explictly account for distance between regions in the modeling process. This is an important point because in many physical processes close proximity between regions does not imply high functional correlation. Brain imaging applications are typically analyzed with the aforementioned modified KL approaches, letting the data learn important patterns of spatial or functional dynamics instead of optimizing hyper-parameters in a mixed model fashion appropriate for spatially indexed functional data. To that end, we develop Bayesian analogues of weak separability Lynch and Chen [2018a,b], Shamshoian et al. [2020] and partial separability [Zapata et al., 2019]. The Bayesian framework offers straightforward inference for all model quantities of interest through the posterior distribution. In addition, we develop informal Bayesian hierarchical goodness of fit tests for assessing the adequacy of the modified probabilistic KL expansions and mean structure.

This article is structured as follows. Section 4.2 presents a probabilistic model as an alternative to weak and partial separability. Section 4.3 discusses choice of priors and goodness of fit assessments. Section 4.4 conducts numerical experiments to evaluate the proposed

methodology in finite samples with comparisons to other pertinent techniques. Section 4.5 applies the proposed methodology to analyze hierarchical EEG data from children with autism spectrum disorder. Section 4.6 summarizes the work and suggests future research directions.

4.2 Probabilistic Models for Region-Referenced Functional Data

Suppose we observe discrete realizations of a stochastic process $Y_{ij}(t) : \mathcal{T} \to \mathbb{R}$, $Y_{ij} \in L_2(\mathcal{T})$, over a compact subset $\mathcal{T} \subset \mathbb{R}$, at points $t \in \{t_1, \ldots, t_n\}$, over regions $j \in \{1, \ldots, R\}$, for subjects $i = 1, \ldots, n$. Let $\boldsymbol{x}_i$ be a $D-$dimensional covariate vector representing exogenous information often assumed to affect the mean structure. Let $\mathbf{Y}_i(t) = (Y_{i1}(t), \ldots, Y_{iR}(t))^\top$ be an $R-$dimensional vector containing measurements from all regions at t. We refer to the following decomposition as the partially separable functional linear model (PSFLM):

$$\mathbf{Y}_i(t) = \sum_{l=1}^{\infty} \boldsymbol{\theta}_{il}\psi_l(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \Sigma_\epsilon) \tag{4.1}$$

$$\boldsymbol{\theta}_{il} = \sum_{j=1}^{R} \boldsymbol{\phi}_{lj}\eta_{ilj} = \phi_l \boldsymbol{\eta}_{il}, \quad \eta_{ilj} \sim t_\nu(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{lj}, \sigma_{lj}^2) \tag{4.2}$$

Here $\epsilon_i(t)$ represents normally distributed measurement error with mean zero constant diagonal variance matrix $\Sigma_\epsilon = \text{diag}(\sigma_{\epsilon 1}^2, \ldots, \sigma_{\epsilon R}^2)$. The functions $\psi_l(t) : \mathcal{T} \to \mathbb{R}$, $\psi_l \in L_2(\mathcal{T})$, are latent basis functions to be estimated from the data, and $\boldsymbol{\theta}_{il} = (\theta_{il1}, \ldots, \theta_{ilR})^\top$ is an $R-$dimensional vector of random effects for the ith subject. We use Bayesian factor analysis to model the θ_{il} terms, with $\boldsymbol{\phi}_{lj} \in \mathbb{R}^R$, $j = 1, \ldots, R$ forming the loading components and η_{ilj} being the subject-specific factors with a $t_\nu(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{lj}, \sigma_{lj}^2)$ distribution. The scaled student's t distribution is favored over the normal distribution for robustness to outliers in the latent space [Lee and Lee, 2020, Kowal and Bourgeois, 2020]. This model is a probabilistic extension to the partially separable KL decomposition from Zapata et al. [2019].

A special case of PSFLM occurs when $\phi_{lj} \equiv \phi_j$ up to sign changes. That is, when the loadings do not depend on l, we recover a probabilistic extension to the weakly separable

KL decomposition [Lynch and Chen, 2018a,b]. We refer to this special case the weakly separable functional linear model (WSFLM) and explore its properties in this article. We note that ideas similar to PSFLM have appeared in the literature in the context of longitudinal-functional data analysis [Park and Staicu, 2015b, Lynch and Chen, 2018a]. We also note that Shamshoian et al. [2020] developed a probabilistic model for longitudinal functional data using weak separability. These modifications of the general multivariate KL decomposition of stochastic processes is are pervasive in the literature on structured functional data analysis. However, to the best of our knowledge, this article is the first to complete Bayesian analysis with these models in the context of region-referenced data. Within the Bayesian framework, we discuss crucial aspects of regularized estimation through ordered shrinkage priors, obtain exact posterior inference, and conduct informal goodness of fit assessments for the covariance structure as shown in Section 4.3.

Let $b_p(t)$, $b = 1, \ldots, P$, be a set of basis functions spanning a function space on $\mathcal{T}$, e.g. B-splines, we assume $\psi_l(t)$ is represented as:

$$\psi_l(t) = \sum_{p=1}^{P} b_p(t)\lambda_{pl} = \boldsymbol{b}(t)\boldsymbol{\lambda}_l. \tag{4.3}$$

Let $\boldsymbol{t} = (t_1, \ldots, t_N)^\top$ be the sampling design at which we observe realizations of our region-referenced process. We follow Kowal and Bourgeois [2020], Kowal [2021] and constrain λ_{pl} so that $\psi_l(\boldsymbol{t})^\top \psi_{l'}(\boldsymbol{t}) = \delta_{ll'}$ and $\phi_l^\top \phi_l = I_R$ where δ is the kronecker delta and I_R is the $R \times R$ identity matrix. These constraints lead to substantially reduced computational burden via likelihood simplifications, which is particularly convenient in the Markov-Chain Monte-Carlo (MCMC) algorithm used for estimation.

These restriction to orthonormal basis functions aid likelihood identifiability and interpretation, and can be implemented efficiently in numerical analysis settings. Implementation details are provided in Appendix 4A.

We compare the above models to existing models in the structured functional data literature. Let θ_{ilj} be the jth component of $\boldsymbol{\theta}_{il}$. If $\boldsymbol{\theta}_{il}$ has a compound symmetric variance-covariance matrix, the above models are similar to the multilevel FPCA of Di et al.

[2009]. To see this, suppose $Y_{ij}(t)$ follows the multilevel FPCA. Then $\text{Cov}(Y_{ij}(s), Y_{ik}(t)) = K_B(s,t)$ for some covariance function $K_B(s,t)$. In contrast, if $Y_{ij}(t)$ follows PSFLM, then $\text{Cov}(Y_{ij}(s), Y_{ik}(t)) = \sum_{l=1}^{\infty} \text{Cov}(\theta_{ilj}, \theta_{ilk})\psi_l(s)\psi_l(t)$. These covariance functions are structurally equivalent if and only if $\text{Cov}(\theta_{ilj}, \theta_{ilk}) = \rho_l$ for some positive sequence $\{\rho_l\}_{l=1}^{\infty}$. In general, under PSFLM $\text{Cov}(\theta_{ilj}, \theta_{ilk})$ depends on j and k, yielding a more flexible model than the multilevel FPCA. Now suppose the variance-covariance matrix of $\boldsymbol{\theta}_{il}$ does not depend on l. Then $\text{Cov}(\boldsymbol{Y}_i(s), \boldsymbol{Y}_i(t)) = \sum_{l=1}^{\infty} \text{Var}(\boldsymbol{\theta}_{il})\psi_l(s)\psi_l(t) = Q\psi_l(s)\psi_l(t)$. This covariance function has a separable form, so clearly WSFLM is a generalization of separability. However, PSFLM and WSFLM are parsimonious special cases of the non-separable model of Li et al. [2020] and the previously mentioned MFPCA decomposition. However, PSFLM and WSFLM provide low dimensional interpretable information on regional and functional dynamics, making them appealing for analyzing variability among marginal dimensions separately.

4.3 Prior Distributions and Assessment of Model Adequacy

4.3.1 Priors Distributions

Priors must be placed on all unknown model parameters. There will typically be enough data so that the posterior for $\sigma^2_{\epsilon j}$ will be insensitive to the choice of prior, so we set $p(\log \sigma_{\epsilon j}) \propto 1$ as a default choice. We follow Kowal and Bourgeois [2020], Kowal [2021] and place a uniform$(2, 128)$ prior on the t degrees of freedom for η_{ilj}. We place independent $t_4(0, 1)$ priors on each element of β_{lj}. We placea Gaussian Markov Random Field prior on $\boldsymbol{\lambda}_l$ by

$$\lambda_l \sim \exp(-.5\zeta_l \boldsymbol{\lambda}_l^{\top} \Omega \boldsymbol{\lambda}_l) \tag{4.4}$$

where Ω is a $P \times P$ known singular roughness penalty matrix and ζ_l controls the smoothness of $\psi_l(t)$. Typically, P will be of moderate size justifying independent uniform priors for $\zeta_l^{1/2}$ [Gelman et al., 2006, Kowal and Bourgeois, 2020]. To ensure adaptive regularization of covariance components across dimensions, we place a Gamma Multiplicative Process Prior [Baladandayuthapani et al., 2008] on the σ^2_{lj} terms from Equation 4.2, so that σ^2_{lj} is

stochastically decreasing in both l and j. Under PSFLM, the prior for σ_{lj} is

$$\sigma_{l1}^{-2} = \prod_{l'=1}^{l} \delta_{l'} \quad \delta_1 \sim \text{Gamma}(a_1, 1), \quad \delta_l \sim \text{Gamma}(a_2, 1) \text{ for } l > 1 \tag{4.5}$$

$$\sigma_{lj}^{-2} = \prod_{l'=1}^{l} \prod_{j'=2}^{j} \delta_{l'}\delta_{lj'} \quad \delta_{lj} \sim \text{Gamma}(a_3, 1) \text{ for } j \geq 2 \tag{4.6}$$

$$a_k \sim \text{Gamma}(2, 1), \quad k = 1, 2, 3 \tag{4.7}$$

Under WSFLM, the prior for σ_{lj}^2 is

$$\sigma_{lj}^{-2} = \sum_{v=1}^{V} d_v a_{vj} b_{vl} \tag{4.8}$$

$$a_{v1} \sim \text{Gamma}(1, 1), \quad a_{vj} \sim \text{Gamma}(1, 1)1(a_{vj} > a_{v(j-1)}) \text{ for } j > 1 \tag{4.9}$$

$$b_{v1} \sim \text{Gamma}(1, 1), \quad b_{vl} \sim \text{Gamma}(1, 1)1(b_{vl} > b_{v(l-1)}) \text{ for } l > 1 \tag{4.10}$$

$$d_1 = 1, \quad d_v \sim \text{Gamma}(1, 1)1(d_v > d_{v-1}) \text{ for } v > 1 \tag{4.11}$$

In practice, the sum in Equation 4.1 cannot have an infinite amount of terms. By truncation, after L terms, the data-generating model for $\boldsymbol{Y}_i(t)$ becomes $\boldsymbol{Y}_i(t) = \sum_{l=1}^{L} \boldsymbol{\theta}_{il}\psi_l(t) + \epsilon_i(t)$. As L increases, the above models become more flexible but tend to overfit for large L. In general, the shrinkage priors shown above reduce overfitting by encouraging the variances of η_{ilr} to be small when η_{ilr} is not needed to model $\boldsymbol{Y}_i(t)$. The shrinkage priors strike a good balance between model flexibility and complexity, which is an important aspect in factor models Bhattacharya and Dunson [2011] and functional regression models [Montagna et al., 2012]. In turn, these priors help make posterior inferences insensitive to the choice of L, provided that L is large enough [Bhattacharya and Dunson, 2011, Kowal and Bourgeois, 2020, Kowal, 2021, Shamshoian et al., 2020, Li et al., 2020].

The PSFLM shrinkage prior in equations (4.5)-(4.7) encourages smaller variances of η_{ilr} with increasing l and j. The structure of the prior is commensurate with the partial separability decomposition [Zapata et al., 2019]. The WSFLM shrinkage prior in equations (4.8)-(4.11) encourages smaller variances of η_{ilr} with increasing l and j. In addition, the d_v

terms limit the constribution of the a_{vj} and b_{vl} terms as v increases. This prior has been carefully constructed to be commensurate with weak separability [Lynch and Chen, 2018a]. Finally, we place independent $N(0,1)$ priors on the elements of ϕ_l. The $N(0,1)$ is a vague prior in this case because the magnitude of any particular entry of ϕ_l (or ϕ) cannot be greater than 1 due to the imposed orthonormality constraint.

We use a Markov-Chain Monte Carlo (MCMC) to simulate samples from the posterior $p(\theta \mid \{\boldsymbol{Y}_i(t), x_i\}_{i=1}^N)$, where θ is any model quantity of interest. We use a combination of Gibbs and Metropolis-Hastings algorithms to sample model parameters from their full conditional distributions. A detailed description of the posterior simulation strategy is summarized in Appendix 4B. After posterior simulation, inference on model quantities of interest is straightforward via post-processing of Monte Carlo samples.

4.3.2 Assessment of Model Adequacy

In this article we are mainly concerned with the goodness of fit in the covariance structure as opposed to the mean structure. Yuan and Johnson [2012] provides theoretical justification to assess the goodness of fit of hierarchical Bayesian models through pivotal discrepancy measures (PDMs). PDMs are functions of model parameters that have an invariant distribution when evaluated at the data-generating (true) model parameter. The key result in the aforementioned article is that PDMs constructed with parameters sampled from the posterior distribution maintain the same invariant distribution. Alternatively, posterior predictive checks [Gelman et al., 1996] are limited to assessing goodness of fit at the data-generating level, and hence have limited use as a diagnostic check for multilevel models [Yuan and Johnson, 2012].

Valid inference of the covariance structure under partial separability assumes $\text{Cov}(\boldsymbol{\theta}_{il}, \boldsymbol{\theta}_{il'}) = 0_{R\times R}$ when $l \neq l'$. [Zapata et al., 2019]. Jiang and Qi [2015] developed a likelihood ratio test to infer whether or not this assumption holds in a frequentist setting

as follows. Construct a matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1L} \\ A_{21} & A_{22} & \cdots & A_{2L} \\ \vdots & \cdots & \cdots & \vdots \\ A_{L1} & A_{L2} & \cdots & A_{LL} \end{pmatrix}$$

such that $A_{jk} = \sum_{i=1}^{n} (\boldsymbol{\theta}_{ij} - \overline{\boldsymbol{\theta}}_{ij})(\boldsymbol{\theta}_{ik} - \overline{\boldsymbol{\theta}}_{ik})^{\top}$ and $\overline{\boldsymbol{\theta}}_{ij} = \sum_{i=1}^{n} \boldsymbol{\theta}_{ij}$. Define

$$W_n = \frac{|A|}{\prod_{l=1}^{L} |A_{ll}|}$$

Then $(\log W_n - \mu_n)/\sigma_n$ converges in distribution to $N(0,1)$ when $\mathrm{Cov}(\boldsymbol{\theta}_{il}, \boldsymbol{\theta}_{il'}) = 0_{R\times R}$ for $r \neq k$ for some sequence of $(\mu_n,\ \sigma_n)$ under some regularity conditions (see Jiang and Qi [2015] Theorem 2). The test rejects the null hypothesis ($\mathrm{Cov}(\boldsymbol{\theta}_{il}, \boldsymbol{\theta}_{il'}) = 0_{R\times R}$) at type I error rate α when $(\log W_n - \mu_n)/\sigma_n \leq z_\alpha$, where z_α is the α quantile of the standard normal distribution.. In the language of Yuan and Johnson [2012], $(\log W_n - \mu_n)/\sigma_n$ is a PDM with a $N(0,1)$ reference distribution. Weak separability assumes $\mathrm{Cov}(\eta_{ijk}, \eta_{ij'k'}) = \delta_{jj'}\delta_{kk'}$ where δ is the kronecker delta. In other words, $\mathrm{Cov}\{\mathrm{vec}(\eta_i)\} = D$, where D is a diagonal matrix. Let $\hat{R}_n$ be the sample correlation coefficient matrix of $\mathrm{vec}(\eta_i)$. Then $(\log|\hat{R}_n| - \mu_n)/\sigma_n$, (μ_n, σ_n) appropriately defined, converges in distribution to $N(0,1)$ when $\mathrm{Cov}(\eta_{ijk}, \eta_{ij'k'}) = 0$ by Corollary 1 of Jiang and Qi [2015]. The rejection region is $(\log|\hat{R}_n| - \mu_n)/\sigma_n \leq -z_\alpha$, where z_α is the same as defined above.

We assess the joint distribution of PDMs using two methods. The first method follows the proposal by Yuan and Johnson [2012] which identifies a quantity p_{min} based on theoretical properties of order statistics. The authors suggest that a p_{min} less than 0.25 implies lack of fit. The second method computes the p-value at the posterior average of PDMs. We call this method p_{mean} and lack of fit is determined when $p_{mean} < \alpha$ for some α type I error. Care must be taken in interpreting p_{min} and p_{mean} since these quantities are not uniform under the null hypothesis. Both methods are computationally trivial to implement with posterior samples, in contrast to the simulation intensive prior predictive method of Dey et al. [1998].

The above methods are simple criteria for determining model adequacy. These choices are not unique, and other criteria, such as the proportion of posterior PDMs exceeding the .95 quantile of the reference distribution, may be adopted to aid in the determination of model adequacy. We discuss operating characteristics of p_{min} and p_{mean} in Section 4.4 and further interpretation of PDMs in Section 4.5. We note that these procedures detect lack of fit in the covariance as a whole. Departures from the hypothesized covariance implied by partial separability or weak separability may still yield useful inferences on some model components. For example, if the data is partially separable, eigenfunction estimates $\hat{\psi}(t)$ assuming weak separability are still valid. It would be interesting to examine exactly *how* the data departs from a hypothesized model. We defer this nuanced aspect to future research.

4.4 Numerical Experiments

We perform an extensive simulation study to investigate operating characteristics of the proposed models and diagnostic measures of model adequacy in several synthetic scenarios. Region-referenced functional data is generated according to 3 scenarios: (1) weak separability, (2) partial separability, and (3) a structure that violates partial separability, denoted as nonseparable. The first L fourier basis functions constitute the eigenfunctions $\psi_l(t)$, $l = 1 \ldots, L$ and region means are set to zero. Under partial and weak separability, $\sigma^2_{lj} = 20j^{-1}\exp(-.5l)$, $\eta_{ilj} \sim N(0, \sigma^2_{lj})$, and $\Sigma_\epsilon = .01I_R$. When data is generated from weak separability, the regional eigenvectors ϕ are eigenvectors of the compound symmetric matrix $\rho_1 11^\top + (1-\rho_1)I_R$. When data is generated from partial separability, the regional eigenvectors ϕ_l are randomly generated orthogonal matrices using techniques from Mezzadri [2006]. In the third scenario we generate data as follows:

1. Let Σ_θ be an $RL \times RL$ matrix.
2. Denote the l, l' $R \times R$ submatrix of Σ_θ as $\Sigma_{\theta,ll'}$.
3. Set $\Sigma_{\theta,ll} = (L-l+1)^2 \cdot (\rho_2 11^\top + (1-\rho_2)I_R)$ for $l = 1, \ldots, L$.
4. Set $\Sigma_{ll'} = (L-l+1)(L-l'+1)\rho_3 11^\top$ for $l \neq l'$, $l, l' = 1, \ldots L$ and $\rho_3 < \rho_2$.

5. Simulate $\theta_i \sim N(0, \Sigma_\theta)$.

Clearly, when $\rho_3 = 0$, Σ_θ is block-diagonal so that the covariance structure of the simulated data reduces to partial separability. When $\rho_3 > 0$, the covariance structure of $\boldsymbol{\theta}_i$ is not block-diagonal, which is the core assumption of partial separability. Therefore ρ_3 acts as a continuous lack of fit parameter, with large values indicating large departures from partial separability and vice versa. We evaluate the fourier eigenfunctions on an equally spaced dense grid in $[0, 1]$ containing t_n points. In the above settings we set $L = 4$, $R = 6$, $\rho_1 = 0.35$, $\rho_2 = 0.6$, $\rho_3 = 0.2$, and $t_n = 60$. We fit the simulated data with a p-spline basis of dimension $P = 15$ and a second order differencing penalty. We use an initialization procedure described in Appendix 4C to obtain initial estimates for ψ, ϕ_l, σ^2_{lj}, Σ_ϵ, and η_{ilj}. Here we focus on covariance structures, and include no covariate information in the simulation, e.g. $x_i = 1$.

We fit weakly and partially separable models to data generated from each scenario (1) - (3). The sample size is set to $n = 50$ and $n = 200$ to examine how estimates and goodness of fit assessments change with sample size. We simulate 300 data sets for each scenario, fitting method, and sample size, resulting in 3,600 total simulated data sets. For each simulated data set, we record the integrated squared error (ISE) of the posterior mean of the regional mean functions and marginal eigenfunctions. Let ISE_μ and ISE_{ψ_l} denote the ISE of the regional mean functions and the lth eigenfunction. ISE_μ and ISE_{ψ_l} are computed by

$$ISE_\mu = \frac{1}{R}\sum_{j=1}^{R}\int_{[0,1]} \hat{\mu}_j(t)^2\,dt, \quad ISE_{\psi_l} = \int_{[0,1]} [\hat{\psi}_l(t) - \psi_l(t)]^2\,dt$$

where $\hat{\mu}_j(t)$ and $\hat{\psi}_l(t)$ are the posterior means of $\mu_j(t)$ and $\psi_l(t)$ respectively. Both integrals are numerically approximated with the 'trapezoid' rule.

We also record the widely applicable information criterion (WAIC; Watanabe and Opper [2010]) for each fitted model, which is broadly used for model selection purposes as in Vehtari et al. [2017]. We will investigate model selection operating characteristics on the basis of WAIC and PDMs. When data is simulated from a weakly or partially separable model, we

tabulate how many times WAIC and PDMs favor fitting with a weakly or partially separable model. WAIC for PSFLM and WSFLM is computed for each simulated data set. In addition, we compute the standard error of the difference of WAIC between PSFLM and WSFLM. Vehtari et al. [2017] compared models by examining whether or not the difference of WAIC exceed twice the standard error of WAIC difference. We follow the same approach and use this criteria to designate when PSFLM or WSFLM is preferred according to WAIC. When no model is preferred by this criteria, model preferance is inconclusive. In this article, WAIC is calculated pointwise as

$$\text{WAIC} = -2(\hat{\text{lppd}} - \hat{p}_{\text{WAIC}})$$
$$\hat{\text{lppd}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(Y_i(t_j) \mid \theta^s)\right), \quad \hat{p}_{\text{WAIC}} = \sum_{i=1}^{n}\sum_{j=1}^{n} V_{s=1}^{S}\big(\log p(Y_i(t_j) \mid \theta^s)\big)$$

where S is the posterior sample size, θ^s denotes the sth posterior sample of all level one parameters, and $V_{s=1}^{S}$ represents sample variance. The quantities $\hat{\text{lppd}}$ and $\hat{p}_{\text{WAIC}}$ are known as the log posterior predictive density and the effective number of parameters, respectively.

Table 4.1 lists mean rejection rates, ISE_{μ}, and ISE_{ψ_l} for $l = 1, 2, 3$ on the 10^{-3} scale. The fourth column of the table, headlined $H_{0,\alpha=.05}$ gives empirical rejection rates at the 5% significance level. For example, the first row has $H_{0,\alpha=.05} = 34\%$ which means of the 300 simulated data sets 34% resulted in $p_{mean} < .05$ for the test of partial separability. Ideally we would like to see empirical rejection rates around the nominal significance level when the simulated data is fit with the assumed model structure. As expected, partial and weak separability rejection rates increase as the sample size grows and the data fitting procedure does not match the data generating process. Using p_{mean} as a decision rule for assessing partial separaility results in very close rejection rates to the nominal type I error. Using p_{mean} as a decision rule for assessing weak separability results in deflated type I error rates, even when $n = 200$. On the other hand, using $p_{min} < .25$ as a decision rule results in higher power and type I error rates across every simulated scenario. These tests are not exact because the $N(0, 1)$ reference distribution is only an approximation because the PDMs are based on asymptotic likelihood ratio tests assuming normality. Nevertheless, the p_{mean}

Table 4.1: Empirical rejection rates using p_{mean} at the $\alpha = 0.05$ and $\alpha = 0.10$ levels under various data-generating truths and fitting with partial and weak separable models. Empirical rejection rates using the $p_{min} < .25$ decision rule are also shown. Estimation errors for ISE_{μ}, ISE_{ψ_l}, $l = 1, 2, 3$ on the 10^{-3} scale are shown as well.

n	Truth	Fit	$H_{0,\,\alpha=.05}$	$H_{0,\,\alpha=.10}$	H_0, p_{min}	$\mu(t)$	$\psi_1(t)$	$\psi_2(t)$	$\psi_3(t)$
50	Nonseparable	Partial	34.00%	48.33%	57.33%	3.33	2.23	25.08	26.56
		Weak	16%	27%	32%	3.08	0.36	12.52	15.57
	Partial	Partial	7%	14.33%	18.67%	2.37	0.17	6.05	8.19
		Weak	92.98%	94.65%	96.66%	2.26	0.18	9.42	13.94
	Weak	Partial	6.67%	10.33%	15%	2.4	0.21	12.19	17.67
		Weak	2.33%	5.33%	7%	2.24	0.19	11.81	16.35
200	Nonseparable	Partial	100%	100%	100%	1.2	13.68	23.22	19.93
		Weak	99%	99.67%	99.67%	1.16	6.82	14.79	15.24
	Partial	Partial	5%	9.67%	13%	0.66	0.04	1.24	2.51
		Weak	100%	100%	100%	0.58	0.06	2.93	5.30
	Weak	Partial	4.67%	8%	11%	0.69	0.06	3.07	5.13
		Weak	2.34%	4.01%	7.69%	0.65	0.06	2.78	4.97

and p_{min} decision rules high power and acceptable type I error rate according to simulation results in Table 4.1.

Table 4.2 compares model selection procedures on the basis of PDMs and WAIC as described previously. Clearly WAIC is unable to distinguish between partially and weakly separable models. This is because WAIC, as defined in this paper, is asymptotically equal to pointwise leave-one-out cross-validation. From a pointwise perspective, weak and partial separable models yield very similar predictions at the data generating level, so WAIC has difficulty distinguishing between the two. In contrast, PDMs targetting hierarchical model structure allow for model selection in this setting. When partially separable data is simulated, PDMs consistently select fitting a partially separable model. When weakly separable data is simulated, PDMs either select fitting with a weakly separable model or give inconclusive results. Inconclusive results are not surprising in this setting because partial separability is a generalization of weak separability, and if results are inconclusive then would likely choose a weakly separable model for easier interpretations.

Table 4.2: Using WAIC and PDMs to perform model selection under partial and weak separability truths. The percentages shown represent the proportion of simulations that select a partially or weakly separable model according to criteria outlined in Section 4.4

n	Truth	Method	Partial	Weak	Inconclusive
50	Partial	WAIC	0%	0%	100%
		PDM	100%	0%	0%
	Weak	WAIC	0%	0.33%	99.67%
		PDM	0.33%	37.67%	62%
200	Partial	WAIC	0.33%	0%	99.67%
		PDM	100%	0%	0%
	Weak	WAIC	0%	0.33%	99.67%
		PDM	1.33%	26.33%	72.33%

4.5 Case Study

Peak alpha frequency (PAF), the frequency at which oscillations in the alpha range demonstrate maximal power, shows well-characterised increases with chronological age during childhood in typically developing children [Somsen et al., 1997, Dustman et al., 1999, Stroganova et al., 1999, Chiang et al., 2011, Cragg et al., 2011, Miskovic et al., 2015]. PAF has been shown to index neural development in TD children [Valdés-Hernández et al., 2010, Segalowitz et al., 2010, Rodríguez-Martínez et al., 2017]. However, a recent study by Dickinson et al. [2018] found that children with ASD did not show the typical increase in PAF with age. Unfortunately, identifying a single PAF induces loss of important spectral information. Furthermore, identifying a single PAF from a noisy spectrogram can be difficult, and Dickinson et al. [2018] used Gaussian curve fitting procedures to circumvent this issue. Treating alpha spectral density as functional data avoids complicated PAF identification procedures and retains more information as opposed to collapsing the entire alpha spectral band to a single point. In this article we aim to quantify the association between age and alpha spectral density, identify principal patterns of variation within group, and validate our analysis with model adequacy checks. Children aged 2-12 years old were recruited throughout Los Angeles by the UCLA Center for Autism Research and Treatment (CART) via community flyers, the CART website, and ongoing CART studies. In this experiment, children had resting state electroencephalography (EEG) recorded while observing bubbles on a computer

screen for two minutes in a dark, sound attenuated room. We follow the data preparation of Scheffler et al. [2019] to extract alpha spectral densities in the frequency domain (6-14 Hz), interpolated over 25 electrodes in the standard 10-20 system.

We average the log spectral density within five ($R = 5$) anatomical brain regions. The brain regions and corresponding electrode label are frontal (Fp1, Fp2, F3, F4, F7, F8, F9, F10), central (C3, Cz, C4), left temporal (T7, T9), right temporal (T8, T10), and occiptal-parietal (O1, O2, Pz, P3, P4, P7, P8, P9, P10). This regional partition is meant to enhance EEG signals and interpretability. We show results from fitting two weakly separable models, one for each group (ASD and TD). We include age of child as a covariate, normalized to have mean zero and standard deviation one. We use a basis of p-splines of dimension 12 and encourage smoothness by a second order differencing penalty. We choose L to explain 95% of variability as described in the initialization procedure in Appendix 4C. The models require $L = 5$ to explain 95% of variability. We run both models for 100,000 iterations, discard the first 25,000 iterations, and keep every 10th iteration for monte-carlo estimation.

Figure 4.1 displays six trajectories overlaid with posterior means for the underlying (denoised) signal. Model fit seems flexible enough to capture salient dynamics of the signal at the data-generating levels. The fitted underlying signals according to PSFLM and WSFLM are virtually identical. This supports the notion that reconstructions of individual trajectories are not sensitive to model specification between PSFLM and WSFLM. Figure 4.2 displays the posterior mean and 95% pointwise posterior bands for the marginal effect of aging one standard deviation on the mean function for each region. Age does not seem associated with log spectral density in expectation for the ASD group. However, age is clearly associated with log spectral density in expectation for the TD group. In the TD group, aging is associated with deflated log spectral density at lower frequencies and inflated log spectral density at higher frequencies. Moreover, this relationship between age and log spectral density is consistent across regions. Essentially, the dominant functional peak shifts to higher frequencies as TD children age, which is corroborated by previous literature. Similar to individual trajectories, the mean function estimates are insensitive to model specification between PSFLM and WSFLM.

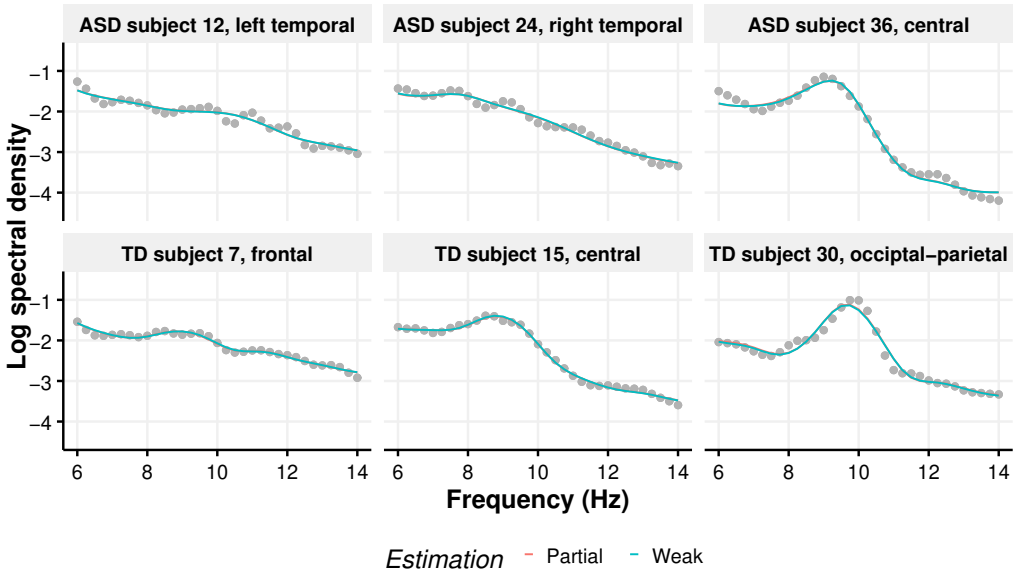


Figure 4.1: Six example log spectral densities for ASD and TD children. Posterior means for the underlying (de-noised) signal are superimposed on top of the observed spectral densities, showing adequate fit at the data generating level.

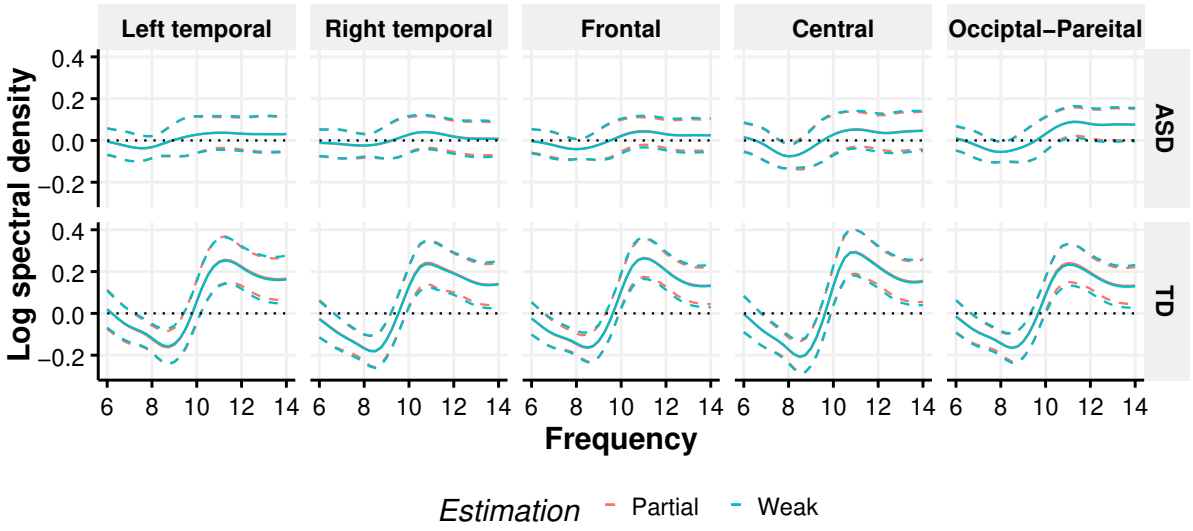


Figure 4.2: Association between aging one standard deviation and log spectral density. Posterior means and 95% pointwise credible intervals are displayed, which show that age does not seem to have a relationship with log spectral density. However, there's clearly a relationship between age and log spectral density for the TD group.

Marginal eigenfunctions for the frequency dimension are shown in Figure 4.3. The first eigenfunction accounts for 36.8% to 45.8% of heterogeneity for the ASD group and 35.4% to 49.8% of heterogeneity for the TD group. The second eigenfunction accounts for 33.2% to 40.1% of heterogeneity for the ASD group and 24.2% to 36.2% of heterogeneity for the TD group. All interval intervals capture 95% of posterior proportion of variability. ASD subjects loading high on the first eigenfunction will tend to have inflated log spectral density at lower frequencies (6 - 8 Hz) and deflated log spectral density at higher frequencies (8 - 14 Hz). The first eigenfunction for the ASD group represents a contrast between low frequencies (6 - 8 Hz) and high frequencies (12 - 14 Hz). TD subjects loading high on the first eigenfunction will have below-average alpha spectral density for higher frequencies (10 - 14 Hz). ASD subjects loading high on the second eigenfunction will tend to have a peaked alpha spectral density around 9 Hz. TD subjects loading high on the second eigenfunction will tend to have a peaked spectral density at 10.5 Hz. This shift in heterogeneity frequency tracks with the overall mean log spectral density across the two groups, indicating a bias-variance relationship. In other words, even though the TD group has inflated log spectral density at higher frequencies, there's still a considerable amount of variability surrounding this inflation at the individual level. The third eigenfunction for both groups contrasts log spectral density at 8-10 Hz and 10-12 Hz.

Under PSFLM, the prinicpal three eigenfunctions account for the ASD group account for 40.7%, 39.6%, and 10.7% of total variability respectively. Under WSFLM, the prinicpal three eigenfunctions account for the ASD group account for 41.5%, 37.7%, and 12.2% of total variability respectively. The discrepancy in the functional form of the first two eigenfunctions according to PSFLM and WSFLM is most likely due to identification issues: the top two eigenfunctions explain nearly the same amount of variability under PSFLM. Repeating the same procedure for the TD group, the top three eigenfunctions explain 40.3%, 33.8%, 13% (PSFLM) and 42.4%, 30.5%, 14.7% (WSFLM) of variability. The proportions of variability explained are more distinguished, resulting in more agreement between the PSFLM and WSFLM eigenfunction estimates.

Principal regional eigenvectors under WSFLM are displayed in Figure 4.4. The principal

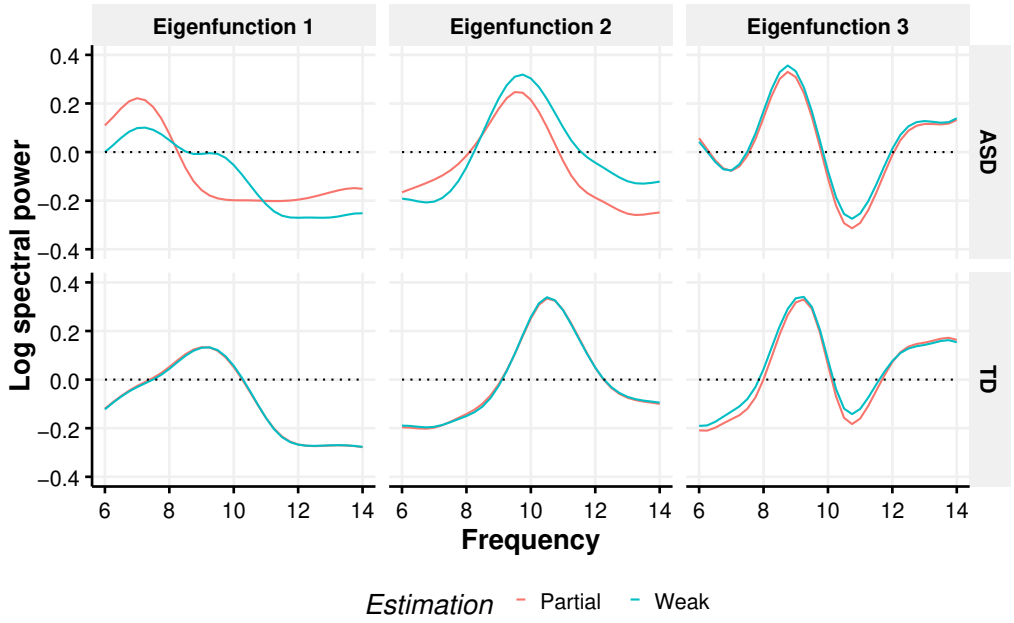


Figure 4.3: Marginal eigenfunctions for the frequency dimension under PSFLM and WSFLM. In the ASD group the first three eigenfunctions explain 40.7%, 39.6%, 10.7% (PSFLM) and 41.5%, 37.7%, 12.2% (WSFLM) of total variability. In the TD group the first three eigenfunctions explain 40.3%, 33.8%, 13.8% (PSFLM) and 42.4%, 30.5%, 14.7% (WSFLM) of total variability. All percent variability explained estimates are monte carlo posterior medians.

regional eigenvector explains about 90% of variability in both groups. Since this eigenvector is nearly constant over each region, it represents an overall magnitude shift in log spectral density. In other words, 90% of the variability can be explained by simple magnitude shifts away from the mean, marginalizing over the frequency dimension. The second principal accounts for approximately 5% of the total variability. This eigenvector is a contrast between the left temporal and right temporal region, which indicates there may be some negative correlation between these regions. However, since the percent variability is so small, it's difficult to assess the validity of this interpretation in a practical sense. The third eigenvector explains approximately 3% of total variability. We omit the interpretation for this eigenvector since it explains such a relatively low amount of variability. Figure 4.4 also displays posterior averaged $\boldsymbol{\theta}_{il}$ under PSFLM. Interpreting regional dynamics from the $\boldsymbol{\theta}_{il}$ is challenging in this setting. This is not too surprising, as PSFLM sacrifices regional interpretability for more flexible covariance estimation compared to WSFLM.

We assess model adequacy using PDMs as explained in Section 4.3.2. Under weak separability, the ASD group has $p_{mean} = .02$ and $p_{min} = .12$. The TD group has $p_{mean} = .038$ and $p_{min} = .27$. Under partial separability, the ASD has $p_{mean} = .01$ and $p_{min} = .08$. The TD group has $p_{mean} = .001$ and $p_{min} = .02$. Based on these small p-values, the assumption of weak or partial separability may not apply to either the ASD or TD data set. In this case, it may be to inaccurate to interpret marginal eigenfunctions as independent sources of heterogeneity. In particular, this analysis may be followed up with more flexible methods, including MPFCA [Happ and Greven, 2018] or vectorizing over regions [Li et al., 2020]. Interpretations will be more difficult but more appropriate with the covariance structure of the data. We note that simply examining the goodness of fit at the data level (e.g., Figure 4.1) is not a valid tool to assess model adequacy in the covariance structure.

4.6 Discussion

We presented Bayesian methods to model region-referenced functional data. The methods differ in the decomposition of the region-referenced covariance function. In particular, we

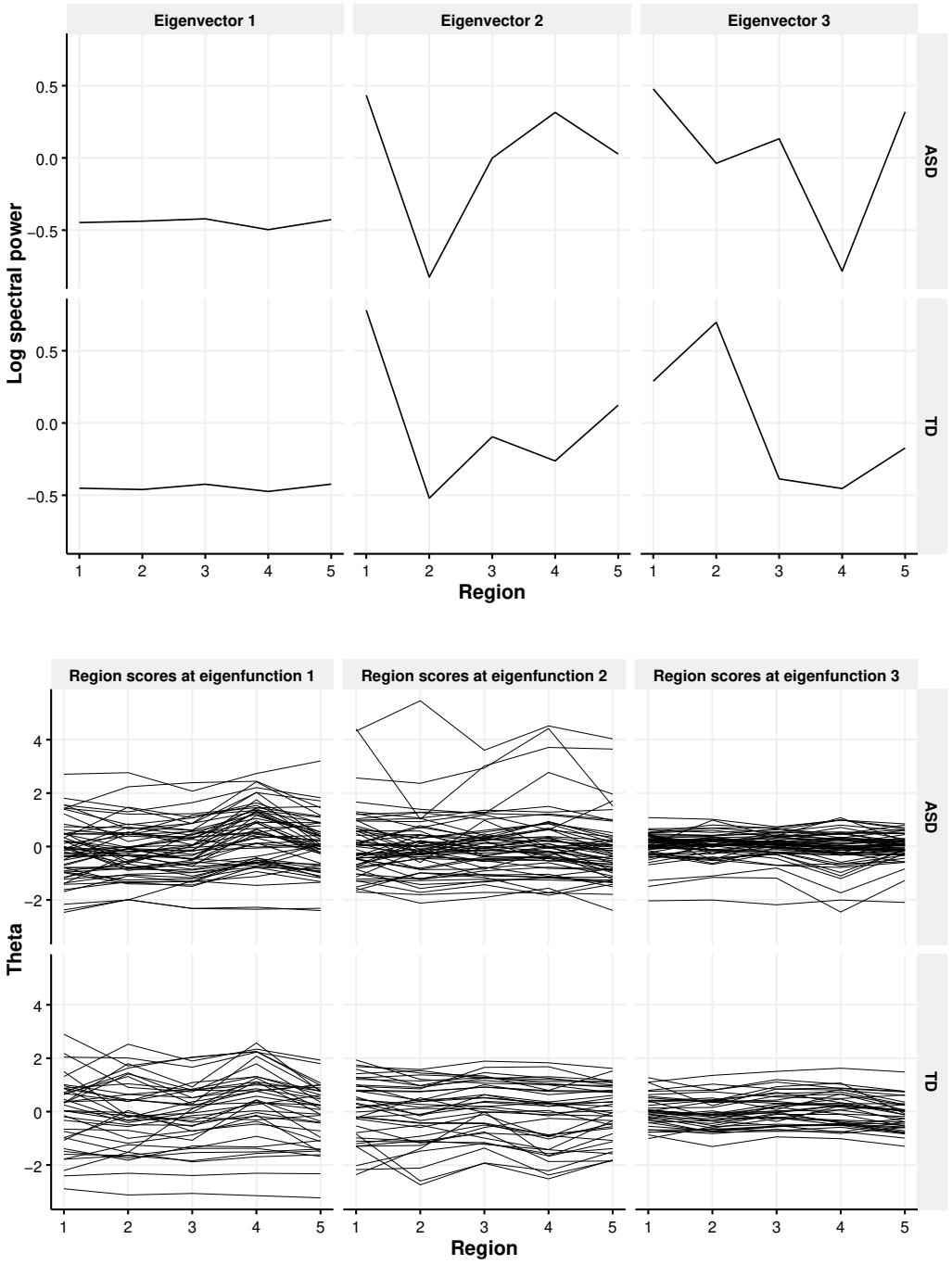


Figure 4.4: Top: Regional eigenvectors under WSFLM. The three eigenvectors explain 90%, 5%, and 3% of variability in both groups. Bottom: posterior averaged θ_{il} estimates for $l = 1, 2, 3$.

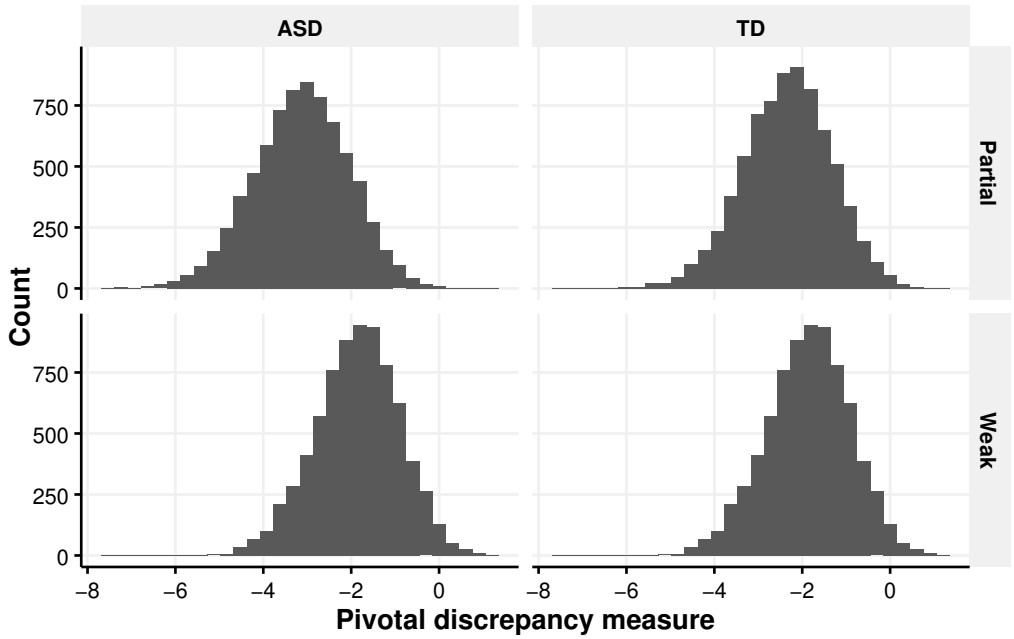


Figure 4.5: Histogram of pivotal discrepancy measures under PSFLM and WSFLM for both groups. The reference distribution is $N(0, 1)$ in all cases.

explored Bayesian models adhering to partial separability and weak separability, where partial separability is a generalization of weak separability. We showed how to construct priors which help prevent overfitting by smoothing functional latent factors and regularizing the influence of superfluous functional and regional latent factors. In our implementation we orthonormalize functional and regional latent factors, leading to greatly reduced computational burden and straightforward interpretations. We also developed pivotal discrepancy measures to assess model adequacy in the covariance structure for both weak and partial separability. We conducted numerical experiments to show that model adequacy decisions based on PDMs have calibrated operating characteristics with respect to false positives and power. We also showed how information criteria, such as WAIC, cannot be used as a model selection tool between partial and weak separability. We applied our methods to analyze alpha spectral density data from children with ASD from a resting state EEG experiment performed at the UCLA CART. In particular, the methods presented in this paper identified age-varying mean functions and a mean-variance relationship between ASD and TD children.

Assessing covariance assumptions validates the interpretation of marginal eigenfunctions and eigenvectors. In this article we showed how PDMs offer a simple solution to assess covariance assumptions based on likelihood ratio tests. We suspect this methodology may be extended other various KL decompositions for hierarchical functional data. For example, assessing covariance structure under the multievel FPCA [Di et al., 2009] using PDMs may allow the level one and level two eigenfunctions to be interpreted as independent sources of variability, rather than solely low dimensional features. Another open problem in the hierarchical functional data literature is selecting a suitable KL decomposition. Choosing a suitable KL decomposition is a non-trivial problem due to inherent complex dependency patterns in hierarchical functional data.

CHAPTER 5

Conclusions

The proposed methods provide effective Bayesian aproaches for modeling functional data in complex settings. We highlight the following advantages of modeling functional data in the Bayesian framework:

1. Adaptive regularization through prior assumptions. All chapters in this dissertation used some form of adaptive regularization. Chapters 2 and 3 modified the Gamma Multiplicative Process prior (GMPP) [Bhattacharya and Dunson, 2011]. Chapter 4 used a combination of GMPP and truncated gamma distributions. Both approaches have shown stable operating characteristics in high dimensional settings with small sample sizes.

2. Straightforward uncertainty quantification. As is typical of Bayesian analyses, inference for key model summaries is conceptually straightforward in all settings considered in this dissertation. To extract Bayesian functional principle components in Chapters 2 and 3, we modified the method of Aguilera and Aguilera-Morillo [2013] to post-process MCMC samples.

3. Joint mean covariance estimation. Empirical approaches typically estimate the mean function under working independence. Then a covariance surface approximation is made based off this mean function. One can iterate between mean and covariance estimations, but the Bayesian framework provides a natural solution to this issue through joint inference.

4. Model adequacy checks and model selection. Model adequacy checks run deep in the Bayesian paradigm. In Chapter 4 we showed how structural covariance assumptions

may be assessed using pivotal discrepancy measures, which serve as diagnostic checks for arbitrary levels in hierarchical models. We also demonstrated how pivotal discrepancy measures may be used as a model selection tool for covariance structure in correlated functional data settings.

The chapters in this dissertation pave the way for further research. Chapter 3 explored Bayesian covariance regression in a very simple setting: dense, one-dimensional, and independent functions. Bayesian covariance regression may be extended to the multivariate or structured functional data setting as in Scheffler et al. [2020b]. Modeling multivariate functions jointly, would yield better performance through Bayesian partial pooling compared to analyzing one functional outcome at a time, as performed in Chapter 3.

More and more strategies for modeling structured functional data are appearing in the FDA literature. Pivotal discrepancy measures may be easily extended to assess covariance assumptions in these complex settings, potentially providing a means to model selection. Using data-level model selection tools (e.g., WAIC) were not effective in assessing covariance assumptions or comparing between models with different covariances. Comparing covariance assumptions between models is a crucial step because covariance assumptions are directly related to interpretable components and inference.

The vast majority of Bayesian FDA methods rely on Gibbs and Metropolis-Hastings sampling, which is a limiting factor for both large evaluation domains and multilevel structure. In addition, implementing these custom algorithms takes up valuable time. Bayesian FDA needs more general purpose MCMC software, perhaps using Variational Hamiltonian Monte Carlo and surrogate functions for likelihood evaluations [Zhang et al., 2018]. It would be particularly interesting to move away from Gibbs and Metropolis-Hastings algorithms, while still retaining the orthonormality constraints used in Chapter 4 for likelihood simplifications.

Appendix 2A: Relationship to Weak Separability

From the main text, we have

$$\begin{aligned} y_i(s,t) &= f_i(s,t) + \epsilon_i(s,t) \\ &= \sum_{j=1}^{q_2}\sum_{k=1}^{q_1} \psi_j(s)\phi_k(t)\eta_{ijk} + r_i(s,t) + \epsilon_i(s,t) \end{aligned}$$

Let $\tilde{f}_i(s,t) = \sum_{j=1}^{q_2}\sum_{k=1}^{q_1}\psi_j(s)\phi_k(t)\eta_{ijk}$. Marginalizing over η_i, the covariance for $\tilde{f}_i(s,t)$ is

$$Cov\{\tilde{f}_i(s,t), \tilde{f}_i(s',t')\} = \sum_{j=1}^{q_2}\sum_{j'=1}^{q_2}\sum_{k=1}^{q_1}\sum_{k'=1}^{q_1}\psi_j(s)\psi_{j'}(s')\phi_k(t)\phi_{k'}(t')Cov(\eta_{ijk}, \eta_{ij'k'})$$

Our model has $Cov(\eta_{ijk}, \eta_{ij'k'}) = 0$ unless $j = j'$ and $k = k'$. Let $h_{jk} = Var(\eta_{ijk})$. The above expression simplifies to

$$Cov\{\tilde{f}_i(s,t), \tilde{f}_i(s',t')\} = \sum_{j=1}^{q_2}\sum_{k=1}^{q_1}\psi_j(s)\psi_j(s')\phi_k(t)\phi_k(t')h_{jk}$$

When $\psi_j(s)$ and $\phi_k(t)$ are chosen to be marginal eigenfunctions of $K_{\mathcal{S}}(s,s')$ and $K_{\mathcal{T}}(t,t')$ respectively, this expression matches the form of a weakly separable covariance (equation 3 Lynch and Chen [2018a]). However, instead of setting $\psi_j(s)$ and $\phi_k(t)$ as marginal eigenfunctions we choose to expand these terms through b-spline expansions. The resulting $\psi_j(s)$ and $\phi_k(t)$ will not be mutually orthogonal, but this is not a primary concern because one can orthogonalize these basis functions through post-processing posterior samples.

Appendix 2B: Gibbs Sampling

In what follows, let $A_{\cdot j}$ and $A_{k\cdot}$ denote the j^{th} column and k^{th} row of matrix A respectively. Let Σ^* a $p_1 \times p_2$ matrix, obtained by extracting the diagonal of Σ. Entries of Σ^* are filled

in column-wise. Let $\Phi(\cdot)$ denote the standard normal cumulative density function and let Gamma* denote the Gamma probability density function (rate parameterization).

- Step 1 Update of Λ:

  1. Update Λ each row at a time by

$$\begin{aligned}
&\pi(\Lambda_{j\cdot}|-) \sim \mathcal{N}_{q_1}(\mu_n, \Lambda_j^{-1}) \\
&\mu_j = \Lambda_j^{-1} \sum_{i=1}^{N} \boldsymbol{\eta}_i \Gamma^\top \text{diag}\{(\Sigma_{j\cdot}^*)^{-1}\}\Theta_{j\cdot i} \\
&\Lambda_j = \sum_{i=1}^{N} \boldsymbol{\eta}_i \Gamma^\top \text{diag}\{(\Sigma_{j\cdot}^*)^{-1}\}\Gamma\boldsymbol{\eta}_i^\top + D_j^{-1} \\
&D_j^{-1} = \text{diag}(\rho_{1j1}\tau_{11}, \ldots, \rho_{1jq_1}\tau_{1q_1})
\end{aligned}$$

     for $j = 1, \ldots, p_1$

  2. Sample ρ_{1jh} from

$$\pi(\rho_{jh}|-) \sim \text{Gamma}\left(\frac{\nu_1 + 1}{2}, \frac{\nu_1}{2} + \frac{\tau_{1h}\lambda_{jh}^2}{2}\right)$$

  3. Sample δ_{11} from

$$\pi(\delta_{11}|-) \sim \text{Gamma}\left(a_{11} + \frac{p_1 q_1}{2}, 1 + \frac{1}{2}\sum_{l=1}^{q_1} \tau_{1l}^{(1)} \sum_{j=1}^{p_1} \rho_{1jl}\lambda_{jl}^2\right)$$

  4. Sample δ_{1h} from

$$\pi(\delta_{1h}|-) \sim \text{Gamma}\left(a_{12} + \frac{p_1}{2}(q_1 - h + 1), 1 + \frac{1}{2}\sum_{l=h}^{q_1} \tau_{1l}^{(h)} \sum_{j=1}^{p_1} \rho_{1jl}\lambda_{jl}^2\right)$$

for $h = 2, \ldots, q_1$, where $\tau_{1l}^{(h)} = \prod_{t=1, t\neq h}^{l} \delta_t$.

5. Sample a_{11}: Draw a uniform random variable u from $U(0,1)$. Draw a random variable x from $N(0,1)1(x+a_{11} > 0)$. Set the proposal of a_{11} equal to $a_{11}^* = x+a_{11}$. Compute

$$A = \frac{\text{Gamma}^*(\delta_{11}, a_{11}^*, 1)\text{Gamma}^*(a_{11}^*, r_1, 1)\Phi(a_{11})}{\text{Gamma}^*(\delta_{11}, a_{11}, 1)\text{Gamma}^*(a_{11}, r_1, 1)\Phi(a_{11}^*)}$$

where Gamma* denotes the Gamma probability density function (rate parameterization). Accept a_{11}^* when $A > u$.

6. Sample a_{12}: Draw a uniform random variable u from $U(0,1)$ Draw a random variable x from $N(0,1)1(x+a_{12} > 0)$. Set the proposal of a_{12} equal to $a_{12}^* = a_{12}+x$. Compute

$$A = \frac{\text{Gamma}^*(a_{12}^*, r_2, 1)\prod_{h=2}^{q_1}\text{Gamma}^*(\delta_{1h}, a_{12}^*, 1)\Phi(a_{12})}{\text{Gamma}^*(a_{12}, r_2, 1)\prod_{h=2}^{q_1}\text{Gamma}^*(\delta_{2h}, a_{12}, 1)\Phi(a_{12}^*)}$$

where Gamma* denotes the Gamma probability density function (rate parameterization). Accept a_{12}^* when $A > u$.

- Step 2 Update of Γ. This is analogous to the update of Λ.

1. Update Γ each row at a time by

$$\begin{aligned}
&\pi(\Gamma_{j\cdot}|-) \sim N_{q_2}(\mu_j, \Lambda_j^{-1}) \\
&\mu_j = \Lambda_j^{-1}\sum_{i=1}^{N}\boldsymbol{\eta}_i^\top\Lambda^\top\text{diag}\{(\Sigma_{\cdot j}^*)^{-1}\}\Theta_{\cdot ji} \\
&\Lambda_j = \sum_{i=1}^{N}\boldsymbol{\eta}_i^\top\Lambda^\top\text{diag}\{(\Sigma_{\cdot j}^*)^{-1}\}\Lambda\boldsymbol{\eta}_i + (D_j^*)^{-1} \\
&(D_j^*)^{-1} = \text{diag}(\rho_{2j1}\tau_{21}, \ldots, \rho_{2jq_2}\tau_{2q_2})
\end{aligned}$$

for $j = 1, \ldots, p_2$

2. Sample ρ_{2jh} from

$$\pi(\rho_{2jh}|-) \sim \text{Gamma}\left(\frac{\nu_2+1}{2}, \frac{\nu_2}{2} + \frac{\tau_{2h}\gamma_{jh}^2}{2}\right)$$

3. Sample δ_{21} from

$$\pi(\delta_{21}|-) \sim \text{Gamma}\left(a_{21} + \frac{p_2 q_2}{2}, 1 + \frac{1}{2}\sum_{l=1}^{q_2} \tau_{2l}^{(1)} \sum_{j=1}^{p_1} \rho_{2jl}\gamma_{jl}^2\right)$$

4. Sample δ_{2h} from

$$\pi(\delta_{2h}|-) \sim \text{Gamma}\left\{a_{22} + \frac{p^2}{2}(q^2 - h + 1), 1 + \frac{1}{2}\sum_{l=h}^{q_1} \tau_{2l}^{(h)} \sum_{j=1}^{p_1} \rho_{2jl}\gamma_{jl}^2\right\}$$

for $h = 2, \ldots, q_2$, where $\tau_{2l}^{(h)} = \prod_{t=1, t\neq h}^{l} \delta_{2t}$

5. Sample a_{21}: Draw a uniform random variable u from $U(0,1)$. Draw a random variable x from $N(0,1)1(x+a_{21} > 0)$. Set the proposal of a_{21} equal to $a_{21}^* = x + a_{21}$. Compute

$$A = \frac{\text{Gamma}^*(\delta_{21}, a_{21}^*, 1)\text{Gamma}^*(a_{21}^*, r_1, 1)\Phi(a_{21})}{\text{Gamma}^*(\delta_{21}, a_{21}, 1)\text{Gamma}^*(a_{21}, r_1, 1)\Phi(a_{21}^*)}$$

where Gamma^* denotes the Gamma probability density function (rate parameterization). Accept a_{21}^* when $A > u$.

6. Sample a_{22}: Draw a uniform random variable u from $U(0,1)$ Draw a random variable x from $N(0,1)1(x+a_{22} > 0)$. Set the proposal of a_{22} equal to $a_{22}^* = a_{22} + x$. Compute

$$A = \frac{\text{Gamma}^*(a_{22}^*, r_2, 1)\prod_{h=2}^{q_2}\text{Gamma}^*(\delta_{2h}, a_{22}^*, 1)\Phi(a_{22})}{\text{Gamma}^*(a_{22}, r_2, 1)\prod_{h=2}^{q_1}\text{Gamma}^*(\delta_{2h}, a_{22}, 1)\Phi(a_{22}^*)}$$

where Gamma^* denotes the Gamma probability density function (rate parameterization). Accept a_{22}^* when $A > u$.

- Step 3 Update of Σ^*:

$$\pi\{(\Sigma^*_{kj})^{-1}|-\} \sim \text{Gamma}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{N}(\Theta_{kji} - \Lambda_{k\cdot}\boldsymbol{\eta}_i\Gamma_{j\cdot}^\top)\right\}$$

- Step 4 Update of φ^{-2}: Let n_{tot} denote the total number of observations.

$$\pi(\varphi^{-2}|-) \sim \text{Gamma}(a_\varphi + \frac{n_{tot}}{2}, b_\varphi + \frac{1}{2}\sum_{i=1}^{N}[y_i\{\boldsymbol{s},\boldsymbol{t}\} - \{B_1(\boldsymbol{s}) \otimes B_2(\boldsymbol{t})\}\text{vec}\{\Theta_i\}]^\top[y_i\{\boldsymbol{s},\boldsymbol{t}\} - \{B_1(\boldsymbol{s}) \otimes B_2(\boldsymbol{t})\}\text{vec}\{\Theta_i\}]$$

- Step 5 Update of $\boldsymbol{\eta}_i$:

$$\begin{aligned}
&\pi\{\text{vec}(\boldsymbol{\eta}_i)|-\} \sim N_{q_1\cdot q_2}(\mu_i, \Lambda_i^{-1}) \\
&\mu_i = \Lambda_i^{-1}(\Gamma^\top \otimes \Lambda^\top)\Sigma^{-1}\text{vec}(\Theta_i) \\
&\Lambda_i = (\Gamma^\top \otimes \Lambda^\top)\Sigma^{-1}(\Gamma \otimes \Lambda)
\end{aligned}$$

- Step 6 Update of Θ_i:

$$\begin{aligned}
&\pi\{\text{vec}(\Theta_i)|-\} \sim N(\mu_i, \Lambda_i^{-1}) \\
&\mu_i = \Lambda_n^{-1}[\varphi^{-2}\{B_1(\boldsymbol{s}) \otimes B_2(\boldsymbol{t})\}y_i\{\boldsymbol{s},\boldsymbol{t}\} + \Sigma^{-1}\{\Gamma \otimes \Lambda\}\text{vec}\{\eta_i\}] \\
&\Lambda_i = \varphi^{-2}\{B_1\boldsymbol{s})^\top B_1(\boldsymbol{s}) \otimes B_2(\boldsymbol{t})^\top B_2(\boldsymbol{t})\} + \Sigma^{-1}
\end{aligned}$$

- Step 7 Update of H: Treat H as a matrix with dimensions $q_1 \times q_2$ and treat β as a 3-dimensional array with dimensions q_1, q_2 d.

$$\pi(h_{jk}^{-1}|-) \sim \text{Gamma}\left\{a_h + \frac{N}{2}, b_h + \frac{1}{2}\sum_{i=1}^{N}(\eta_{ijk} - \beta_{jk\cdot}x_i)^2\right\}$$

- Step 8: Update of β: Let $\nu_{jk} = (\eta_{1jk}, \ldots, \eta_{Njk})^\top$. Let X be of dimension $N \times d$

obtained by row-stacking x_i.

$$\beta_{jk\cdot} \sim N_d[h_{jk}^{-1} X^\top \nu_{jk}, \{h_{jk}^{-1} X^\top X + \text{diag}(\omega_{jk\cdot})^{-1}\}^{-1}]$$

where $\omega_{jk\cdot}$ is the corresponding d-dimensional vector of entries in E, the prior variance entries for β.

- Step 9 Update of E: Treat E as an array with dimensions q_1, q_2, d.

$$\pi(E_{jkd}|-) \sim \text{Gamma}\{1, 1/2 \cdot (1 + \beta_{jkd}^2)\}$$

Appendix 2C: Simulation Details and Additional Results

In this section we remind the reader of the simulation scheme and provide additional details. We performed a numerical experiment to asses mean and covariance estimation. We use an equally spaced dense grid of 10 longitudinal time points and 20 functional time points with $s \in [0, 1]$ and $t \in [0, 1]$. We study three cases. The three cases are

1. $K_{\mathcal{S}}(s, s') = \sum_{j=1}^{2} \lambda_j \psi_j(s) \psi_j(s')$ with eigenvalues $\lambda_j = \frac{1}{j^2 \pi^2}$ and eigenfunctions
   $\psi_j(s) = \sqrt{2} \sin(j\pi s)$
   $K_{\mathcal{T}}(t, t') = \sigma^2 \left(1 + \frac{\sqrt{3}|t-t'|}{\rho}\right) \exp\left(-\frac{\sqrt{3}|t-t'|}{\rho}\right)$.
   $\mu(s, t) = \sqrt{\frac{1}{5\sqrt{s+1}}} \sin(5t)$.
   Note: $K_{\mathcal{T}}(t, t')$ has the form of a Matèrn covariance function.

2. $K_{\mathcal{S}}(s, s') = \sum_{j=1}^{2} \lambda_j \psi_j(s) \psi_j(s')$ with eigenvalues $\lambda_j = \frac{1}{(j-1/2)^2 \pi^2}$ and eigenfunctions
   $\psi_j(s) = \sqrt{2} \sin\{(j - 1/2)\pi s\}$.
   $K_{\mathcal{T}}(t, t') = \sum_{k=1}^{50} \lambda_k \phi_k(t) \phi_k(t')$ where $\lambda_k = k^{-2\alpha}$ and $\phi_k(t) = \cos(k\pi t)$.
   $\mu(s, t) = 5\sqrt{1 - (s - .5)^2 - (t - .5)^2}$.
   Note: $K_{\mathcal{S}}(s, s')$ is the Brownian motion covariance function used in Xiao et al. [2016].

3. $K\{(s, t), (s', t')\} = \frac{1}{(t - t')^2 + 1} \exp\left\{-\frac{(s - s')^2}{(t - t')^2 + 1}\right\}$.
   $\mu(s, t) = \sqrt{1 + \sin(\pi s) + \cos(\pi t)}$.

Note: this covariance function is a special case of a stationary non-separable covariance function Gneiting [2002].

The data generating truth for cases 1 and 2 can be described by the following manner. Let $B^{(1)}(s)$ be a $10 \times p_2$ matrix, where each column contains a b-spline evaluated at 10 equally spaced points from 0 to 1. Let $e_1(s), e_2(s), \ldots, e_{q_2}(s)$ be the first q_2 eigenfunctions of $K_{\mathcal{S}}$ and let $\xi_1, \xi_2, \ldots, \xi_{q_2}$ be the first q_2 eigenvalues of $K_{\mathcal{S}}(s, s')$ evaluated over a 10×10 grid of equally spaced points. Set the jth column ($\Gamma_{\cdot j}$) as the least squares solution to $\sqrt{\xi_j} e_j(t) = B^1(s)\Gamma_{\cdot j}$. Once Γ is generated, it will hold that $K_{\mathcal{S}}(s, s') \approx B^{(1)}(s)\Gamma\Gamma^\top B^{(1)}(s')^\top$ where Γ is $p_2 \times q_2$. A similar method is used to generate Λ so that Λ is of dimension $p_1 \times q_1$. Next, we set $h_{jk} = \exp(-\sqrt{.01j + .1k})$ and let $\boldsymbol{h} = (h_{11}, h_{12}, \ldots, h_{q_1,q_2})'$. Then $H = \text{diag}(\text{vec}(\boldsymbol{h}))$. Finally, the data generating covariance function is $K\{(s,t),(s',t')\} = \{B^{(1)}(s)\Gamma \otimes B^{(2)}\Lambda B^{(1)}(t)\}H\{B^{(1)}(s')\Gamma \otimes B^{(2)}(t')\Lambda\}^\top$. The mean and covariance functions $\mu(s,t)$ and $K\{(s,t),(s',t')\}$ in case 3 are not projected on b-splines and are simply evaluated on the longitudinal functional domain. In our simulations we set $\sigma^2 = 1$, $\rho = 0.5$, and $\alpha = 0.5$. $B^{(1)}(s)$ is chosen to be a cubic b-spline with knots at $s = (1/3, 2/3)$, resulting in a 10×6 basis matrix. $B^{(2)}(t)$ is also chosen to be a b-spline with knots at $t = (1/3, 2/3)$, resulting in a 20×6 basis matrix. The simulations have $(q_1, q_2) = (3, 3)$ as the generating truth, and we fit using $(q_1, q_2) = (6, 6)$ and $(p_1, p_2) = (8, 9)$. In doing so, we simply enlarge the model space and assess the model's ability to regularize a misspecified model. Hyper-parameters are set as follows: $\nu_1 = 5$, $\nu_2 = 5$, $r_1 = 1$, $r_2 = 2$, $a_\sigma = .5$, $b_\sigma = .5$, $a_h = 1$, $b_h = 1$, $a_\varphi = .0001$, and $b_\varphi = .0001$. For each simulation, we compute the mean, covariance surace, marginal covariance surfaces, and two eigenfunctions associated with the marginal covariances with the two largest eigenvalues. We use sample sizes $n = 30$ and $n = 60$ to compare finite sample properties. For all quantities of interest we report relative mean integrated squared error. For a function f with domain D and estimator $\hat{f}$, this means $RE(\hat{f}, f) = \int_D \{\hat{f}(u) - f(u)\}^2 du / \int_D f^2(u) du$. We run 1000 simulations for each sample size for each case and report the 50%, 10%, and 90% quantiles of $RE(\hat{f}, f)$. In addition to the metrics reported in the main paper, we also compare the Bayesian method to the product FPCA in estimating marginal covariance functions ($K_{\mathcal{S}}(s, s')$, $K_{\mathcal{T}}(t, t')$), two principal longitudinal eigenfunctions ($\psi_j(s)$,

Table A2.1: Numerical experiment comparing the proposed method to the Product FPCA for estimating functionals of the covariance structure for case 1. We report the 50th percentile of the relative error, with the numbers in the parantheses denoting the 10th and 90th percentiles of the relative error.

		Bayes	Product
$n = 30$	$K_{\mathcal{S}}(s, s')$	.032 (.009, .149)	.052 (.031, .126)
	$K_{\mathcal{T}}(t, t')$	.044 (.013, .170)	.039 (.012, .132)
	$\psi_1(s)$	.003 (.001, .013)	.006 (.004, .014)
	$\psi_2(s)$	.012 (.003, .034)	.013 (.004, .042)
	$\phi_1(t)$	.006 (.001, .030)	.010 (.001, .043)
	$\phi_2(t)$	.017 (.005, .051)	.022 (.006, .063)
$n = 60$	$K_{\mathcal{S}}(s, s')$	.015 (.004, .061)	.040 (.028, .086)
	$K_{\mathcal{T}}(t, t')$	.020 (.006, .074)	.023 (.008, .076)
	$\psi_1(s)$	.002 (.000, .005)	.005 (.003, .009)
	$\psi_2(s)$	.005 (.002, .017)	.006 (.006, .018)
	$\phi_1(t)$	.003 (.001, .014)	.005 (.001, .022)
	$\phi_2(t)$	.008 (.002, .021)	.014 (.005, .034)

$j = 1, 2$) and two principal functional eigenfunctions ($\phi_k(t)$, $k = 1, 2$). Results are listed in Tables A2.1, A2.2, A2.3 for cases 1, 2, and 3 respectively. Across all three cases the Bayesian method and product FPCA have similar relative errors with no clear preferred method for point estimates.

We conduct a small simulation aimed at assessing the performance of the proposed information criteria in Appendix D. We considered the following data-generating mechanism: covariance case 2, 20 longitudinal points, 20 functional points, $N = 30$, $(p_1, p_2) = (10, 10)$, $(q_1, q_2) = (4, 4)$, and $\varphi^2 = .025$. We fit candidate models with $(p_1, p_2) = (5, 5), (10, 10)$, and $(15, 15)$. We keep the number of latent factors as $(4, 4)$ in estimation, as the model is robust to the number of latent factors, due to adaptive penalization. Table A2.4 displays averaged information criteria over 1,000 simulations. The $(p_1, p_2) = (10, 10)$ row contains the smallest information criteria across all three metrics, giving strong indication that several alternative criteria tend to select an appropriate number of basis functions.

We perform a simulation to assess sensitivity to p_1, p_2, q_1, and q_2 in the context of mean and covariance point estimation and coverage rates. In this experiment, we use 20 longitudinal and 20 functional points with a sample size of 60. The true model has $(p_1, p_2) = (12, 12)$ and $(q_1, q_2) = (3, 3)$. We fit using $(p_1, p_2) = (10, 10), (12, 12), (14, 14)$ and $(16, 16)$. We denote

Table A2.2: Numerical experiment comparing the proposed method to the Product FPCA for estimating functionals of the covariance structure for case 2. We report the 50th percentile of the relative error, with the numbers in the parantheses denoting the 10th and 90th percentiles of the relative error.

		Bayes	Product
$n = 30$	$K_{\mathcal{S}}(s, s')$	.032 (.007, .164)	.040 (.006, .167)
	$K_{\mathcal{T}}(t, t')$	.028 (.005, .145)	.035 (.005, .155)
	$\psi_1(s)$	.002 (.000, .008)	.006 (.004, .014)
	$\psi_2(s)$	.015 (.004, .057)	.026 (.010, .070)
	$\phi_1(t)$	.001 (.000, .004)	.005 (.004, .009)
	$\phi_2(t)$	.015 (.006, .040)	.040 (.025, .063)
$n = 60$	$K_{\mathcal{S}}(s, s')$	.016 (.003, .080)	.019 (.003, .078)
	$K_{\mathcal{T}}(t, t')$	.015 (.002, .071)	.017 (.004, .069)
	$\psi_1(s)$	.001 (.000, .003)	.005 (.003, .009)
	$\psi_2(s)$	.007 (.002, .030)	.019 (.010, .049)
	$\phi_1(t)$	.000 (.000, .002)	.005 (.004, .007)
	$\phi_2(t)$	.006 (.002, .014)	.037 (.027, .049)

Table A2.3: Numerical experiment comparing the proposed method to the Product FPCA for estimating functionals of the covariance structure for case 3. We report the 50th percentile of the relative error, with the numbers in the parantheses denoting the 10th and 90th percentiles of the relative error.

		Bayes	Product
$n = 30$	$K_{\mathcal{S}}(s, s')$	.036 (.008, .146)	.035 (.007, .141)
	$K_{\mathcal{T}}(t, t')$	.030 (.005, .138)	.032 (.006, .127)
	$\psi_1(s)$	.003 (.000, .016)	.005 (.002, .020)
	$\psi_2(s)$	.005 (.001, .018)	.024 (.019, .041)
	$\phi_1(t)$	.002 (.000, .010)	.003 (.001, .011)
	$\phi_2(t)$	.005 (.001, .015)	.010 (.005, .021)
$n = 60$	$K_{\mathcal{S}}(s, s')$	.020 (.004, .073)	.019 (.004, .070)
	$K_{\mathcal{T}}(t, t')$	.017 (.003, .067)	.016 (.003, .065)
	$\psi_1(s)$	.002 (.000, .008)	.003 (.002, .011)
	$\psi_2(s)$	.003 (.000, .009)	.021 (.019, .029)
	$\phi_1(t)$	.001 (.000, .005)	.002 (.001, .006)
	$\phi_2(t)$	.002 (.001, .008)	.007 (.005, .013)

Table A2.4: Information criteria for case 2. Each (p_1, p_2) combination is repeated 1000 times. The table reports the .5, .1, and .9 quantiles of the information criteria over 1000 simulations. Each number is on the 10^4 scale.

p_1, p_2	DIC	BIC 1	BIC 2
(5, 5)	1.29 (1.20, 1.37)	1.49 (1.40, 1.57)	1.59 (1.51, 1.67)
(10, 10)	1.22 (1.14, 1.29)	1.40 (1.32, 1.47)	1.50 (1.43, 1.57)
(15, 15)	1.24 (1.16, 1.31)	1.41 (1.33, 1.48)	1.52 (1.43, 1.59)

Table A2.5: Mean integrated squared errors for $q_1, q_2 = (3, 3)$

Model	$\mu(s,t)$	$\psi_1(s)$	$\psi_2(s)$	$\phi_1(t)$	$\phi_2(t)$
50%	.014	.006	.064	.006	.021
60%	.015	.006	.030	.006	.019
70%	.016	.006	.021	.006	.019
80%	.016	.007	.017	.006	.019

these models as 50%, 60%, 70%, and 80% respectively. We also fit using $(q_1, q_2) = (3, 3)$ and $(q_1, q_2) = (6, 6)$. Measurement error variance is set to .025 and we simulate from case 1 (see above). We obtain relative mean squared errors and coverage rates (using a 95% nominal value) for $\mu(s,t)$ and the first two marginal eigenfunctions of $K_{\mathcal{S}}(s,s')$ and $K_{\mathcal{T}}(t,t')$. When fit with $(q_1, q_2) = (3, 3)$, Table A2.5 shows the relative mean squared errors do not seem overly sensitive to choice of p_1 and p_2. However, coverage rates from Table A2.6 do not obtain the nominal 95% rate, with coverage rate for the first marginal eigenfunction of $K_{\mathcal{S}}(s,s')$ at .450. In contrast, when fit with $(q_1, q_2) = (6, 6)$, all coverage rates are above the nominal 95%, indicating valid probabilistic inference. Furthermore, the relative mean squared errors in Table A2.7 do not depend on choice of p_1 or p_2. The message of this simulation is that the choice of p_1 or p_2 will not have a big impact on mean and covariance recovery, provided a sufficiently rich set of basis functions is selected, and that that q_1 and q_2 are sufficiently large. In this experiment, each (p_1, p_2) and (q_1, q_2) combination were repeated 500 times. Hyper-parameters are the kept the same from above.

We also designed a simulation study to compare to a related procedure, Longitudinal Functional Principal Components (LFPCA) Greven et al. [2010], in estimation and inference in the global mean $\mu(s,t)$. LFPCA adds a random a random intercept function and random slope function for each longitudinal object, thereby accounting for random variability over repeated visits in a semiparametric fashion. To estimate the global mean, the authors use bivariate penalized smoothing with restricted maximum likelihood (REML). Maximizing this criterion is appealing because it has been found to be relatively robust to misspecification of the working correlation structure Krivobokova and Kauermann [2007]. In this experiment, the design grid is common over $n = 50$ subjects, with 20 equally spaced $s \in [0, 1]$ and 20

equally spaced $t \in [0,1]$, making for 1000 separate curves. The mean function is set to $\mu(s,t) = 0.5(s/4 - t/20)^2$. We generate simulated data from two scenarios: (1) Bayesian data generating model and (2) LFPCA data generating model. The Bayesian data generating model (1) is

$$\begin{aligned}
K_{\mathcal{S}}(s,s') &= \sum_{j=1}^{3} \lambda_j \psi_j(s)\psi_j(s') \\
\lambda_j &= 1/(j^2\pi^2) \\
\psi_1(s) &= \sqrt{2}\cos(2\pi s) \\
\psi_2(s) &= \sqrt{2}\sin(2\pi s) \\
\psi_3(s) &= \sqrt{2}\cos(4\pi s) \\
K_{\mathcal{T}}(t,t') &= \left(1 + \frac{\sqrt{3}|t-t'|}{0.5}\right)\exp\left(-\frac{\sqrt{3}|t-t'|}{0.5}\right)
\end{aligned}$$

The rest of the Bayesian data generating is specified as in the previous simulations for cases 1, 2, and 3. The LFPCA data generating model is

$$Y_i(s,t) = \mu(s,t) + X_{i,0}(t) + s \cdot X_{i,1}(t) + U_i(s,t) + \epsilon_i(s,t)$$

LFPCA induces dimension reduction by setting $X_{i,0}(t) = \sum_{k=1}^{N_X} \xi_{ik}\phi_k^0(t)$, $X_{i,1}(t) = \sum_{k=1}^{N_X} \xi_{ik}\phi_k^1(t)$, and $U_i(s,t) = \sum_{k=1}^{N_U} \zeta_{isk}\phi_k^U(t)$. We set $N_X = N_U = 4, \xi_{ik} \sim N(0,\lambda_k)$, $\zeta_{isk} \sim N(0,\nu_k)$ for all i, s, and k, where $\lambda_k = \nu_k = 0.5^{k-1}$ The eigenfunctions are

$$\begin{array}{lll}
\phi_1^0(t) = \sin(2\pi t) & \phi_1^1(t) = 1/\sqrt{2} & \phi_1^U(t) = 1 \\
\phi_2^0(t) = \cos(2\pi t) & \phi_2^1(t) = \sin(6\pi t) & \phi_2^U(t) = \sqrt{3}(2t-1) \\
\phi_3^0(t) = \sin(4\pi t) & \phi_3^1(t) = \cos(6\pi t) & \phi_3^U(t) = \sqrt{5}(6t^2 - 6t + 1) \\
\phi_4^0(t) = \cos(4\pi t) & \phi_4^1(t) = \sin(8\pi t) & \phi_4^U(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1)
\end{array}$$

For both data generating schemes $\epsilon_i(s,t)$ is normally distributed with mean zero and standard deviation .05. There are six scenarios in this experiment:

1. B-B: Data generated from Bayes model, fit by Bayes model.

2. B-R: Data generated from Bayes model, fit by REML. Inference performed by asymptotic standard errors given from the R package `mgcv` Wood [2006].

3. B-R-Boot: Data generated from Bayes model, fit by REML. Inference performed by 100 bootstrap replicates.

4. LFPCA-B: Data generated from LFPCA, estimation and inference performed by Bayes model.

5. LFPCA-R: Data generated from LFPCA, fit by REML. Inference performed as in scenario 2.

6. LFPCA-R-Boot: Data generated from LFPCA, fit by REML. Inference performed by 100 bootstrap replicates.

REML fitting is performed via the *gamm* function in the R package `mgcv`. Both the Bayesian method and REML use 12 cubic b-splines for nonparametric estimation. To evaluate estimation accuracy, we compute $\int_S \int_T \{\hat{\mu}(s,t) - \mu(s,t)\}^2 \, dtds / \int_S \int_T \mu(s,t)^2 \, dtds$. To evaluate inference, we examine pointwise coverage at the nominal 5% type I error rate. Due to the symmetric nature of the mean function function, we use symmetric bounds computed by $\hat{\mu}(s,t) \pm \alpha^* \text{SD}\{\hat{\mu}(s,t)\}$, where $\text{SD}\{\hat{\mu}(s,t)\}$ is estimated by sampling the posterior distribution (5000 iterations with 500 burnin), asymptotic standard errors as in `mgcv`, or 100 bootstrap replicates. Each scenario is completed 1000 times. Table A2.9 reports the associated estimation errors and coverage rates.

The Bayesian method has comparable relative error compared to the REML fit whether the data is generated from the Bayesian data generating model or the LFPCA data generating model. Clearly relying on `mgcv` standard errors will yield pointwise confidence intervals which are too narrow. Either using a bootstrap or MCMC will correct this issue. Since mean estimation is largely unaffected by choice of fitting method, users should decide between the proposed method or LFPCA depending on their analytic goals. For example, the proposed

Table A2.6: Coverage for $q_1, q_2 = (3, 3)$

Model	$\mu(s,t)$	$\psi_1(s)$	$\psi_2(s)$	$\phi_1(t)$	$\phi_2(t)$
50%	.960	.620	.238	.974	.720
60%	.928	.562	.494	.984	.742
70%	.902	.526	.632	.982	.750
80%	.834	.450	.698	.980	.756

Table A2.7: Mean integrated squared errors for $q_1, q_2 = (6, 6)$

Model	$\mu(s,t)$	$\psi_1(s)$	$\psi_2(s)$	$\phi_1(t)$	$\phi_2(t)$
50%	.015	.004	.013	.005	.009
60%	.015	.004	.013	.005	.010
70%	.015	.004	.013	.005	.010
80%	.015	.004	.014	.006	.010

Table A2.8: Coverage for $q_1, q_2 = (6, 6)$

Model	$\mu(s,t)$	$\psi_1(s)$	$\psi_2(s)$	$\phi_1(t)$	$\phi_2(t)$
50%	.992	.982	.972	.982	.970
60%	.996	.976	.974	.972	.970
70%	.994	.976	.978	.978	.964
80%	.988	.980	.972	.980	.962

Table A2.9: Comparing mean estimation and coverage over different data generating mechanisms and fitting methods. B-B refers to data generated from the model in this paper, fit by Bayes. B-R refers to data generated from the model in this paper, fit by REML. B-R-Boot refers to data generated from the model in this paper, fit by REML (bootstrap pointwise confidence intervals). LFPCA-B refers to data generated from LFPCA, fit by the proposed method. LFPCA-R refers to data generated from LFPCA, fit by REML. LFPCA-R-Boot refers to data generated from LFPCA, fit by REML (bootstrap pointwise confidence intervals). The table reports the 50%, 10%, and 90% percentiles over 1000 simulations.

	B-B	B-R	B-R-Boot
$\mu(s,t)$	.039 (.015, .097)	.036 (.011, .096)	.036 (.011, .096)
Coverage	.992 (.835, 1.00)	.418 (.212, .660)	.995 (.818, 1.00)
	LFPCA-B	LFPCA-R	LFPCA-R-Boot
$\mu(s,t)$	.063 (.032, .130)	.043 (.015, .112)	.043 (.032, .112)
Coverage	.978 (.870, 1.00)	.442 (.215, .720)	.998 (.870, 1.00)

method can yield inference on the longitudinal patterns of variation, while LFPCA cannot. Conversely, LFPCA decomposes variation into a random intercept function, random slope function, and a subject-visit random function. Eigen-analyses of these random functions are interpretable but are not obtainable in the framework of the proposed method.

Appendix 2D: Model Selection for Case Studies

In this paper we select an appropriate number of splines by minimizing information criteria. We consider the deviance information criterion (DIC) and two variations of the Bayesian information criterion, BIC and BIC_h Delattre et al. [2014]. The information criteria are calculated as follows. Obtain the posterior-averaged mean function $\hat{\mu}(s,t)$ and the posterior-averaged covariance function $\hat{K}\{(s,t),(s',t')\}$. Evaluate these functions on a grid $\boldsymbol{s}^* \times \boldsymbol{t}^* \in \mathcal{S} \times \mathcal{T}$ common to all subjects in the study. Let the vector $\hat{\mu}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i = \mu(\boldsymbol{x}_i,\boldsymbol{s}^*,\boldsymbol{t}^*)$ denote the conditional mean function for subject i and let $\hat{K}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}$ be the common covariance function over all subjects. Marginalizing over $\boldsymbol{\eta}_i$ and $\boldsymbol{\zeta}_i$, the log-likelihood for subject i is

$$\begin{aligned} l_i = &-\frac{N_i}{2}\log(2\pi) - \frac{1}{2}|\hat{K}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}| \\ &-\frac{1}{2}\{y_i(\boldsymbol{s}^*,\boldsymbol{t}^*) - \hat{\mu}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i\}^\top \hat{K}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}^{-1}\{y_i(\boldsymbol{s}^*,\boldsymbol{t}^*) - \hat{\mu}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i\} \end{aligned}$$

where N_i denotes the total number of functional longitudinal observations for subject i. Let $\hat{\mu}^{(r)}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i$ be the r^{th} sample of the posterior mean function evaluated at observed $\boldsymbol{s}^*$ and $\boldsymbol{t}^*$ for subject i. Analogously, let $\hat{K}^{(r)}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}$ be the r^{th} posterior draw of $K\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}$. Let

$$\begin{aligned} g_i^{(r)} = &-\frac{N_i}{2}\log(2\pi) - \frac{1}{2}|\hat{K}^{(r)}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}| \\ &-\frac{1}{2}(y_i(\boldsymbol{s}^*,\boldsymbol{t}^*) - \hat{\mu}^{(r)}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i)^\top \hat{K}^{(r)}\{(\boldsymbol{s}^*,\boldsymbol{t}^*),(\boldsymbol{s}^{*\prime},\boldsymbol{t}^{*\prime})\}^{-1}\{y_i(\boldsymbol{s}^*,\boldsymbol{t}^*) - \hat{\mu}^{(r)}(\boldsymbol{s}^*,\boldsymbol{t}^*)_i\} \end{aligned}$$

Then DIC is equal to

$$\begin{aligned} DIC &= -2\Big\{ \sum_{i=1}^{N} l_i - 2 \cdot \Big(\sum_{i=1}^{N} l_i - \frac{1}{R} \sum_{r=1}^{R} \sum_{i=1}^{N} g_i^{(r)} \Big) \Big\} \\ &= -2(L - P) \end{aligned}$$

where $L = \sum_{i=1}^{N} l_i$ and $P = 2 \cdot (\sum_{i=1}^{N} l_i - \frac{1}{R} \sum_{r=1}^{R} \sum_{i=1}^{N} g_i^{(r)})$. The quantity L represents a goodness of fit and the quantity P represents the effective number of parameters. BIC is conceptually similar to DIC, except BIC explicitly counts the number of parameters in a model to penalize model complexity. We use two versions of BIC. The first version does not distinguish between random or fixed parameters for counting purposes. The parameters used in this calculation are Γ, Λ, H, β, Σ, and φ^2. The number of paramaters is $n_{par} = p_2 \cdot q_2 + p_1 \cdot q_1 + q_1 \cdot q_2 + q_1 \cdot q_2 + p_1 \cdot p_2 + 1$. BIC is calculated as

$$\text{BIC} = -2L + n_{par} \log(N)$$

BIC_h is a hybrid information criteria which distinguishes between fixed and random parameters. BIC_h is computed as

$$BIC_h = -2L + n_{par-random} \log(N) + n_{par-fixed} \log(n_{tot})$$

where $n_{par-fixed}$ counts the number of fixed parameters, $n_{par-random}$ counts the number of random parameters, and n_{tot} counts the total number of observed longitudinal functional data over all subjects. In the proposed probabilistic model, it's unclear whether Γ and Λ should be counted as fixed parameters or random parameters since both appear in the mean and covariance processes. We choose to count them as fixed parameters. Therefore the random parameters include H, Σ, and φ^2 and the fixed parameters include Γ, Λ, and β. The number of random parameters is $q_1 \cdot q_2 + p_1 \cdot p_2 + 1$ and the number of fixed parameters is $p_2 \cdot q_2 + p_1 \cdot q_1 + q_1 \cdot q_2$.

We now discuss choosing the number of b-spline bases used in estimation for the fertility

and EEG case studies. Recall that $f_i(s,t) = \sum_{j=1}^{q_2} \sum_{k=1}^{q_1} \psi_j(s)\phi_k(t)\eta_{ikj} + r_i(s,t)$, with $\psi_j(s) = \sum_{l=1}^{p_2} \gamma_{lj} b^{(1)}(s)$ and $\phi_k(t) = \sum_{m=1}^{p_1} \lambda_{mk} b^{(2)}(t)$. The prior on Γ and Λ encourages near-zero loadings for basis coefficients of $\psi_j(s)$ and $\phi_k(t)$, while simultaneously able to drop out entire $\psi_j(s)$ and $\phi_k(t)$. Hence, this prior is robust against specifying q_1 and q_2 too large. However, we noticed sensitivity in smoothness of $f_i(s,t)$ by varying p_1 and p_2, which are the degrees of freedom for $b^{(2)}(t)$ and $b^{(1)}(s)$ respectively. Selecting degrees of freedom for b-splines can be accomplished via reversible jump MCMC Green [1995], latent indicators Thompson and Rosen [2008], among many other methods. In this paper we choose to select the degrees of freedoms by minimizing information criteria. We use the deviance information criterion (DIC), and two versions of the Bayesian information criterion (BIC & BIC_h).

Model selection for the fertility data is as follows. We vary p_1 and p_2 from $(p_1, p_2)^\top = (11, 14)^\top$ to $(p_1, p_2)^\top = (44, 56)^\top$. We also let q_1 and q_2 grow with p_1 and q_2. In total we compute information criteria for six models. Information criteria values are based on 5,000 iterations of 4 independent Markov chains, after discarding 1,000 draws for burn-in. The model with $(p_1,\, p_2,\, q_1,\, q_2)^\top = (22, 28, 11, 10)^\top$ minimizes all three information criteria, giving strong indication that this model balances the goodness of fit and model complexity moreso than the other five models. This model has a moderate number of degrees of freedom, indicating that setting p_1 and p_2 too high can lead to overfitting. This is not suprising because the fertility rate varies as a smooth function as a function of age and calendar year.

Model selection for the EEG case study uses the same number of iterations, chains, and burn-in. We let p_1 and p_2 vary from $(p_1, p_2)^\top = (10, 14)^\top$ to fourteen more complex models. The model with $(p_1, p_2, q_1, q_2)^\top = (20, 56, 10, 28)^\top$ minimized all three information criteria in both TD and ASD groups, giving evidence that this model seems to balance goodness of fit and model complexity. The high degrees of freedom for trial time indicates that condition differentiation varies rapidly from trial to trial. However, the moderate number of degrees of freedom for within trial time indicates that the within trial condition differentiation changes more smoothly. This is expected because this data is preprocessed with a 30 trial sliding window as described in the main text.

Table A2.10: Model description and information criteria for fertility data. The black box highlights the model with the smallest information criteria.

Model	(p_1, p_2)	(q_1, q_2)	DIC	BIC	BIC_h
1	(11, 14)	(6,5)	-79584	-70976	-66485
2	(15, 19)	(8,8)	-125732	-116136	-109381
3	(22, 28)	(11, 10)	-145781	-135006	-125986
4	(29, 37)	(13, 14)	-142638	-130166	-118147
5	(35, 45)	(16, 18)	-134634	-120161	-104534
6	(44, 56)	(18, 20)	-125504	-109257	-91506

Table A2.11: Model description and information criteria for the TD group in the EEG implicit learning study.

Model	(p_1, p_2)	(q_1, q_2)	DIC	BIC	BIC_h
1	(10, 14)	(5,7)	32386	40921	45584
2	(13, 19)	(6,9)	16767	25186	31128
3	(19, 28)	(9, 14)	-229	8717	18181
4	(25, 38)	(12, 19)	-16421	-6618	6596
5	(28, 42)	(14, 21)	-25416	-15675	-543
6	(30, 45)	(15, 22)	-29894	-19915	-3803
7	(37, 56)	(18, 28)	-65231	-33544	-54380
8	(37, 15)	(18, 7)	41286	50165	59179
9	(37, 20)	(18, 10)	28260	36967	47670
10	(37, 25)	(18, 12)	17101	25378	37207
11	(10, 56)	(5, 28)	-60983	-48876	-34468
12	(15, 56)	(7, 28)	-89911	-78659	-63262
13	(20, 56)	(10, 28)	-96729	-85843	-68962
14	(25, 56)	(12, 28)	-85646	-74844	-56974
15	(30, 56)	(15, 28)	-76871	-65849	-46496

Table A2.12: Model description and information criteria for the ASD group in the EEG implicit learning study.

Model	(p_1, p_2)	(q_1, q_2)	DIC	BIC	BIC_h
1	(10, 14)	(5,7)	47203	55873	60539
2	(13, 19)	(6,9)	28813	37426	43370
3	(19, 28)	(9, 14)	4003	11895	21364
4	(25, 38)	(12, 19)	-17713	-8591	4630
5	(28, 42)	(14, 21)	-22473	-12891	2247
6	(30, 45)	(15, 22)	-26997	-16926	-806
7	(37, 56)	(18, 28)	-67619	-57589	-36741
8	(37, 15)	(18, 7)	56749	65547	74566
9	(37, 20)	(18, 10)	34387	42837	53546
10	(37, 25)	(18, 12)	24240	32226	44061
11	(10, 56)	(5, 28)	-30508	-18227	-3812
12	(15, 56)	(7, 28)	-76282	-64538	-49133
13	(20, 56)	(10, 28)	-102624	-91275	-74386
14	(25, 56)	(12, 28)	-90221	-79561	-61682
15	(30, 56)	(15, 28)	-80708	-70094	-50731

Appendix 2E: Missing and Sparse Functional Data

The main text presents a model with data observed on a common grid for simplicity. In this section we discuss accounting for missing or sparse functional responses. There are three cases to consider: missingness over longitudinal time (case 1), missingness over functional time (case 2), and missingness over both longitudinal and functional time (case 3). All three cases can be handled by imputing missing observations during sampling, and this is the approach we take in the EEG case study which had a total of 37 trials missing for the ASD group and 77 trials missing for the TD group.

Let $\tilde{s}_i$ denote the missing longitudinal times for subject i. Let $B_{i1}(\tilde{s})$ denote the rows of $B_1(s)$ corresponding to longitudinal times $\tilde{s}_i$. For case 1, we impute $Y_i(\tilde{s}_i, t) \sim N((B_{i1}(\tilde{s}_i) \otimes B(t))\text{vec}(\Theta_i), \varphi^2)$. The imputed observations $Y_i(\tilde{s}_i, t)$ are treated as observed data in the Gibbs sampler.

Cases 2 is handled in a slightly different manner. Here we assume each subject has complete longitudinal data but sparse or missing functional data. Suppose subject i has missing functional data $\tilde{t}_{i1}, \ldots, \tilde{t}_{iJ}$ over J longitudinal times. Form the matrix B_i in the

following manner:

$$\begin{pmatrix} B_1(s_1) \otimes B_2(\tilde{t}_{i1}) \\ B_1(s_2) \otimes B_2(\tilde{t}_{i2}) \\ \vdots \\ B_1(s_J) \otimes B_2(\tilde{t}_{iJ}) \end{pmatrix}$$

Then impute the response by sampling from $N(B_i \text{vec}(\Theta_i), \varphi^2)$.

For case 3 the matrix B_i is formed in a similar manner. Suppose subject i has missing longitudinal time $\tilde{s}_i = (s_{i1}, \ldots s_{im})$ and missing functional observations $\tilde{t}_{i1}, \ldots t_{il}$. Let $\hat{s}_i$ be the sorted longitudinal times from least to greatest. The vector $\hat{s}_i$ has length m equal to number of missing longitudinal time plus number of longitudinal times with missing functional observations. Let t be the entire grid of functional time. If $\hat{s}_{ij}$ is an entire missing longitudinal time point, let $\tilde{t}_{ij} = t$. Otherwise, let $\tilde{t}_{ij}$ be the grid of missing functional observations. Construct B_i as

$$\begin{pmatrix} B_1(\hat{s}_{i1}) \otimes B_2(\tilde{t}_{i1}) \\ \vdots \\ B_1(\hat{s}_{im}) \otimes B_2(\tilde{t}_{im}) \end{pmatrix}$$

As in cases 1 and 2, impute the missing data by sampling from $N(B_i \text{vec}(\Theta_i), \varphi^2)$.

Alternatively, one can alter the update for Θ_i to accommodate missing or sparse observations over longitudinal or functional domains. Suppose subject i has observed longitudinal times $s_{i1}, \ldots, s_{in}$ with corresponding observed functional times $\boldsymbol{t_{i1}}, \ldots \boldsymbol{t_{in}}$. Define B_i to be

$$\begin{pmatrix} B_1(s_{i1}) \otimes B_2(\boldsymbol{t_{i1}}) \\ \vdots \\ B_1(s_{in}) \otimes B_2(\boldsymbol{t_{in}}) \end{pmatrix}$$

Sample Θ_i by

$$\pi(\text{vec}(\Theta_i)) \sim N(\mu_i, \Lambda_i^{-1})$$
$$\mu_i = \Lambda_i^{-1}\varphi^{-2}B_i y_i + \Sigma^{-1}(\Gamma \otimes \Lambda)\text{vec}(\eta_i))$$
$$\Lambda_i = \varphi^{-2}B_i^\top B_i + \Sigma^{-1}$$

Appendix 2F: Post-processing MCMC samples

The main text discusses posterior inference for the mean $\mu(\boldsymbol{x}_i, s, t)$. In this section we give details on posterior inference for eigenfunctions of $K_{\mathcal{T}}(t, t')$ and $K_{\mathcal{S}}(s, s')$, defined in the main text. The total covariance $K\{(s,t),(s',t')\}$ can be expressed as

$$K\{(s,t),(s',t')\} = \sum_{j=1}^{q_1}\sum_{k=1}^{q_2} h_{jk}\psi_j(s)\psi_j(s')\phi_k(t)\phi_k(t') + \sum_{j=1}^{p_1}\sum_{k=1}^{p_2}\sigma_{jk}b_j^{(1)}(s)b_j^{(1)}(s')b_k^{(2)}(t)b_k^{(2)}(t')$$

Direct computation of the marginal covariance function $K_{\mathcal{S}}(s, s')$ yields

$$\begin{aligned}
K_{\mathcal{S}}(s,s') &= \int_{\mathcal{T}} K\{(s,t),(s',t)\}dt \\
&= \int_{\mathcal{T}} \{\sum_{j=1}^{q_1}\sum_{k=1}^{q_2} h_{jk}\psi_j(s)\psi_j(s')\phi_k(t)\phi_k(t) + \sum_{j=1}^{p_1}\sum_{k=1}^{p_2}\sigma_{jk}b_j^{(1)}(s)b_j^{(1)}(s')b_k^{(2)}(t)b_k^{(2)}(t)\}dt \\
&= \sum_{j=1}^{q_1}\psi_j(s)\psi_j(s')\sum_{k=1}^{q_2} h_{jk}\int_{\mathcal{T}}\phi_k(t)\phi_k(t)dt + \sum_{j=1}^{p_1}\sigma_{jk}b_j^{(1)}(s)b_j^{(1)}(s')\int_{\mathcal{T}} b_k^{(2)}(t)b_k^{(2)}(t)dt \\
&= \sum_{j=1}^{q_1}\psi_j(s)\psi_j(s')\sum_{k=1}^{q_2} h_{jk}\nu_k + \sum_{j=1}^{p_1}\sigma_{jk}\chi_k b_j^{(1)}(s)b_j^{(1)}(s')
\end{aligned}$$

where $\nu_k = \int_{\mathcal{T}}\phi_k(t)\phi_k(t)dt$ and $\chi_k = \int_{\mathcal{T}} b_k^{(2)}(t)b_k^{(2)}(t)dt$. A similar expression can be derived for $K_{\mathcal{T}}(t, t')$. Computing marginal kernels with the above expression has much better scalability than first forming $K\{(s,t),(s',t')\}$ and then computing the required integrals $\int_{\mathcal{T}} K\{(s,t),(s',t)\}dt$ or $\int_{\mathcal{S}} K\{(s,t),(s,t')\}ds$ directly. In our implementation we use the 'trapz' function from the R package 'pracma' to numerically approximate the above integrals.

Appendix 3A: Markov-Chain Monte Carlo Sampling Algorithm

In this section we give a detailed Markov-Chain Monte Carlo (MCMC) algorithm to sample from the posterior. Let N be the number of independent functional responses and assume all response functions are observed on a common grid $T = \{t_1, \ldots, t_n\}$. Let B be an $n \times p$ matrix with $B_{ij} = b_j(t_i)$. Let X be an $N \times r(d_1)$ matrix with row i equal to $\boldsymbol{b}^x(\boldsymbol{x}_i)$. Let Y be an $N \times n$ matrix with $Y_{ij} = y_i(t_j)$, so that each row represents one discretized functional response. Let Γ_j be an $N \times N$ diagonal matrix with rth diagonal element equal to η_{rj} for $j = 1, \ldots, k$.

1. Update β:

   Let $\Omega_r = \tau_{1xr}\tilde{K}_r + \tau_{1tr}\tilde{K}$ if $p_r > 1$. Otherwise set $\Omega_r = \tau_{1tr}K$. Construct

   $\Omega = \text{blkdiag}(\Omega_1, \ldots, \Omega_R)$.

   Let $C = \sigma^{-2}X^\top X \otimes B^\top B + \Omega$

   Let $A = \sigma^{-2}\text{vec}[\{B^\top Y^\top - B^\top B(\sum_{j=1}^k \Lambda_j X^\top \Gamma_j)\}X]$

   Sample $\text{vec}(\beta) \sim N(C^{-1}A, C^{-1})$

2. Update Λ_j:

   Let $\Omega_r = \tau_{2xr}\tilde{K}_r + \tau_{2tr}\tilde{K} + \tau^*_{rj}\phi_{rj}$ if $p_r > 1$. Otherwise set $\Omega_r = \tau_{2tr}K + \tau^*_{rj}\phi_{rj}$.

   Construct $\Omega = \text{blkdiag}(\Omega_1, \ldots, \Omega_R)$

   Let $C = \sigma^{-2}X^\top \Gamma_j^2 X \otimes B^\top B + \Omega$

   Let $A = \sigma^{-2}\text{vec}[\{B^\top Y^\top - B^\top B(\beta + \sum_{j' \neq j} \Lambda_{j'} X^\top \Gamma_{j'})\}\Gamma_j X]$

   Sample $\Lambda_j \sim N(C^{-1}A, C^{-1})$

3. Update η_{ij}:

   Let $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ik})$.

   Let $\ddot{X}_i$ be a $p \times k$ matrix with column j equal to $\Lambda_j \boldsymbol{b}^x(\boldsymbol{x}_i)$.

   Let $C = \sigma^{-2}\ddot{X}_i^\top B^\top B\ddot{X}_i + I_k$, where I_k is the $k \times k$ identity matrix.

   Let $A = \sigma^{-2}\ddot{X}_i^\top\{B^\top Y_{i\cdot} - B^\top B\beta\boldsymbol{b}^x(\boldsymbol{x}_i)\}$

   Sample $\boldsymbol{\eta}_i \sim N(C^{-1}A, C^{-1})$

4. Update σ^2:

Let $A = Y^\top - B\beta X^\top + \sum_{j=1}^{k} B\Lambda_j X^\top \Gamma_j$

Sample $\sigma^{-2} \sim \text{Gamma}(Nn/2 + a_\epsilon, A \odot A/2 + b_\epsilon)$, where $\odot$ denotes element-wise multiplication.

5. Update τ_{1tr}, τ_{1xr}, τ_{2tr}, τ_{2xr}:

   Sample $\tau_{1tr} \sim \text{Gamma}\{\text{rank}(\tilde{K})/2 - 0.5, \text{vec}(\beta_r)^\top \tilde{K} \text{vec}(\beta_r)/2\}$

   Sample $\tau_{1xr} \sim \text{Gamma}\{\text{rank}(\tilde{K}_r)/2 - 0.5, \text{vec}(\beta_r^\top)\tilde{K}_r \text{vec}(\beta_r)/2\}$ (if $p_r > 1$)

   Sample $\tau_{2trj} \sim \text{Gamma}\{\text{rank}(\tilde{K})/2 - 0.5, \text{vec}(\Lambda_{rj})^\top \tilde{K} \text{vec}(\Lambda_{rj})/2\}$

   Sample $\tau_{2xrj} \sim \text{Gamma}\{\text{rank}(\tilde{K}_r)/2 - 0.5, \text{vec}(\Lambda_{rj}^\top)\tilde{K}_r \text{vec}(\Lambda_{rj})/2\}$ (if $p_r > 1$)

6. Update ϕ_{rj}:

   Let ϕ_{irj} denote the ith diagonal element of ϕ_{rj}.

   Let λ_{irj} be the ith element of $\text{vec}(\Lambda_{rj})$.

   Sample $\phi_{irj} \sim \text{Gamma}(a_\phi + 0.5, \tau_{rj}^* \lambda_{irj}^2/2 + b_\phi)$

7. Update δ_{r1}:

   Let $A = \text{vec}(\Lambda_{r1}^\top)\phi_{r1}\text{vec}(\Lambda_{r1})$

   Let $B = \sum_{j=2}^{k} \tau_{rj}^* \text{vec}(\Lambda_{rj})^\top \phi_{rj} \text{vec}(\Lambda_{rj})$

   Sample $\delta_{r1} \sim \text{Gamma}\{kp_r p/2 + a_{r0}, (A + B)/2 + 1\}$

8. Update δ_{rj}:

   Let $A = \sum_{j'=1}^{k} \tau_{rj'}^{*(j)} \text{vec}(\Lambda_{rj'})^\top \phi_{rj'} \text{vec}(\Lambda_{rj})$, where $\tau_{rj'}^{*(j)} = \tau_{rj'}^*$ if $j \neq j'$ and 1 otherwise.

   Sample $\delta_{rj'} \sim \text{Gamma}\{p_r p(k - j' + 1)/2 + a_{r1}, A/2 + 1\}$

9. Update a_{r0}:

   Let $\text{Gamma}^*(x, a, b)$ denote the Gamma density evaluated at x with shape a and rate b.

   Let $\phi(x)$ denote the standard normal cumulative distribution function evaluated at x.

   Sample candidate $a_{r0}^* \sim N(a_{r0}, 1)$ until $a_{r0}^* > 0$.

   Compute $A = \dfrac{\text{Gamma}^*(\delta_{r1}, a_{r0}^*, 1) \cdot \text{Gamma}^*(a_{r0}^*, 2, 1) \cdot \phi(a_{r0})}{\text{Gamma}^*(\delta_{r1}, a_{r0}, 1) \cdot \text{Gamma}^*(a_{r0}, 2, 1) \cdot \phi(a_{r0}^*)}$

   Sample $U \sim \text{Unif}(0, 1)$

   If $U \leq A$, accept candidate a_{r0}^*.

10. Update a_{r1}:

    Let $\delta^*_{r2} = \prod_{j=2}^{k} \delta_{rj}$

    Let $\text{Gamma}^*(x, a, b)$ denote the Gamma density evaluated at x with shape a and rate b.

    Let $\phi(x)$ denote the standard normal cumulative distribution function evaluated at x.

    Sample candidate $a^*_{r1} \sim N(a_{r1}, 1)$ until $a^*_{r1} > 0$.

    Compute $A = \dfrac{\text{Gamma}^*(\delta^*_{r2}, a^*_{r1}, 1) \cdot \text{Gamma}^*(a^*_{r1}, 2, 1) \cdot \phi(a_{r1})}{\text{Gamma}^*(\delta^*_{r2}, a_{r1}, 1) \cdot \text{Gamma}^*(a_{r1}, 2, 1) \cdot \phi(a^*_{r1})}$

    Sample $U \sim \text{Unif}(0, 1)$

    If $U \leq A$, accept candidate a^*_{r1}.

11. Update missing values of Y:

    Suppose $y_i(t_j)$ is missing.

    Let $\mu = \boldsymbol{b}(t_j)\beta\boldsymbol{b}^x(\boldsymbol{x}_i) + \sum_{j=1}^{k} \boldsymbol{b}(t_j)\Lambda_j\boldsymbol{b}^x(\boldsymbol{x}_i)\eta_{ij}$

    Sample $y_i(t_j) \sim N(\mu, \sigma^2)$

Appendix 3B: Additional Details on Posterior Inference

In this section we give details on post-processing MCMC samples to obtain eigenfunctions as in a usual FPCA. Latent functional factors $\psi_j(t, \boldsymbol{x})$ are not orthonormal and cannot be directly interpreted as function principal components. To orthonormalize $\psi_j(t, \boldsymbol{x})$, one may evaluate posterior draws of $c(t, t', \boldsymbol{x})$ on an arbitrary domain $t, t' \in \mathcal{T}$ followed by a spectral decomposition to yield orthonormal modes of variation. However, this procedure may be computationally intensive because $\mathcal{T}$ is in theory infinite-dimensional. To alleviate this burden, we borrow methods from Aguilera and Aguilera-Morillo [2013] to obtain orthonormal modes of variation by performing a spectral decomposition on a much lower dimensional matrix. To obtain $\tilde{\psi}^{(m)}(t, \boldsymbol{x})$, the mth draw of orthonormal posterior modes of variation, we

1. Compute $p \times p$ matrix Ψ, where $\Psi_{ij} = \int_{\mathcal{T}} b_i(t)b_j(t)dt$.
2. Compute $\tilde{\Lambda}_{\boldsymbol{x}} = \sum_{j=1}^{k} \Lambda_j \boldsymbol{b}^x(\boldsymbol{x})\boldsymbol{b}^x(\boldsymbol{x})^\top \Lambda_j^\top$
3. Set $\tilde{\Lambda}_{\boldsymbol{x}} = \Psi^{1/2}\tilde{\Lambda}_{\boldsymbol{x}}\Psi^{1/2}$

4. Perform a spectral decomposition on $\tilde{\Lambda}_{\boldsymbol{x}}$ to obtain $\tilde{\psi}_j(\boldsymbol{x})$

5. Set $\tilde{\psi}_j(t, \boldsymbol{x}) = \boldsymbol{b}(t)\Psi^{-1/2}\tilde{\psi}_j(\boldsymbol{x})$

6. If desired, postprocessed principal scores are equal to $\int_{\mathcal{T}} \tilde{\psi}_j(t, \boldsymbol{x}_i) r_i(t, \boldsymbol{x}_i)\, dt$. For sparse functional response settings, principal scores are based on conditional expectation [Yao et al., 2005].

Step 1 may be pre-computed before any posterior samples are evaluated. We compare the mth posterior eigenfunction to the running average in l_2 norm. If the norm is smaller after multiplying the estimated eigenfunction by -1, we multiply the estimated eigenfunction by -1. This process ensures all posterior samples of eigenfunctions are oriented correctly so that means and interval calculations are sensible.

Up until this point we have only discussed pointwise credible intervals. However, pointwise intervals are not appropriate for making probabilistic statements of an entire function. Simultaneous credible bands are more appropriate for entire function inference and are easily computable using posterior samples assuming normality as detailed in Krivobokova et al. [2010] and Crainiceanu et al. [2007]. Suppose we desire a simultaneous credible band of some functional $f(\cdot)$ observed at points $t_1, \ldots, t_N$. Let $\mu_f(t_i)$ and $\sigma_f(t_i)$ be the estimated pointwise mean and standard deviation respectively. Let $f^{(m)}(\cdot)$ be the mth realization of $f(\cdot)$ drawn from the posterior out of M total samples. By assuming approximate normality and deriving the $(1-\alpha)$ sample quantity c_b of

$$\max_{i=1,\ldots,N} \left| \frac{f^{(m)}(t_i) - \mu_f(t_i)}{\sigma_f(t_i)} \right|, \; m = 1, \ldots, M$$

a simultaneous credible region is given by the hyperrectangular

$$\big[\mu_f(t_i) - c_b \cdot \sigma_f(t_i), \mu_f(t_i) + c_b \cdot \sigma_f(t_i)\big], \; i = 1 \ldots, N$$

In our implenetation on github we include the option to compute pointwise and simultaneous bands for latent subject-specific functions, covariate-adjusted means, and covariate-adjusted

eigenfunctions.

Appendix 3C: Additional Details on the Case Studies

In this section we provide details setting up the design matrices and penalty matrices for the ASD case study [Dickinson et al., 2018] and sleep wave data from the Sleep Heart Health Study (SHHS) [Quan et al., 1997]. We also discuss computing low dimensional covariance summaries $g(t, \boldsymbol{x})$ and associated inference.

The ASD case study [Dickinson et al., 2018] has three covariates excluding an intercept term: diagnostic group, age of child, and a group by age interaction term. In terms of notation used in Web Appendix A, the design matrix X has dimension 97 by 12. The first column of X is an intercept column with repeating ones. The second column is a group indicator with a 1 if the child is diagnosed with ASD and 0 otherwise. The next five columns expand age of child by p-splines. The next five columns expand age of child of child by p-splines row-wise multiplied by 0 if the child is in the TD diagnostic group. In terms of notation from Sections 2 and 3 of the main manuscript, $\boldsymbol{x}_1 = \{\text{Intercept}\}$, $\boldsymbol{x}_2 = \{\text{Group}\}$, $\boldsymbol{x}_3 = \{\text{Age}\}$, and $\boldsymbol{x}_4 = \{\text{Group by age interaction}\}$. We also expand the frequency dimension into a set of 12 p-splines. The frequency dimension is associated with a 12×12 second order smoothing matrix K with rank 10. The age dimension is associated with a 5×5 second order smoothing matrix with rank 3. Let M^- denote the generalized inverse of a matrix M. The 12×1 coefficient matrices β_1 and β_2 have priors $N\{0, (\tau_{1t1}K)^-\}$ and $N\{0, (\tau_{1t2}K^-)\}$. The 12×5 coefficient matrices β_3 and β_4 have priors $N\{0, (\tau_{1t3}\tilde{K} + \tau_{1x3}\tilde{K}_{\text{age}})^-\}$ and $N\{(0, (\tau_{1x4}\tilde{K}_{\text{age}} + \tau_{1t4}\tilde{K})^-\}$ after vectorization. Here $\tilde{K}_{\text{age}} = K_{\text{age}} \otimes I_{12\times 12}$ and $\tilde{K} = I_{5\times 5} \otimes K$. Priors for Λ_{rj} are identical after replacing τ_{1tr} and τ_{1xr} by τ_{2trj} and τ_{2xrj}.

The low dimensional summaries $g(t, \boldsymbol{x}_r)$ are designed to help users quantify heterogeneity by covariates. Since the functional argument is frequency, abbreviated as ω, we will write $g(\omega, \boldsymbol{x}_r)$ in place of $g(t, \boldsymbol{x}_r)$. In this example we compute four $g(\omega, \boldsymbol{x}_r)$ to quantify heterogeneity. The first quantifies heterogeneity for a 70 month old child with ASD. A single posterior sample is computed by $g(\omega, \boldsymbol{x}_r) = \sum_{j=1}^{8}[\boldsymbol{b}(\omega)\Lambda_k \boldsymbol{b}^x(\boldsymbol{x}^1)]^2$, where $\boldsymbol{x}^1$ is the covariate vector for

a 70 month old child in the ASD group. The second quantifies heterogeneity for a 70 month old TD child. A single posterior sample is computed by $g(\omega, \boldsymbol{x}_r) = \sum_{j=1}^{8}[\boldsymbol{b}(\omega)\Lambda_k \boldsymbol{b}^x(\boldsymbol{x}^2)]^2$, where $\boldsymbol{x}^2$ is the covariate vector for a 70 month old child in the TD group. These two quantities represent the total amount of heterogeneity for a typical child in the ASD and TD groups respectively. The third quantifies the additional amount of heterogeneity over age in the ASD group. A single posterior sample is computed as $\frac{1}{10}\sum_{m=1}^{10}\sum_{j=1}^{8}[\boldsymbol{b}(\omega)\Lambda_k \boldsymbol{b}^x(\boldsymbol{x}^{1m} - \boldsymbol{x}^1)]^2$, where $\boldsymbol{x}^{1m}$ denotes the covariate vector for a child in the ASD group at mth equally spaced age between the 10th and 90th percentile of all ages. The fourth quantifies the additional amount of heterogeneity over age in the TD group. A single posterior sample is computed as $\frac{1}{10}\sum_{m=1}^{10}\sum_{j=1}^{8}[\boldsymbol{b}(\omega)\Lambda_k \boldsymbol{b}^x(\boldsymbol{x}^{2m} - \boldsymbol{x}^2)]^2$, where $\boldsymbol{x}^{2m}$ denotes the covariate vector for a child in the TD group at the mth equally spaced age between the 10th and 90th percentile of all ages. The latter two summaries are designed to integrate additional heterogeneity over age compared to the group respective baseline at 70 months old. Web Figure A3.1 displays pointwise posterior medians of the four summaries described above. Baseline heterogeneity is similar between ASD and TD groups. The TD baseline heterogeneity curve shows that the heterogeneity is concentrated between childrens' alpha spectral densities at 6 Hz and at 8 - 11 Hz. The ASD baseline curve shows shows that variability is conentrated at 6 Hz and peaked around 9.5 Hz. There does not seem to be a relationship between heterogeneity and age. The SHHS case study has three covariates of interest excluding an intercept term: hypertension group, age, and a group by age interaction. The design matrix X has dimension 5258 by 16, which accounts for 5258 patients with an intercept, group indicator, and seven p-spline basis functions for age and age$\times$ group interaction. Following a similar setup as the previous case study, $\boldsymbol{x}_1 = \{\text{Intercept}\}$, $\boldsymbol{x}_2 = \{\text{Group}\}$, $\boldsymbol{x}_3 = \{\text{Age}\}$, and $\boldsymbol{x}_4 = \{\text{Group by age interaction}\}$. We expand the sleep time dimension into a set of 24 p-splines associated with a 24×24 second order smoothing matrix K with rank 22. The age dimension is associated with a second order smoothing matrix K_{age} with rank 5. The 24×1 coefficient matrices β_1 and β_2 have priors $N(\{0, (\tau_{1t1}K^-)\}$ and $N\{0, (\tau_{1t2}K)^-)\}$. The 24×7 coefficient matrices β_3 and β_4 have have priors $N\{0, (\tau_{1t3}\tilde{K} + \tau_{1x3}\tilde{K}_{\text{age}})^-\}$ and $N\{0, (\tau_{1t4}\tilde{K} + \tau_{1x4}\tilde{K}_{\text{age}})^-\}$ after vectorization. Here $\tilde{K}_{\text{age}} = K_{\text{age}} \otimes I_{24\times 24}$ and $\tilde{K} = I_{7\times 7} \otimes K$. Priors for Λ_{rj} are identical after replacing τ_{1tr}

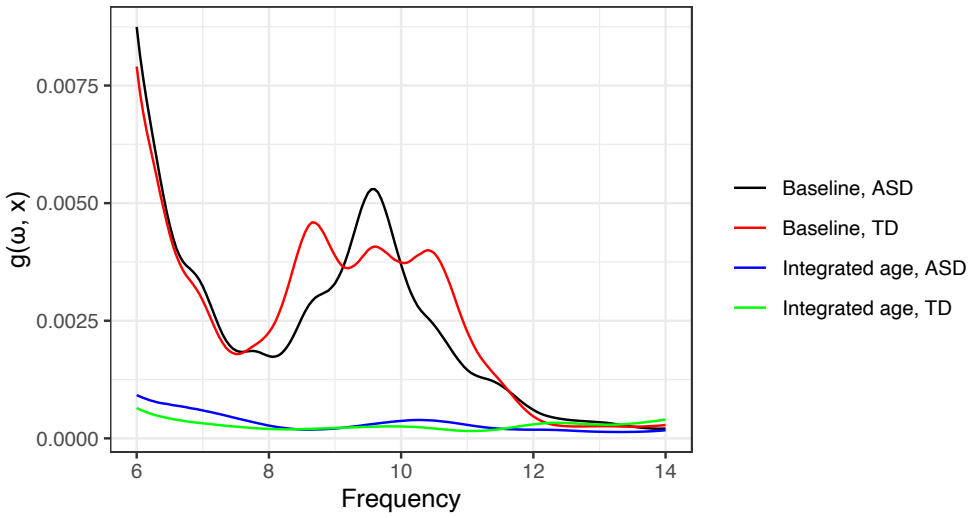


Figure A3.1: Posterior medians of low dimensional summaries $g(\omega, \boldsymbol{x}_r)$. The age effect is averaged over 10 equally spaced ages. Heterogeneity is largest around 6 Hz. TD heterogeneity is large around 8 - 11 Hz and ASD heterogeneity is more peaked around 9.5 Hz.

and τ_{1xr} by τ_{2trj} and τ_{2xrj}. Low dimensional covariance summaries $g(t, \boldsymbol{x}_r)$ are calculated in the same manner as in the ASD case study, except we sum over 12 latent factors as opposed to 8. The baseline age for both groups is 63 years old. Web Figure A3.2 displays the posterior mean of the low dimensional summaries. Heterogeneity between subjects' relative delta power spectral density is highest at about 1 hour of sleep (epoch 100 - 120). The figure also shows that heterogeneity does not depend on age, similar to the ASD case study.

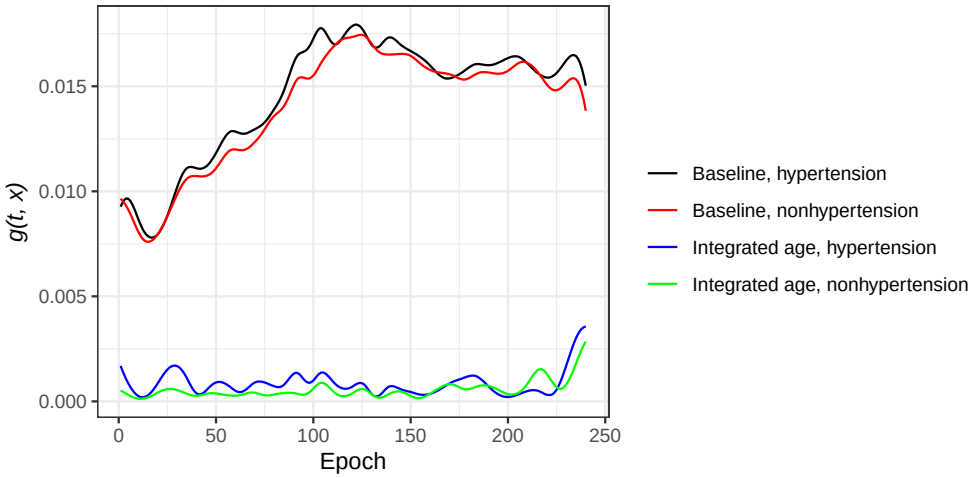


Figure A3.2: Posterior medians of low dimensional summaries $g(\omega, \boldsymbol{x}_r)$. The age effect is averaged over 10 equally spaced ages. Heterogeneity curves have a similar profile over both hypertension and nonhypertension groups. Heterogeneity is is largest around the first hour of sleep (epoch 100 - 120) for both groups. Heterogeneity is not influenced by age in either group.

Appendix 4A: Orthonormal Constraints and Likelihood Simplifications

Orthonormality constraints are imposed on both data-driven basis functions $\boldsymbol{\psi}_l = (\psi_l(t_1), \ldots, \psi_l(t_n))'$ and regional bases $\boldsymbol{\phi}_{lj}$ during sampling. For every posterior sample, these constraints force $\boldsymbol{\psi}_l'\boldsymbol{\psi}_{l'} = \delta_{ll'}$ and $\boldsymbol{\phi}_{lj}'\boldsymbol{\phi}_{lj'} = \delta_{jj'}$ where δ_{ij} is the Kronecker delta. We follow Kowal et al. [2017], Kowal and Bourgeois [2020], Kowal [2021] and use a two-step procedure to impose orthonormality conditions: at every posterior sample of $\boldsymbol{\psi}_l$ (or $\boldsymbol{\phi}_{lj}$), we (1) *condition* on $C_{l,\psi}\boldsymbol{\psi}_l = 0$ (or $C_{l,\phi_l}\boldsymbol{\phi}_{lj} = 0$) and (2) rescale to enforce unit l_2 norm. The matrices $C_{l,\psi}$ and C_{j,ϕ_l} are equal to $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-1})'$ and $(\boldsymbol{\phi}_{l1}, \ldots, \boldsymbol{\phi}_{l,j-1})'$ respectively. Note that these orthogonal constraints only apply when $l > 1$ (or $j > 1$), so that $\boldsymbol{\psi}_1$ (or $\boldsymbol{\phi}_{l1}$) has no constraints. Conditioning on constraints is a natural idea in Bayesian statistics, and produces optimal properties for constrained penalized regression.

Let $\boldsymbol{b}_p = (b_p(t_1), \ldots, b_p(t_n))'$ and $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_P)$. To condition on $C_{l,\psi}\boldsymbol{\psi}_l = 0$ during the posterior sample of $\boldsymbol{\psi}_l$, we first sample $\boldsymbol{\lambda}_l$ without constraints as $\boldsymbol{\lambda}_l \sim N(Q_{\lambda,l}^{-1}\boldsymbol{b}_{\lambda,l}, Q_l^{-1})$ for some matrix $Q_{\lambda,l}^{-1}$ and vector $\boldsymbol{b}_{\lambda,l}$. Since $\boldsymbol{\lambda}_l$ and $C_{l,\psi}\boldsymbol{\psi}_l = C_{l,\psi}B\boldsymbol{\lambda}_l$ have joint distribution

$$\begin{pmatrix} \boldsymbol{\lambda}_l \\ C_{l,\psi}B\boldsymbol{\lambda}_l \end{pmatrix} \sim N\left(\begin{pmatrix} Q_{\lambda,l}^{-1}\boldsymbol{b}_{\lambda,l} \\ C_{l,\psi}BQ_{\lambda,l}^{-1}\boldsymbol{b}_{\lambda,l} \end{pmatrix}, \begin{pmatrix} Q_{\lambda,l}^{-1} & Q_{\lambda,l}^{-1}B'C_{\psi,l}' \\ C_{\psi,l}BQ_{\lambda,l}^{-1} & C_{\psi,l}BQ_{\lambda,l}^{-1}B'C_{\psi,l}' \end{pmatrix}\right)$$

so $\boldsymbol{\lambda}_l \,|\, C_{\psi,l}B\boldsymbol{\lambda}_l = 0$ is again multivariate normal with mean

$$Q_{\lambda,l}^{-1}\boldsymbol{b}_{\lambda,l} - Q_{\lambda,l}^{-1}B'C_{\psi,l}'(C_{\psi,l}BQ_{\lambda,l}^{-1}B'C_{\psi,l}')^{-1}C_{\psi,l}BQ_{\lambda,l}^{-1}\boldsymbol{b}_{\lambda,l}$$

and variance-covariance matrix

$$Q_{\lambda,l}^{-1} - Q_{\lambda,l}^{-1}B'C_{\psi,l}'(C_{\psi,l}BQ_{\lambda,l}^{-1}B'C_{\psi,l}')^{-1}C_{\psi,l}BQ_{\lambda,l}^{-1}$$

Now consider the random variable $\boldsymbol{\lambda}_l^* = \boldsymbol{\lambda}_l - Q_{\psi,l}^{-1}B'C_{\psi,l}'(C_{\psi,l}BQ_{\lambda,l}^{-1}B'C_{\psi,l}')^{-1}C_{\psi,l}B\boldsymbol{\lambda}_l$. One

can work out the algebra to show $\boldsymbol{\lambda}_l^*$ has the same mean and variance-covariance above. Since the the multivariate normal distribution is idenified by its mean and covariance, $\boldsymbol{\lambda}_l^*$ is equal in distribution to $\boldsymbol{\lambda}_l \,|\, C_{\psi,l} B \boldsymbol{\lambda}_l = 0$. Moreover, sampling $\boldsymbol{\lambda}_l^*$ requires fewer floating point operations than sampling $\boldsymbol{\lambda}_l \,|\, C_{\psi,l} B \boldsymbol{\lambda}_l = 0$, so sampling $\boldsymbol{\lambda}_l^*$ is the preferred method. By construction, $\boldsymbol{\psi}_l = B\boldsymbol{\lambda}_l^*$, is orthogonal to $\boldsymbol{\psi}_{l'}$ for $l' > l$. After constructing $\boldsymbol{\psi}_l$, we overwrite $\boldsymbol{\psi}_l$ with its normalized version, $\boldsymbol{\psi}_l / ||\boldsymbol{\psi}_l||$, where $||\cdot||$ denotes the Euclidean norm. As $\boldsymbol{\psi}_l$ are sampled in succession, $\boldsymbol{\psi}_l' \boldsymbol{\psi}_{l'} = \delta_{ll'}$, as desired. Ensuring orthonormality of $\boldsymbol{\phi}_{lj}$ follows a similar strategy.

Aside from enhancing interpretability of $\psi_l(t)$ and $\boldsymbol{\phi}_{lj}$, orthonormalization leads to likelihood simplifications, greatly improving computational scalability. Let $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))$. Consider the likelihood $p(Y_i \,|\, \boldsymbol{\psi}_l, \phi_l, \eta_i, \Sigma_\epsilon)$. As a function of η_i, we have that

$$p(Y_i \,|\, \boldsymbol{\psi}_l, \phi_l, \eta_i, \Sigma_\epsilon) \tag{5.1}$$

$$\propto \exp\left\{ -.5 \left[\mathrm{vec}(Y_i) - \sum_{l=1}^{L} (\boldsymbol{\psi}_l \otimes \phi_l) \boldsymbol{\eta}_{il} \right]' I_{t_n} \otimes \Sigma_\epsilon^{-1} \left[\mathrm{vec}(Y_i) - \sum_{l=1}^{L} (\boldsymbol{\psi}_l \otimes \phi_l) \boldsymbol{\eta}_{il} \right] \right\} \tag{5.2}$$

$$\propto \exp\left\{ \mathrm{vec}(\Sigma_\epsilon^{-1} Y_i)' \sum_{l=1}^{L} (\boldsymbol{\psi}_l \otimes \phi_l) \boldsymbol{\eta}_{il} \right\} \exp\left\{ -.5 \left[\sum_{l=1}^{L} \boldsymbol{\eta}_{il}' \phi_l' \Sigma_\epsilon^{-1} \phi_l \boldsymbol{\eta}_{il} \right] \right\} \tag{5.3}$$

Equation 5.3 follows from Equation 5.2 because $\boldsymbol{\psi}_l' \boldsymbol{\psi}_{l'} = \delta_{ll'}$. Note that removing this constraint would make a double sum involving $\boldsymbol{\psi}_l$ appear in the last term of the likelihood in Equation 5.3. Clearly the orthonormality constraint simplifies the likelihood $p(Y_i \,|\, \boldsymbol{\psi}_l, \phi_l, \eta_i, \Sigma_\epsilon)$, thus enabling more efficient sampling of η_i.

Appendix 4B: Markov-Chain Monte Carlo Sampling Algorithm

In this section we give a detailed Markov-Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution. The github repository github.com/jshamsho/rrbfda includes a C++ implementation of all MCMC algorithms. The algorithms described here use param-

eter expansion to sample from conjugate conditional posteriors. Specifically, we replace $\eta_{ijl} \sim t_\nu(\boldsymbol{x}_i'\boldsymbol{\beta}_{lj}, \sigma_{lj}^2)$ with $\eta_{ilj} \sim N(\boldsymbol{x}_i\boldsymbol{\beta}_{lj}, \sigma_{lj}^2\xi_{\eta_{ilj}})$ and $\xi_{\eta_{ilj}}^{-1} \sim \text{Gamma}(\nu/2, \nu/2)$. Let β_{ljd} be the dth element of $\boldsymbol{\beta}_{lj}$. We also replace $\beta_{ljd} \sim t_4(0,1)$ priors with $\beta_{ljd} \sim N(0, \sigma_{\beta_{ljd}}^2)$ and $\sigma_{\beta_{ljd}}^{-2} \sim \text{Gamma}(4/2, 4/2)$. We will describe our sampling algorithms for parameters common to both PSFLM, specific to PSFLM, and specific to WSFLM in this order.

MCMC algorithms for parameters common to PSFLM and WSFLM

1. Update ζ_l

   Sample $\zeta_l \sim \text{Gamma}(a, b)1(\zeta_l < 10^7)$

   $a = -.5 + .5\text{rank}(\Omega)$

   $b = .5\boldsymbol{\lambda}_l'\Omega\boldsymbol{\lambda}_l$

2. Update ν

   Standard Metropolis-Hastings update with a proposal distribution

   $\text{Uniform}(\max(2, \nu - 2), \min(128, \nu + 2))$

3. Update $\boldsymbol{\beta}_{lj}$

   Let $\boldsymbol{\eta}_{lj} = (\eta_{1lj}, \ldots, \eta_{nlj})'$, $\Sigma_{\eta_{lj}} = \text{diag}(\sigma_{lj}^2\xi_{\eta_{ilj}}, \ldots, \sigma_{lj}^2\xi_{\eta_{ilj}})$, $\Sigma_{\beta_{lj}} = \text{diag}(\sigma_{\beta_{lj1}}^2, \ldots, \sigma_{\beta_{ljD}}^2)$

   Sample $\boldsymbol{\beta}_{lj} \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = X'\Sigma_{\eta_{lj}}^{-1}X + \Sigma_{\beta_{lj}}^{-1}$

   $\boldsymbol{b} = X'\Sigma_{\eta_{lj}}^{-1}\boldsymbol{\eta}_{lj}$

4. Update $\sigma_{\beta_{ljd}}^2$

   Sample $\sigma_{\beta_{ljd}}^{-2} \sim \text{Gamma}(a, b)$

   $a = .5(4 + 1)$

   $b = .5(4 + \beta_{ljd}^2)$

5. Update $\xi_{\eta_{ilj}}$

   Sample $\xi_{\eta_{ilj}}^{-1} \sim \text{Gamma}(a, b)$

   $a = .5(\nu + 1)$

   $b = .5(\nu + (\eta_{ilj} - \boldsymbol{x}_i'\boldsymbol{\beta}_{lj})^2\sigma_{lj}^{-2})$

MCMC algorithms for PSFLM

1. Update $\boldsymbol{\psi}_l$

   Let $B = (\boldsymbol{b}(t_1), \ldots, \boldsymbol{b}(t_n))'$, $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$

   Sample $\boldsymbol{\lambda}_l \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = \text{tr}(\phi_l \Sigma_\epsilon^{-1} \phi_l \sum_{i=1}^n \boldsymbol{\eta}_{il}\boldsymbol{\eta}_{il}')B'B + \zeta_l \Omega$

   $\boldsymbol{b} = \sum_{i=1}^n B'Y_i\Sigma_\epsilon^{-1}\phi_l\boldsymbol{\eta}_{il} - \sum_{i=1}^n \sum_{l'=1,l'\neq l}^L B'B\boldsymbol{\lambda}_{l'}\boldsymbol{\eta}_{il'}'\phi_{l'}'\Sigma_\epsilon^{-1}\phi_l\boldsymbol{\eta}_{il}$

   Overwrite $\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l - Q^{-1}B'C_{\psi,l}'(C_{\psi,l}BQ^{-1}B'C_{\psi,l}')^{-1}C_{\psi,l}B\boldsymbol{\lambda}_l$

   Where $C_{\psi,l} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{l-1})'$

   Set $\boldsymbol{\psi}_l = B\boldsymbol{\lambda}_l$

   Overwrite $\boldsymbol{\psi}_l = \boldsymbol{\psi}_l / ||\boldsymbol{\psi}_l||$

2. Update $\boldsymbol{\phi}_{lj}$

   Let $\boldsymbol{\phi}_j = (\boldsymbol{\phi}_{1j}', \ldots, \boldsymbol{\phi}_{Lj}')'$, $\eta_{ij} = \text{diag}(\eta_{i1j}, \ldots, \eta_{iLj})$, $\psi = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L)$

   Let $M_{\phi_j} = (\boldsymbol{\phi}_{1j}, \ldots, \boldsymbol{\phi}_{Lj})$

   Sample $\boldsymbol{\phi}_j \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = (\sum_{i=1}^n \eta_{ij}\eta_{ij}) \otimes \Sigma_\epsilon^{-1} + I_{RL}$

   $\boldsymbol{b} = \sum_{i=1}^n \text{vec}(\Sigma_\epsilon^{-1}Y_i'\psi\eta_{ij}) - \sum_{i=1}^n \sum_{j'=1,j'\neq j}^R \text{vec}(\Sigma_\epsilon^{-1}M_{\phi_j}\eta_{ij'}\eta_{ij})$

   Overwrite $\boldsymbol{\phi}_j = \boldsymbol{\phi}_j - Q^{-1}C_{\phi,j}'(C_{\phi,j}Q^{-1}C_{\phi,j}')^{-1}C_{\phi,j}Q^{-1}\boldsymbol{\phi}_j$

   Where $C_{\phi,j} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{j-1})'$

   Overwrite $\boldsymbol{\phi}_{lj} = \boldsymbol{\phi}_{lj} / ||\boldsymbol{\phi}_{lj}||$

3. Update η_{ilj}

   Let $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$, $\Sigma_{\eta_{il}} = \text{diag}(\sigma_{l1}^2\xi_{\eta_{il1}}, \ldots, \sigma_{lR}^2\xi_{\eta_{ilR}})$, $\beta_l = (\boldsymbol{\beta}_{l1}, \ldots, \boldsymbol{\beta}_{lR})'$

   Sample $\boldsymbol{\eta}_{il} \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = \phi_l'\Sigma_\epsilon^{-1}\phi_l + \Sigma_{\eta_{il}}^{-1}$

   $\boldsymbol{b} = \phi_l'\Sigma_\epsilon^{-1}Y_i'\boldsymbol{\psi}_l + \Sigma_{\eta_{il}}^{-1}\beta_l\boldsymbol{x}_i$

4. Update Σ_ϵ

   Let $\sigma_{\epsilon j}^2$ be the jth diagonal element of Σ_ϵ

   Let $\phi_{ljj'}$ be the j'th element of $\boldsymbol{\phi}_{lj}$

Sample $\sigma_{\epsilon j}^{-2} \sim \text{Gamma}(a, b)$

Where $a = .5nt_n$

$b = .5 \sum_{i=1}^{n} \sum_{t=t_1}^{t_n} (Y_{ij}(t) - \sum_{l=1}^{L} \sum_{j'=1}^{R} \psi_l(t)\phi_{ljj'}\eta_{ilj'})^2$

5. Update δ_1

   Let $\boldsymbol{\eta}_{lj} = (\eta_{1lj}, \ldots, \eta_{nlj})'$, $D_{\xi_{\eta_{lj}}} = \text{diag}(\xi_{\eta_{1lj}}, \ldots, \xi_{\eta_{nlj}})$

   Set $\delta_1 = 1$

   Compute $\sigma_{l1}^{-2} = \prod_{l'=1}^{l} \delta_{l'}$

   Compute $\sigma_{lj}^{-2} = \prod_{l'=1}^{l} \prod_{j'=2}^{j} \delta_{l'}\delta_{l'j'}$, $j > 1$

   Sample $\delta_1 \sim \text{Gamma}(a, b)$

   $a = a_1 + .5nRL$

   $b = .5 + \sum_{l=1}^{L} \sum_{r=1}^{R} (\boldsymbol{\eta}_{lj} - X\boldsymbol{\beta}_{lj})' D_{\xi_{\eta_{lj}}}^{-1} (\boldsymbol{\eta}_{lj} - X\boldsymbol{\beta}_{lj}) \sigma_{lj}^{-2}$

6. Update δ_l, $l > 1$

   Let $\boldsymbol{\eta}_{l'j} = (\eta_{1l'j}, \ldots, \eta_{nl'j})'$, $D_{\xi_{\eta_{l'j}}} = \text{diag}(\xi_{\eta_{1l'j}}, \ldots, \xi_{\eta_{nl'j}})$

   Set $\delta_l = 1$

   Compute $\sigma_{l'1}^{-2} = \prod_{l''=1}^{l'} \delta_{l''}$

   Compute $\sigma_{l'j}^{-2} = \prod_{l''=1}^{l'} \prod_{j'=2}^{j} \delta_{l''}\delta_{l''j'}$

   Sample $\delta_l \sim \text{Gamma}(a, b)$

   $a = a_2 + .5n(L - l' + 1)R$

   $b = .5 + \sum_{l'=l}^{L} (\boldsymbol{\eta}_{l'j} - X\boldsymbol{\beta}_{l'j})' D_{\xi_{\eta_{l'j}}}^{-1} (\boldsymbol{\eta}_{\boldsymbol{l'j}} - X\boldsymbol{\beta}_{l'j}) \sigma_{l'j}^{-2}$

7. Update δ_{lj}, $j > 1$

   Let $\boldsymbol{\eta}_{lj'} = (\eta_{1lj'}, \ldots, \eta_{nlj'})'$, $D_{\xi_{\eta_{lj'}}} = \text{diag}(\xi_{\eta_{1lj'}}, \ldots, \xi_{\eta_{nlj'}})$

   Set $\delta_{lj} = 1$ Compute $\sigma_{l1}^{-2} = \prod_{l'=1}^{l} \delta_{l'}$

   $\sigma_{lj'}^{-2} = \prod_{l'=1}^{l} \prod_{j''=2}^{j'} \delta_{l'}\delta_{l'j''}$

   Sample $\delta_{lj} \sim \text{Gamma}(a, b)$

   $a = a_3 + .5n(R - j + 1)$

   $b = .5 + \sum_{j'=j}^{R} \sigma_{lj'}^{-2} (\boldsymbol{\eta}_{lj'} - X\boldsymbol{\beta}_{lj'})' D_{\xi_{\eta_{lj'}}}^{-1} (\boldsymbol{\eta}_{lj'} - X\boldsymbol{\beta}_{lj'})$

8. Update a_1, a_2, a_3

   Standard Metropolis-Hastings updates with proposal distributions

$a_v \sim \text{Uniform}(\max(a_v - 0.5, 0), a_v + 0.5),\ v = 1, 2, 3$

MCMC algorithms for WSFLM

1. Update $\boldsymbol{\psi}_l$

   Let $\phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_R)$, $B = (\boldsymbol{b}(t_1), \ldots, \boldsymbol{b}(t_n))'$, $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$

   Sample $\boldsymbol{\lambda}_l \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = \text{tr}(\phi'\Sigma_\epsilon^{-1}\phi \sum_{i=1}^n \boldsymbol{\eta}_{il}\boldsymbol{\eta}_{il}')B'B + \zeta_l\Omega$

   $\boldsymbol{b} = \sum_{i=1}^n B'Y_i\Sigma_\epsilon^{-1}\phi\boldsymbol{\eta}_{il} - \sum_{l'=1,l'\neq l}^L B'B\boldsymbol{\lambda}_{l'}\boldsymbol{\eta}_{il'}'\phi'\Sigma_\epsilon^{-1}\phi\boldsymbol{\eta}_{il}$

   Overwrite $\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l - Q^{-1}B'C_{\psi,l}'(C_{\psi,l}BQ^{-1}B'C_{\psi,l}')^{-1}C_{\psi,l}B\boldsymbol{\lambda}_l$

   Where $C_{\psi,l} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{l-1})'$

   Set $\boldsymbol{\psi}_l = B\boldsymbol{\lambda}_l$

   Overwrite $\boldsymbol{\psi}_l = \boldsymbol{\psi}_l / ||\boldsymbol{\psi}_l||$

2. Update $\boldsymbol{\phi}_j$

   Let $\boldsymbol{\eta}_{ij}^* = (\eta_{i1j}, \ldots, \eta_{iLj})'$, $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$, $\psi = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L)$

   Sample $\boldsymbol{\phi} \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = (\sum_{i=1}^n \boldsymbol{\eta}_{ij}^{*'}\boldsymbol{\eta}_{ij}^*)\Sigma_\epsilon^{-1} + I_R$

   $\boldsymbol{b} = \sum_{i=1}^n \text{vec}(\Sigma_\epsilon^{-1}Y_i'\psi\boldsymbol{\eta}_{ij}^* - \sum_{j'=1,j'\neq j}^R \boldsymbol{\phi}_{j'}\boldsymbol{\eta}_{ij'}^{*'}\boldsymbol{\eta}_{ij}^*)$

   Overwrite $\boldsymbol{\phi}_j = \boldsymbol{\phi}_j - Q^{-1}C_{\phi,j}'(C_{\phi,j}Q^{-1}C_{\phi,j}')^{-1}C_{\phi,j}Q^{-1}\boldsymbol{\phi}_j$

   Where $C_{\phi,j} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{j-1})'$

   Overwrite $\boldsymbol{\phi}_j = \boldsymbol{\phi}_j / ||\boldsymbol{\phi}_j||$

3. Update η_{ilj}

   Let $\boldsymbol{\eta}_i = \text{vec}(\eta_{i1}, \ldots, \eta_{iL})$, $\phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_R)$, $\beta = (\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}, \ldots, \boldsymbol{\beta}_{LR})$

   $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$, $\psi = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L)$, $\Sigma_{\eta_i} = \text{diag}(\sigma_{11}^2\xi_{\eta_{i11}}, \sigma_{12}^2\xi_{\eta_{i12}}, \ldots, \sigma_{LR}^2\xi_{\eta_{iLR}})$

   Sample $\boldsymbol{\eta}_i \sim N(Q^{-1}\boldsymbol{b}, Q^{-1})$

   $Q = I_L \otimes \phi'\Sigma_\epsilon^{-1}\phi + \Sigma_{\eta_i}^{-1}$

   $\boldsymbol{b} = \text{vec}(\phi'\Sigma_\epsilon^{-1}Y_i'\psi) + \Sigma_{\eta_i}^{-1}\beta\boldsymbol{x}_i$

4. Update Σ_ϵ^{-1}

   Let $\sigma_{\epsilon j}^2$ be the jth diagonal element of Σ_ϵ

Let $\phi_{jj'}$ be the j'th element of $\boldsymbol{\phi}_j$

Sample $\sigma_{\epsilon j}^{-2} \sim \text{Gamma}(a, b)$

Where $a = .5nt_n$

$b = .5 \sum_{i=1}^{n} \sum_{t=t_1}^{t_n} (Y_{ij}(t) - \sum_{l=1}^{L} \sum_{j'=1}^{R} \psi_l(t)\phi_{jj'}\eta_{ilj'})^2$

5. Update a_{vj}, b_{vl}, d_v

   Standard Metropolis-Hastings updates with proposal distributions

   $a_{v1} \sim N(a_{v1}, .025a_{v1})1(a_{v1} > 10^{-5})$

   $a_{vj} \sim N(a_{vj}, .025a_{vj})1(a_{vj} > a_{v,j-1})$, $j > 1$

   $b_{v1} \sim N(b_{v1}, .025b_{v1})1(b_{v1} > 10^{-5})$

   $b_{vl} \sim N(b_{vl}, .025b_{vl})1(b_{vl} > b_{v,l-1})$, $l > 1$

   $d_v \sim N(d_v, .05d_v)1(d_v > d_{v-1})$, $v > 1$

Appendix 4C: Markov-Chain Monte Carlo Initialization Scheme

Parameter initialization greatly reduces the need for a long burn-in at the begin of the sampling process. We initialize the parameters for $\boldsymbol{\lambda}_j$, ϕ_l, Σ_ϵ, and η_{ilj}. The initialization scheme for PSFLM is very similar to that of WSFLM. The algorithms require either a % variability explained criteria (e.g., 95%) or a pre-specified number of latent eigenfunctions L. Starting with PSFLM, the important steps are

1. Create the data matrix $Y_i = (\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_n))'$. Stack these matrices along the column dimension to generate $Y = (Y_1, \ldots, Y_n)$.

2. Use a one-dimensional functional principal components analysis (FPCA), e.g., Xiao et al. [2013], to extract a pre-specified number of eigenfunctions or explain a % variability threshold. Either way, retain L eigenfunctions $\boldsymbol{\psi}_l$. In our implementation, we use $.2t_n$ knots rounded down to the nearest integer. We use the function `fpca.face` in the R package Goldsmith et al. [2020].

3. Regress the eigenfunctions one at a time on a pre-specified basis matrix B to obtain $\boldsymbol{\lambda}_l$, $l = 1, \ldots, L$. Simply set $\boldsymbol{\lambda}_l = (B'B)^{-1}B'\boldsymbol{\psi}_l$.

4. Extract all $\boldsymbol{\theta}_{il}$ coefficients from the one-dimensional FPCA. Arrange the $\boldsymbol{\theta}_{il}$ coefficients in a matrix $\theta_l = (\boldsymbol{\theta}_{1l}, \ldots, \boldsymbol{\theta}_{nl})'$. Perform principal components analysis (PCA) on θ_l for each l. Retain all loadings to form the matrix ϕ_l. Retain all individual scores to form η_{il}.

5. Reconstruct the smoothed data by $\boldsymbol{Y}_i(t) = \sum_{l=1}^{L} \boldsymbol{\psi}\boldsymbol{\eta}'_{il}\phi'_l$. Compute all residuals between smoothed data and observed data, one region at a time. Set $\sigma^2_{\epsilon j}$ to the sample variance of the residuals for the jth region.

Steps 1, 2, and 3 for initializing WSFLM are exactly the same as PSFLM. The rest of the steps for initializing parameters for WSFLM are

4. Stack Y_i along the row dimension to generate $Y = (Y_1, \ldots, Y_n)'$. Perform a PCA on this matrix, extracting regional eigenvectors $\boldsymbol{\phi}_j$, $j = 1, \ldots, R$.

5. Construct a matrix of tensor terms, $E = (\boldsymbol{\psi}_1 \otimes \boldsymbol{\phi}_1, \boldsymbol{\psi}_1 \otimes \boldsymbol{\phi}_2, \ldots, \boldsymbol{\psi}_L \otimes \boldsymbol{\phi}_R)$. Vectorize Y_i column by column to obtain a vector $\tilde{\boldsymbol{Y}}_i$.

6. Compute $\text{vec}(\eta_i) = (E'E)^{-1}E'\tilde{\boldsymbol{Y}}_i$. Rearrange the $\text{vec}(\eta_i)$ to obtain η_{ilj}.

7. Reconstruct the smoothed data $Y_i = \sum_{l=1}^{L}\sum_{j=1}^{R} \boldsymbol{\psi}_l\boldsymbol{\phi}_j\eta_{ilj}$.
   Compute all residuals smoothed data and observed data, one region at a time. Set $\sigma^2_{\epsilon j}$ to the sample variance of residuals for the jth region.

In summary, these algorithms efficiently initialize $\boldsymbol{\lambda}_l$, $\boldsymbol{\psi}_l$, ϕ_l, $\boldsymbol{\theta}_{il}$, and η_{ilj}. This initialization scheme is useful for (1) reducing the need for a long burn-in during posterior sampling and (2) a heuristic for selecting the number of latent $\boldsymbol{\psi}_l$. In particular, we use these algorithms to select L explaining 95% of variability in the EEG applied case study. One should note that the priors on η_{ilj} are designed to prevent overfitting with superfluous $\boldsymbol{\psi}_l$. Therefore, the only penalty of large L is wasted computation, not overfitting.

Bibliography

A. Aguilera and M. Aguilera-Morillo. Penalized pca approaches for b-spline expansions of smooth functional data. *Applied Mathematics and Computation*, 219(14):7805 – 7819, 2013. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2013.02.009. URL `http://www.sciencedirect.com/science/article/pii/S0096300313001252`.

T. Akerstedt, K. Hume, D. Minors, and J. Waterhouse. Good sleep–its timing and physiological sleep characteristics. *Journal of sleep research*, 6(4):221, 1997.

V. Baladandayuthapani, B. K. Mallick, and R. J. Carroll. Spatially adaptive Bayesian regression splines. *Journal of Computational and Graphical Statistics*, 14:378–394, 2005.

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

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

A. Bhattacharya and D. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98: 291–306, 2011.

R. J. Boik. Spectral models for covariance matrices. *Biometrika*, 89(1):159–182, 2002. ISSN 00063444. URL `http://www.jstor.org/stable/4140565`.

M. H. Bonnet. Sleep restoration as a function of periodic awakening, movement, or electroencephalographic change. *Sleep*, 10(4):364–373, 1987.

J. Buckelmüller, H.-P. Landolt, H. Stassen, and P. Achermann. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience*, 138(1):351–356, 2006.

H. Cardot. Conditional functional principal components analysis. *Scandinavian journal of statistics*, 34(2):317–335, 2007.

C. Carroll, A. Gajardo, Y. Chen, X. Dai, J. Fan, P. Z. Hadjipantelis, K. Han, H. Ji, H.-G. Mueller, and J.-L. Wang. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2020. URL `https://CRAN.R-project.org/package=fdapace`. R package version 0.5.5.

P. E. Castro, W. H. Lawton, and E. Sylvestre. Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28(4):329–337, 1986.

K. Chen and H. Müller. Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500):1599–1609, 2012. doi: 10.1080/01621459.2012.734196. URL `https://doi.org/10.1080/01621459.2012.734196`.

K. Chen, P. Delicado, and H. Müller. Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society, Series B*, 79(1):177–196, 2017. doi: 10.1111/rssb.12160. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12160`.

A. Chiang, C. Rennie, P. Robinson, S. Van Albada, and C. Kerr. Age trends and sex differences of alpha rhythms including split alpha peaks. *Clinical Neurophysiology*, 122(8): 1505–1517, 2011.

S. Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432):1313–1321, 1995.

J.-M. Chiou, Y.-T. Chen, and Y.-F. Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pages 1571–1596, 2014.

J. M. Chiou, Y. F. Yang, and Y. T. Chen. Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis*, 146:301–312, 2016.

L. Cragg, N. Kovacevic, A. R. McIntosh, C. Poulsen, K. Martinu, G. Leonard, and T. Paus. Maturation of eeg power spectra in early adolescence: a longitudinal study. *Developmental science*, 14(5):935–943, 2011.

C. Crainiceanu, D. Ruppert, R. Carroll, A. Joshi, and B. Goodner. Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16:265–288, 2007.

C. M. Crainiceanu, B. S. Caffo, C.-Z. Di, and N. M. Punjabi. Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *Journal of the American Statistical Association*, 104(486):541–555, 2009.

A. Cuevas, M. Febrero, and R. Fraiman. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, 51(2):1063 – 1074, 2006. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2005.10.012. URL `http://www.sciencedirect.com/science/article/pii/S0167947305002793`.

M. Delattre, M. Lavielle, and M. Poursat. A note on bic in mixed-effects models. *Electron. J. Statist.*, 8(1):456–475, 2014. doi: 10.1214/14-EJS890. URL `https://doi.org/10.1214/14-EJS890`.

D. K. Dey, A. E. Gelfand, T. B. Swartz, and P. K. Vlachos. A simulation-intensive approach for checking hierarchical models. *Test*, 7(2):325–346, 1998.

C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *Ann. Appl. Stat.*, 3(1):458–488, 2009. doi: 10.1214/08-AOAS206. URL `https://doi.org/10.1214/08-AOAS206`.

A. Dickinson, C. DiStefano, D. Senturk, and S. S. Jeste. Peak alpha frequency is a neural marker of cognitive function across the autism spectrum. *European Journal of Neuroscience*, 47(6):643–651, 2018.

R. E. Dustman, D. E. Shearer, and R. Y. Emmerson. Life-span changes in eeg spectral amplitude, amplitude variability and mean frequency. *Clinical neurophysiology*, 110(8): 1399–1409, 1999.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

F. Ferraty, I. Van Keilegom, and P. Vieu. On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics*, 37(2):286–306, 2010. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/41000880.

B. N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

B. N. Flury. Two generalizations of the common principal component model. *Biometrika*, 74(1):59–69, 03 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.1.59. URL https://doi.org/10.1093/biomet/74.1.59.

E. B. Fox and D. B. Dunson. Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research*, 16(1):2501–2542, 2015.

A. M. Franks and P. Hoff. Shared subspace models for multi-group covariance estimation. *Journal of Machine Learning Research*, 20(171):1–37, 2019.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002. doi: 10.1198/016214502760047113. URL https://doi.org/10.1198/016214502760047113.

J. Goldsmith, S. Greven, and C. Crainiceanu. Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51, 2013.

J. Goldsmith, V. Zipunnikov, and J. Schrack. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353, 2015.

J. Goldsmith, X. Liu, J. Jacobson, and A. Rundle. New insights into activity patterns in children, found using functional data analyses. *Medicine and science in sports and exercise*, 48(9):1723–1729, 2016. doi: 10.1249/MSS.0000000000000968. URL `https://www.ncbi.nlm.nih.gov/pubmed/27183122`.

J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss. *refund: Regression with Functional Data*, 2020. URL `https://CRAN.R-project.org/package=refund`. R package version 0.1-23.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711. URL `https://dx.doi.org/10.1093/biomet/82.4.711`.

S. Greven and F. Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2):1–35, 2017.

S. Greven, C. Crainiceanu, B. Caffo, and D. Reich. Longitudinal functional principal component analysis. *Electron. J. Statist.*, 4:1022–1054, 2010. doi: 10.1214/10-EJS575. URL `https://doi.org/10.1214/10-EJS575`.

W. Guo. Functional mixed effects models. *Biometrics*, 58(1):121–128, 2002.

C. Happ and S. Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018.

K. Hasenstab, C. A. Sugar, D. Telesca, K. McEvoy, S. Jeste, and D. Şentürk. Identifying longitudinal trends within eeg experiments. *Biometrics*, 71(4):1090–1100, 2015.

K. Hasenstab, A. Scheffler, D. Telesca, C. A. Sugar, S. Jeste, C. DiStefano, and D. Şentürk. A multi-dimensional functional principal components analysis of eeg data. *Biometrics*, 73 (3):999–1009, 2017.

HFD. Human fertility database. *Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria)*, 2019. URL `https://www.humanfertility.org`.

B. Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

P. D. Hoff. A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):971–992, 2009.

P. D. Hoff and X. Niu. A covariance regression model. *Statistica Sinica*, pages 729–753, 2012.

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

H. H. Jasper. The ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.*, 10:370–375, 1958.

S. Javaheri, Y. Y. Zhao, N. M. Punjabi, S. F. Quan, D. J. Gottlieb, and S. Redline. Slow-wave sleep is associated with incident hypertension: the sleep heart health study. *Sleep*, 41(1):zsx179, 2018.

S. S. Jeste, N. Kirkham, D. Senturk, K. Hasenstab, C. Sugar, C. Kupelian, E. Baker, A. J. Sanders, C. Shimizu, A. Norona, T. Paparella, S. F. Freeman, and S. P. Johnson. Electrophysiological evidence of heterogeneity in visual statistical learning in young children with asd. *Developmental Science*, 18(1):90–105, 2015. doi: 10.1111/desc.12188.

C.-R. Jiang, J.-L. Wang, et al. Covariate adjusted functional principal components analysis for longitudinal data. *The Annals of Statistics*, 38(2):1194–1226, 2010.

T. Jiang and Y. Qi. Likelihood ratio tests for high-dimensional normal distributions. *Scandinavian Journal of Statistics*, 42(4):988–1009, 2015.

K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34, 1946.

D. R. Kowal. Dynamic regression models for time-ordered functional data. *Bayesian Analysis*, 1(1):1–29, 2021.

D. R. Kowal and D. C. Bourgeois. Bayesian function-on-scalars regression for high-dimensional data. *Journal of Computational and Graphical Statistics*, pages 1–10, 2020.

D. R. Kowal, D. S. Matteson, and D. Ruppert. A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112(518):733–744, 2017.

T. Krivobokova and G. Kauermann. A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480):1328–1337, 2007.

T. Krivobokova, T. Kneib, and G. Claeskens. Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863, 2010.

M. G. Kundu, J. Harezlak, and T. W. Randolph. Longitudinal functional models with structured penalties. *Statistical modelling*, 16(2):114–139, 2016.

S. Lang and A. Brezger. Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004.

J. Lee and J. Lee. Robust sparse bayesian infinite factor models, 2020.

W. Lee, M. F. Miranda, P. Rausch, V. Baladandayuthapani, M. Fazio, J. C. Downs, and J. S. Morris. Bayesian semiparametric functional mixed models for serially correlated functional data, with application to glaucoma data. *Journal of the American Statistical Association*, 114(526):495–513, 2019a. doi: 10.1080/01621459.2018.1476242. URL `https://doi.org/10.1080/01621459.2018.1476242`.

W. Lee, M. F. Miranda, P. Rausch, V. Baladandayuthapani, M. Fazio, J. C. Downs, and J. S. Morris. Bayesian semiparametric functional mixed models for serially correlated functional data, with application to glaucoma data. *Journal of the American Statistical Association*, 114(526):495–513, 2019b. doi: 10.1080/01621459.2018.1476242. URL `https://doi.org/10.1080/01621459.2018.1476242`.

B. Li, L. Bruyneel, and E. Lesaffre. A multivariate multilevel gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in medicine*, 33(11):1877–1899, 2014.

Q. Li, J. Shamshoian, D. Şentürk, C. Sugar, S. Jeste, C. DiStefano, D. Telesca, et al. Region-referenced spectral power dynamics of eeg signals: A hierarchical modeling approach. *Annals of Applied Statistics*, 14(4):2053–2068, 2020.

M. Loève. Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162, 1946.

C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. The autism diagnostic observation schedule— generic: A standard measure of social and communication deficits associated with the spectrum of autism. *ournal of autism and developmental disorders*, 30(3):205–223, 2000.

B. Lynch and K. Chen. A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika*, 105(4):815–831, 2018a.

B. Lynch and K. Chen. A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika*, 105(4):815–831, 2018b.

B. A. Mander, J. R. Winer, and M. P. Walker. Sleep and human aging. *Neuron*, 94(1):19–36, 2017.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 83:69–70.

F. Mezzadri. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

V. Miskovic, X. Ma, C.-A. Chou, M. Fan, M. Owens, H. Sayama, and B. E. Gibb. Developmental changes in spontaneous electrocortical activity and network organization from early to late childhood. *Neuroimage*, 118:237–247, 2015.

S. Montagna, S. Tokdar, B. Neelon, and D. Dunson. Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, 68:1064–1073, 2012.

J. Morris, V. Baladandayuthapani, R. Herrick, P. Sanna, and H. Gutstein. Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The Annals of Applied Statistics*, 5(2A):894–923, 2011.

J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199, 2006. doi: 10.1111/j.1467-9868.2006.00539.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2006.00539.x`.

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

J. S. Morris, M. Vannucci, P. J. Brown, and R. J. Carroll. Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98(463):573–583, 2003b. doi: 10.1198/016214503000000422. URL `https://doi.org/10.1198/016214503000000422`.

R. M. Neal. Regression and classification using gaussian process priors (with discussion). *Bayesian statistics 6*, pages 475–501, 1999.

A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

S. Park and A. Staicu. Longitudinal functional data analysis. *Stat*, 4(1):212–226, 2015a. doi: 10.1002/sta4.89. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.89.

S. Y. Park and A.-M. Staicu. Longitudinal functional data analysis. *Stat*, 4(1):212–226, 2015b.

F. Perrin, J. Pernier, O. Bertrand, and J. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2): 184–187, 1989.

S. Purcell. *luna: Analysis of Sleep Signal Data*, 2020. https://github.com/remnrem/luna, http://zzz.bwh.harvard.edu/luna.

S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

A. Quintero and E. Lesaffre. Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal*, 59(5):1047–1066, 2017.

J. O. Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences*, 4, 2004.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

S. Ren, H. Lai, W. Tong, M. Aminzadeh, X. Hou, and S. Lai. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, 37(9):1487–1498, 2010. doi: 10.1080/02664760903046102. URL https://doi.org/10.1080/02664760903046102.

E. Rodríguez-Martínez, F. Ruiz-Martínez, C. B. Paulino, and C. M. Gómez. Frequency shift in topography of spontaneous brain rhythms from childhood to adulthood. *Cognitive neurodynamics*, 11(1):23–33, 2017.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392, 2009.

A. Scheffler, D. Telesca, Q. Li, C. Sugar, C. Distefano, S. Jeste, and D. Şentürk. Hybrid principal components analysis for region-referenced longitudinal functional eeg data. *Biostatistics*, 21(1):139–157, 2020a.

A. W. Scheffler, D. Telesca, C. A. Sugar, S. Jeste, A. Dickinson, C. DiStefano, and D. Şentürk. Covariate-adjusted region-referenced generalized functional linear model for eeg data. *Statistics in Medicine*, 38(30):5587–5602, 2019.

A. W. Scheffler, A. Dickinson, C. DiStefano, S. Jeste, and D. Şentürk. Covariate-adjusted hybrid principal components analysis. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 391–404. Springer, 2020b.

F. Scheipl, A.-M. Staicu, and S. Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015.

J. R. Schott. Some tests for common principal component subspaces in several groups. *Biometrika*, 78(4):771–777, 1991. ISSN 00063444. URL `http://www.jstor.org/stable/2336928`.

J. R. Schott. Partial common principal component subspaces. *Biometrika*, 86(4):899–908, 1999. ISSN 00063444. URL `http://www.jstor.org/stable/2673593`.

S. J. Segalowitz, D. L. Santesso, and M. K. Jetha. Electrophysiological changes during adolescence: a review. *Brain and cognition*, 72(1):86–100, 2010.

J. Shamshoian, D. Şentürk, S. Jeste, and D. Telesca. Bayesian analysis of longitudinal and multidimensional functional data. *Biostatistics*, 2020.

J. Shi and T. Choi. *Gaussian Process Regression Analysis for Functional Data*. New York: Chapman and Hall/CRC, 2011.

J. Shi, B. Wang, E. Will, and R. West. Mixed-effects gaussian process functional regression models with application to dose–response curve prediction. *Statistics in medicine*, 31(26): 3165–3177, 2012.

J. Q. Shi, R. Murray-Smith, and D. Titterington. Hierarchical gaussian process mixtures for regression. *Statistics and computing*, 15(1):31–41, 2005.

J. Q. Shi, B. Wang, R. Murray-Smith, and D. M. Titterington. Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723, 2007. doi: 10.1111/j.1541-0420.2007.00758.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2007.00758.x.

H. Shou, V. Zipunnikov, C. M. Crainiceanu, and S. Greven. Structured functional principal component analysis. *Biometrics*, 71(1):247–257, 2015. doi: 10.1111/biom.12236. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12236.

R. J. Somsen, B. J. van't Klooster, M. W. van der Molen, H. M. van Leeuwen, and R. Licht. Growth spurts in brain maturation during middle childhood as indexed by eeg power spectra. *Biological psychology*, 44(3):187–209, 1997.

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

T. Stroganova, E. Orekhova, and I. Posikera. Eeg alpha rhythm in infants. *Clinical Neurophysiology*, 110:997–1012, 1999.

A. Suarez and S. Ghosal. Bayesian estimation of principal components for functional data. *Bayesian Anal.*, 12(2):311–333, 2017. doi: 10.1214/16-BA1003. URL https://doi.org/10.1214/16-BA1003.

W. K. Thompson and O. Rosen. A bayesian model for sparse functional data. *Biometrics*, 64(1):54–63, 2008.

T. Tokuda, B. Goodrich, I. Van Mechelen, A. Gelman, and F. Tuerlinckx. Visualizing distributions of covariance matrices. *Columbia Univ., New York, USA, Tech. Rep*, pages 18–18, 2011.

P. A. Valdés-Hernández, A. Ojeda-González, E. Martínez-Montes, A. Lage-Castellanos, T. Virués-Alba, L. Valdés-Urrutia, and P. A. Valdes-Sosa. White matter architecture

rather than cortical surface area correlates with the eeg alpha rhythm. *Neuroimage*, 49 (3):2328–2339, 2010.

E. Van Cauter, R. Leproult, and L. Plat. Age-related changes in slow wave sleep and rem sleep and relationship with growth hormone and cortisol levels in healthy men. *Jama*, 284 (7):861–868, 2000.

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

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

M. P. Walker. The role of slow wave sleep in memory processing. *Journal of Clinical Sleep Medicine*, 5(2 suppl):S20–S26, 2009.

B. Wang and J. Q. Shi. Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association*, 109(507):1123–1133, 2014. doi: 10.1080/01621459.2014.889021. URL `https://doi.org/10.1080/01621459.2014.889021`.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

S. Watanabe and M. Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in Graphical Models*, 1998.

S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2006. ISBN 9781420010404. URL `https://books.google.com/books?id=xhPNBQAAQBAJ`.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

L. Xiao, Y. Li, and D. Ruppert. Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, pages 577–599, 2013.

L. Xiao, L. Huang, J. A. Schrack, L. Ferrucci, V. Zipunnikov, and C. M. Crainiceanu. Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics*, 16(2):352–367, 2015.

L. Xiao, V. Zipunnikov, D. Ruppert, and C. Crainiceanu. Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26 (1):409–421, 2016. ISSN 1573-1375. doi: 10.1007/s11222-014-9485-x. URL https://doi.org/10.1007/s11222-014-9485-x.

J. Yang, H. Zhu, T. Choi, and D. Cox. Smoothing and mean–covariance estimation of functional data with a bayesian hierarchical model. *Bayesian Anal.*, 11(3):649–670, 2016a. doi: 10.1214/15-BA967. URL https://doi.org/10.1214/15-BA967.

J. Yang, H. Zhu, T. Choi, D. D. Cox, et al. Smoothing and mean–covariance estimation of functional data with a bayesian hierarchical model. *Bayesian Analysis*, 11(3):649–670, 2016b.

J. Yang, D. Cox, J. Lee, and P. R. T. Choi. Efficient bayesian hierarchical functional data analysis with basis function approximations using gaussian-wishart processes. *Biometrics*, 73(4):1082–1091, 2017. doi: 10.1111/biom.12705. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12705.

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

Y. Yuan and V. E. Johnson. Goodness-of-fit diagnostics for bayesian hierarchical models. *Biometrics*, 68(1):156–164, 2012.

J. Zapata, S.-Y. Oh, and A. Petersen. Partial separability and functional graphical models for multivariate gaussian processes. *arXiv preprint arXiv:1910.03134*, 2019.

C. Zhang, B. Shahbaba, and H. Zhao. Variational hamiltonian monte carlo via score matching. *Bayesian Analysis*, 13(2):485–506, 2018.

J. Zhang, G. J. Siegle, W. D'Andrea, and R. T. Krafty. Interpretable principal components analysis for multilevel multivariate functional data, with application to eeg experiments. *arXiv preprint arXiv:1909.08024*, 2019.

L. Zhang, V. Baladandayuthapani, H. Zhu, K. A. Baggerly, T. Majewski, B. A. Czerniak, and J. S. Morris. Functional car models for large spatially correlated functional datasets. *Journal of the American Statistical Association*, 111(514):772–786, 2016.

Y. Zhao, B. Wang, S. Mostofsky, B. Caffo, and X. Luo. Covariate assisted principal regression for covariance matrix outcomes. *bioRxiv*, page 425033, 2018.

V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu. Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*, 20(4):852–873, 2011.

V. Zipunnikov, S. Greven, H. Shou, B. Caffo, D. S. Reich, and C. Crainiceanu. Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The annals of applied statistics*, 8(4):2175, 2014.